

Sztuczne sieci neuronowe

Marek Grochowski

Wydział Fizyki, Astronomii i Informatyki Stosowanej UMK

<http://www.fizyka.umk.pl/~grochu/nn>
grochu@is.umk.pl

11 lipca 2023

Plan

1. Perceptrony, sieci wielowarstwowe jednokierunkowe (MLP)
2. Metody uczenia i algorytm wstecznej propagacji błędu
3. Radialne funkcje bazowe (RBF) i metody aproksymacji
4. Samoorganizacja, sieci SOM, uczenie konkurencyjne
5. Sieci dynamiczne: model Hopfielda, maszyny Boltzmana
6. Głębokie sieci neuronowe (DNN)
7. Sieci splotowe (CNN)
8. Sieci rekurencyjne (RNN) i uczenie sekwencji
9. Autoenkodery, wykrywanie cech, DBN
10. Sieci typu GAN

Literatura I

-  Osowski S., *Sieci neuronowe w ujęciu algorytmicznym*, Wydawnictwo Naukowo-Techniczne, Warszawa 1996
-  Tadeusiewicz R., *Sieci neuronowe*, Akademicka Oficyna Wydawnicza RM, Warszawa 1993
-  Ryszard Tadeusiewicz, Tomasz Gąciarz, Barbara Borowik, Bartosz Lepe, *Odkrywanie właściwości sieci neuronowych przy użyciu programów w języku C#*, Polska Akademia Umiejętności, 2008
-  J. Żurada, M. Barski, W., Jędruch *Sztuczne sieci neuronowe*, Wydawnictwo Naukowe PWN 1996

Literatura II



Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning*. MIT Press, 2016,
<http://www.deeplearningbook.org>



Michael Nielsen, *Neural Networks and Deep Learning*,
<http://neuralnetworksanddeeplearning.com/>



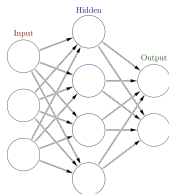
Julien Vitay, *Neurocomputing*,
<https://julien-vitay.net/lecturenotes-neurocomputing/>



Denny Britz, *Deep Learning Glossary*,
<http://www.wildml.com/deep-learning-glossary/>

Sztuczne Sieci Neuronowe

Systemy obliczeniowe inspirowane biologicznymi sieciami neuronowymi składające się ze zbioru prostych jednostek przetwarzających sygnał (neuronów) komunikujących się między sobą za pomocą połączeń (wagi połączeń).



- model matematyczny $f(\mathbf{x}; \mathbf{W}) = \mathbf{y}$
- systemy uczące się, adaptowanie wag $\mathbf{W}$, Uczenie Maszynowe
- systemy inteligentne: Inteligencja Obliczeniowa, Sztuczna Inteligencja
- równoległe przetwarzanie sygnału

Inteligencja obliczeniowa

Computational Intelligence (CI) zajmuje się rozwiązywaniem problemów, które nie są efektywnie algorytmizowalne.

Cechą wielu systemów CI jest rozwiązywanie zadań na podstawie znanych przykładów, uczenie się z empirycznych danych zamiast programowania rozwiązania.

Problemy niealgorytmizowalne:

- zagadnienie jest zbyt złożone, np. problemy NP-trudne
- modelowany proces może zawierać trudne do zdefiniowania niejasności lub może być stochastyczny z natury
- procesu nie da się opisać przez zrozumiałe zasady
- warunki mogą się zmieniać, algorytm musi się dostosować do nowej sytuacji

Problemy efektywnie niealgorytmizowalne

Problemy NP-trudne - Liczba kroków algorytmu dla złożonych sytuacji rośnie w sposób szybszy niż jakikolwiek wielomian liczby elementów (złożoności specyfikacji problemu).

Przykład: problem komiwojażera



Rysunek 1.1. Najkrótsza trasa premiera przebiegająca przez wszystkie miasta wojewódzkie w podziale administracyjnym z 1975 roku (A. Adrański)

Tabela 1.1. Czasy znalezienia przez komputer, wykonujący 100 miliardów operacji na sekundę, najkrótszej trasy podróży premiera (zob. ćwiczenie 1.3)

Liczba województw	Czas znajdowania najkrótszej trasy
17	3.5 minuty
25	$2 \cdot 10^5$ lat
49	$4 \cdot 10^{42}$ lat

Dla 100 miejsz mamy 100! (liczba 161 cyfrowa)

Rysunek: M. Sysło, „Algorytmy”

Problemy niealgorytmizowalne - przykłady

- rozumienie sensu zdań,
- działania twórcze, decyzje intuicyjne,
- rozpoznawanie twarzy i obrazów,
- rozpoznawanie pisma ręcznego,
- rozpoznawanie mowy i sygnałów, percepcja,
- sterowanie robotem, nieliniowymi układami,
- diagnostyka medyczna, planowanie terapii.

CI i sztuczna inteligencja (AI)

Klasyczne rozumienie AI i CI:

CI - percepcja i sterowanie: zachowania sensomotoryczne – sieci neuronowe i uczenie maszynowe

AI - wyższe czynności poznawcze: logika, język, rozumowanie, rozwiązywanie problemów.

Good old-fashion **Artificial Intelligence (GOFAI)** to część CI posługująca się symboliczną reprezentacją wiedzy, zajmuje się rozumowaniem, tworzeniem systemów ekspertowych.

CI i sztuczna inteligencja (AI)

Obecnie „sztuczna inteligencja” używana jest na określenie wszystkiego, co się kojarzy z inteligencją maszyn i algorytmów, które naśladują inteligentne zachowania człowieka.

- **Strong AI, Artificial General Intelligence** - agent świadomie działający i uczący się działać w środowisku (podobnie do człowieka)
- **Weak AI** - maszyna rozwiązująca (ucząca się rozwiązywać) konkretne zadanie

Efekt AI (twierdzenie Teslera)

„AI is whatever hasn't been done yet”

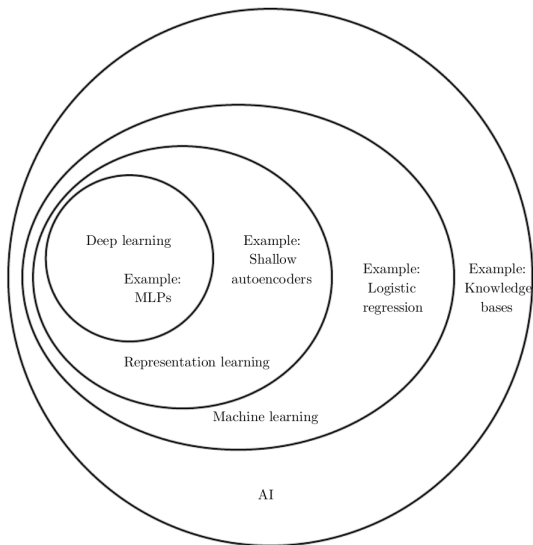
Uczenie maszynowe

- **Uczenie maszynowe** (Machine learning, ML) - systemy, które uczą się rozwiązywać problemy na podstawie dostarczonych przypadków uczących (zebranych danych)
- **Generalizacja** - dobrze wyuczony model działa poprawnie także dla przypadków, których nie było w danych treningowych
- Gdy dane się zmieniają system można dostosować trenując go na nowych danych



Rysunek: <https://xkcd.com/>

ANN i sztuczna inteligencja (AI)

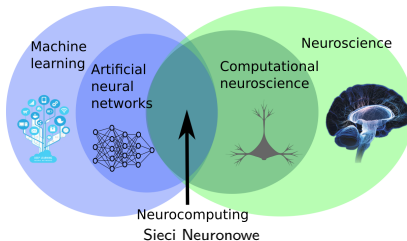


Gooffellow, 2016 [1]

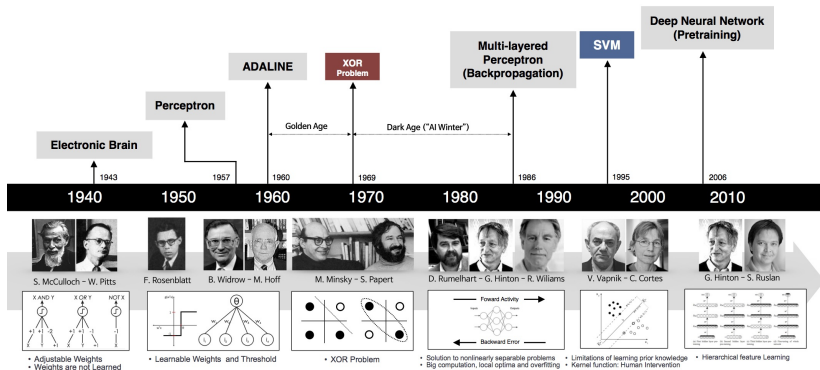
Obliczenia neuronowe

- **modelowanie działania mózgu** - symulacje komputerowe, modele procesów kognitywnych
- przetwarzanie równoległe bardzo wydajne przy rozwiązywaniu niektórych problemów
- odporność na uszkodzenia, oddziaływania lokalne, uszkodzenie części sieci nie musi powodować drastycznych skutków
- rozwiązywanie praktycznych problemów za pomocą metod inspirowanych systemami biologicznymi

-> **Sztuczne Sieci Neuronowe (Artificial Neural Networks)**



Zródło: J. Vitay [7]



https://beamandrew.github.io/deeplearning/2017/02/23/deep_learning_101_part1.html

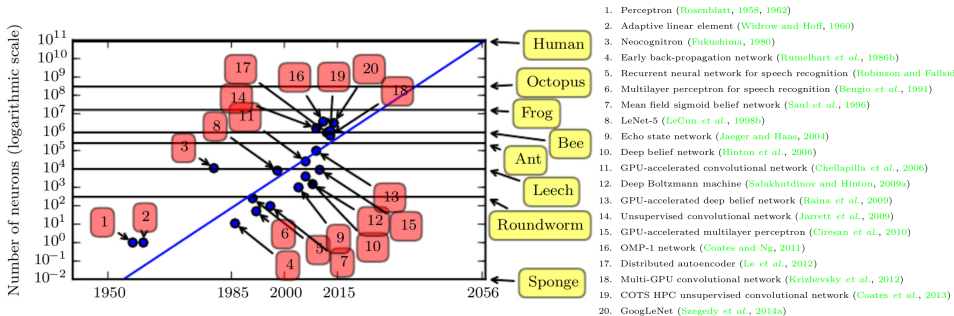
Rys historyczny I

- 1938 N. Rashevsky, neurodynamika - sieci neuronowe jako układ dynamiczny
- 1943 W. McCulloch, W. Pitts, sieci neuronowe jako układy logiczne
- 1949 D. Hebb, uczenie się na poziomie synaptycznym (pierwsza reguła uczenia)
- 1952 A. Hodgkin, A. Huxley, szczegółowy matematyczny model neuronu kałamarnicy olbrzymiej (nagroda Nobla 1963)
- 1958 F. Rosenblatt, Perceptron - sieć jako funkcja matematyczna
J. von Neuman, The computer and the brain
- 1960 B. Widrow, M. Hoff, Adeline
- 1967 K. Steinbuch, E. Schmitt, Macierze uczące się
- 1969 M. Minsky, S. Papert, książka „Perceptrony” - wykazali ograniczenia sieci liniowych
- 1973 Chr. von der Malsburg, samoorganizacja w układzie nerwowym

Rys historyczny II

- 1982 J. Hopfield, model Hopfielda
T. Kohonen, samoorganizacja map topograficznych mózgu
- 1983 K. Fukushima, S. Miyake, T. Ito, Neokognitron, głęboka sieć neuronowa, inspiracja późniejszych sieci spłotowych
- 1985 D. Ackley, G. Hinton, T. Sejnowski, maszyny Boltzman
- 1986 D. Rumelhart, G. Hinton, R. Williams, wsteczna propagacja błędów
- 1987 G. Carpenter, S. Grossberg, model ART;
W. Freeman, Chaos w mózgu
- 1990 T. Poggio, F. Girosi, sieci RBF, teoria regularyzacji
- 1995 V. Vapnik, SVM i „koniec ery Data Mining”
- 1997 S. Hochreiter, J. Schmidhuber, sieć rekurencyjna LSTM
- 1998 Y. LeCun, sieci spłotowe w analizie obrazów
- 2005 GPU do sieci spłotowych, bardzo duże sieci, głębokie uczenie

Ilość neuronów w modelach neuronowych



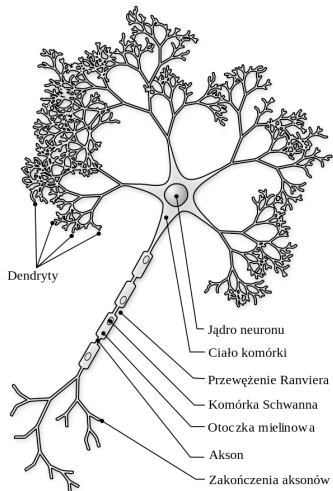
Liczba neuronów podwaja się co 2.5 roku

Goofellow, 2016 [1]

Mózg i sieć neuronowa

- 10^{11} neuronów (100 miliardów), do 100Hz
- gęstość połączeń około 10^4 synaps na neuron
- Receptory - dostarczają sygnały wejściowe (bodźce wewnętrzne i sygnał ze zmysłów)
- Sygnał wyjściowy z układu nerwowego steruje efektorami (reakcja na bodziec)
- obliczenia równoległe
- układy biologiczne zbyt złożone do bezpośredniego modelowania - jedynie ogólne zasady organizacji i sposobu funkcjonowania układów biologicznych neuronów mogą służyć za inspirację
- Sztuczna sieć neuronowa to model matematyczny, który można widzieć jako złożenie wielu (nieliniowych) funkcji i często nie ma nic wspólnego z sieciami biologicznymi

Neuron biologiczny



Rysunek: [wikipedia.org](https://pl.wikipedia.org/wiki/Neuron)

Neuron biologiczny

- **Dendryty** - wejście neuronu, odbierają sygnał od sąsiadujących neuronów, impulsy mogą być pobudzające lub hamujące
- **Ciało komórki (soma)**, akumuluje napływające w danym momencie sygnały
- **Akson** (wyjście), emituje sygnał, gdy potencjał w somie osiągnie odpowiedni próg
- **Synapsy** - styki pomiędzy dendrytami i aksonami, pobudzenie napływające z aksonu powoduje uwolnienie chemicznych neurotransmiterów, które wpływają na zachowanie połączeń neuronów
- Działanie neuronów biologicznych jest bardzo złożonym procesem, neurony w sztucznych sieciach są dużym uproszczeniem

👉 Schematic animation of spike generation in neural cell
(J. Piersa)

Logika progowa neuronu

Neuron zbiera dochodzące do niego sygnały x_i obliczając sumę ważoną tych sygnałów (wagi w_i określane są przez procesy synaptyczne), tworząc z nich sumaryczny sygnał wejściowy:

$$I(t) = \sum_i w_i x_i(t)$$

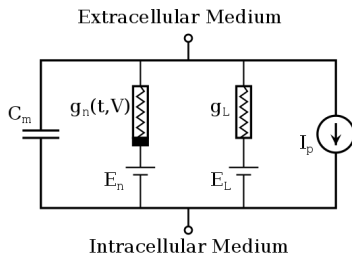
Sumowanie przestrzenne i czasowe potencjałów zgromadzonych na błonie komórkowej daje całkowitą aktywację elementu:

$$a(t) = F(I(t), a(t-1))$$

Aktywacja, razem z progiem wzbudzenia T , określa sygnał wyjściowy

$$o(t) = f(a(t), T)$$

Model Hodgkin–Huxley (1952)

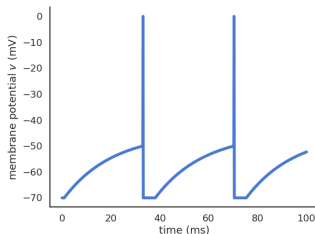


- model oparty na przewodnictwie
- zestaw nieliniowych równań różniczkowych, które przybliżają charakterystykę elektryczną komórek pobudliwych, takich jak neurony i komórki mięśniowe
- system dynamiczny w czasie ciągłym
- całkowity prąd przepływający przez membranę

$$I = C_m \frac{dV_m}{dt} + g_K (V_m - V_K) + g_{Na} (V_m - V_{Na}) + g_I (V_m - V_I)$$

Neurony impulsujące (*spiking neurons*)

Sumuj i odpal: *Leaky integrate-and-fire* (LIF; Llapicque, 1907)
neurony generują impuls, gdy sumaryczny potencjał membrany przekroczy próg



$$C \frac{dv}{dt} = -g_L (v - V_L) + I$$

Gdy $v > V_T$ neuron emituje impuls a następnie odpoczywa

Neurony kodowane częstością wzburzeń

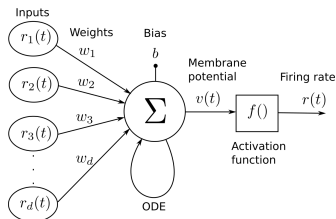
Rate-coded neurons

- potencjał membrany $v(t)$

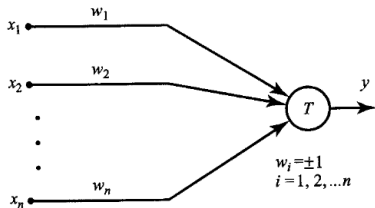
$$\tau \frac{dv(t)}{dt} + v(t) = \sum_{i=1}^d w_{i,j} r_i(t) + b$$

- częstość wzburzeń (*firing rate*) $r(t)$ odwzorowuje potencjał membrany w pojedynczą wartość za pomocą **funkcji aktywacji**

$$r(t) = f(v(t))$$

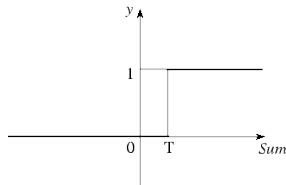


Neuron McCulloch, Pitts (1943-49)



stan wyjściowy neuronu

$$o(\mathbf{x}) = \begin{cases} 1 & \text{gdy } \sum_{i=1}^n w_i x_i \geq T \\ 0 & \text{gdy } \sum_{i=1}^n w_i x_i < T \end{cases}$$



grafika: www.wikipedia.org

Neuron McCulloch, Pitts (1943-49)

- neurony mogą być tylko w dwóch stanach, aktywne albo nieaktywne, sygnał wejściowy i wyjściowy binarny

$$x_i \in \{0, 1\} \quad o(\mathbf{x}) \in \{0, 1\}$$

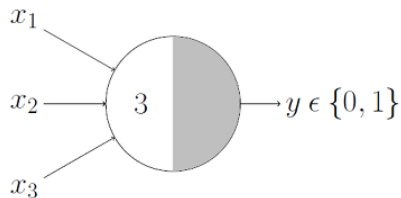
- sygnały dochodzą przez synapsy pobudzające i hamujące

$$w_{ij} \in \{-1, +1\}$$

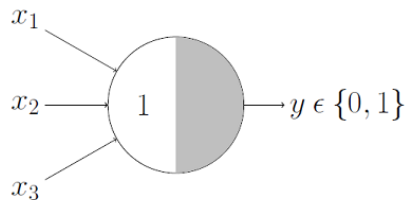
- każdy neuron ma ustalony próg pobudzenia $T > 0$
- sygnały sumują się w pewnym kwancie czasu i gdy przekroczą próg T to neuron się aktywuje
- brak uczenia się - wagi i próg mają stałą wartość
- odpowiedni dobór wag pozwala zrealizować funkcje logiczne: NOT, OR, AND bądź NOR i NAND -> sieć neuronów może zrealizować dowolną funkcję logiczną

Realizacja funkcji logicznych

AND



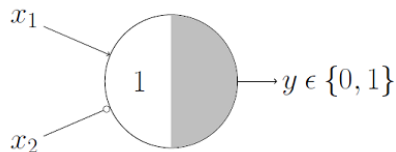
OR



NOT



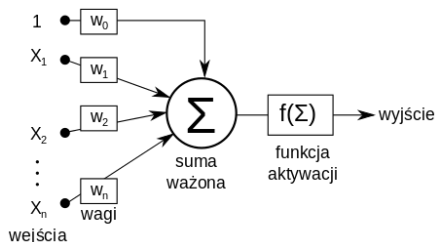
hamowanie sygnału wejściem x_2



Źródło: Akshay L. Chandra,  McCulloch-Pitts Neuron

$$x_1 \text{ AND } !x_2^*$$

Ogólny schemat neuronu



$$o(\mathbf{x}) = f\left(\sum_{i=1}^n w_i x_i + w_0\right) = f(\mathbf{w}^T \mathbf{x})$$

gdzie

$\mathbf{x} = [1, x_1, x_2, \dots, x_n]$ wielowymiarowy sygnał wejściowy $\mathbf{x} \in \mathbb{R}^{n+1}$

$\mathbf{w} = [w_0, w_1, w_2, \dots, w_n]$ wagi połączeń $\mathbf{w} \in \mathbb{R}^{n+1}$

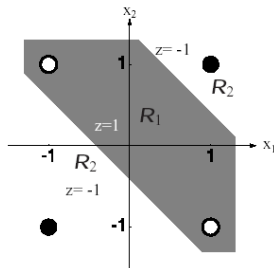
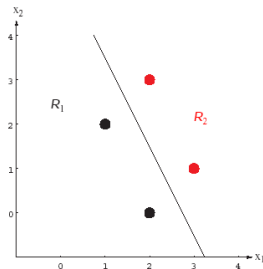
w_0 - wyraz wolny (bias), stały sygnał wejściowy $1 \cdot w_0$

Interpretacja geometryczna - separowalność

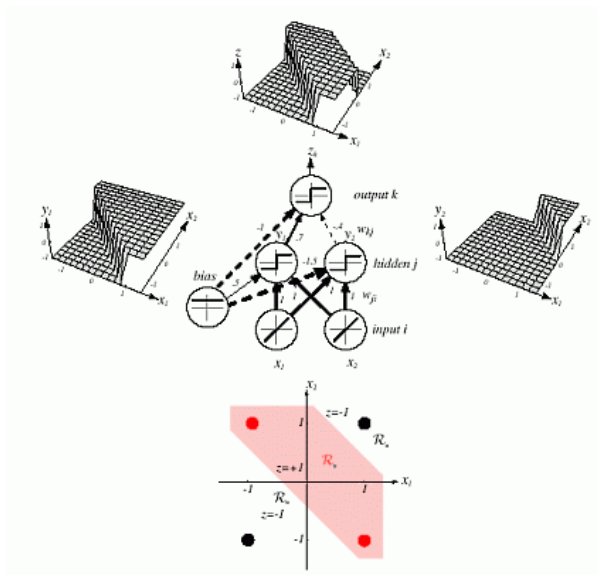
Neuron z progiem realizuje separowalność liniową, hiperpłaszczyzna dzieli przestrzeń wejściową $\mathbb{R}^n$ na dwie części

Przykład neuronu z 2 wejściami

$$o(\mathbf{x}) = \begin{cases} 1 & \text{gdy } w_1x_1 + w_2x_2 \geq \theta \\ 0 & \text{gdy } w_1x_1 + w_2x_2 < \theta \end{cases}$$



XOR



Reguła Hebba (1949)

$$\Delta w_i = \eta \cdot o(\mathbf{x}) \cdot x_i = \eta f(\mathbf{w}^T \mathbf{x}) x_i$$

„Kiedy akson komórki A jest dostatecznie blisko by pobudzić komórkę B i wielokrotnie w sposób trwały bierze udział w jej pobudzaniu, procesy wzrostu lub zmian metabolicznych zachodzą w obu komórkach tak, że sprawność neuronu A jako jednej z komórek pobudzających B, wzrasta.”

„neurons that fire together wire together”

Plastyczność układu (uczenie)

Uczenie się: zmiana wag synaptycznych w_i w czasie

Reguła Hebba

$$\Delta w_i = \eta o(\mathbf{x}) x_i = \eta f(\mathbf{w}^T \mathbf{x}) x_i$$

Reguła delta (Widrowa-Hoffa)

$$\Delta w_i = \eta (y - o(\mathbf{x})) x_i$$

gdzie y to spodziewana odpowiedź dla sygnału wejściowego $\mathbf{x}$

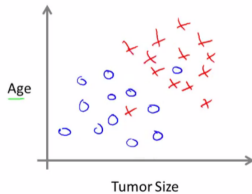
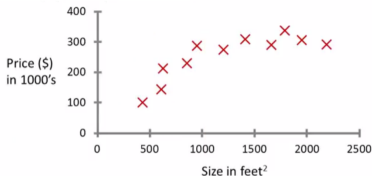
Rodzaje ucznia

- **Uczenie bez nadzoru** (*unsupervised learning*)
uczenie wyłącznie na podstawie sygnały wejściowego $\mathbf{X}$,
odkrywanie ciekawych struktur w przestrzeni danych
wejściowych.
Uczenie spontaniczne, odpowiada uczeniu w okresie
niemowlęcym.
- **Uczenie nadzorowane** (*supervised learning*) - dla każdego
sygnału wejściowego x dana jest pożądana odpowiedź y .
Uczenie „szkolne”.
- **Uczenie z krytykiem**, ze „wzmocnieniem” (*reinforcement learning*) - uczenie się zachowań, które przynoszą zysk po
dłuższym czasie (np. autonomiczne roboty, gra z
przeciwnikiem).
Uczenie dojrzałe (nabieranie „mądrości”).

Uczenie nadzorowane

Każdy przypadek x ze zbioru treningowego posiada przypisaną wartość y

- **Regresja (aproksymacja)** - obiekt wyjściowy y jest liczbą rzeczywistą lub wektorem liczb (np. temperatura)
- **Klasyfikacja** - obiekt wyjściowy y jest etykietą klasy (np. nazwa obiektu na zdjęciu)



Rysunek: Mohamad Ghassany, <http://www.mghassany.com/MLcourse/introduction.html>

Uczenie nadzorowane

- Model

$$f(\mathbf{x}; \mathbf{W}) = \mathbf{y}$$

realizuje mapowanie zdefiniowane przez parametry sieci $\mathbf{W}$ (wagi połączeń) wektora treningowego $\mathbf{x}$ do wektora wyjściowego $\mathbf{y}$ (pożądanego wyjścia)

- Uczenie najczęściej polega na takim doborze parametrów $\mathbf{W}$ aby zminimalizować różnice pomiędzy wyjściem sieci $f(\mathbf{x})$ a wartością pożądaną $\mathbf{y}$, np.:

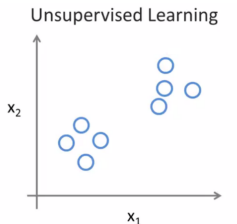
$$(f(\mathbf{x}; \mathbf{W}) - \mathbf{y})^2$$

- Celem uczenia nadzorowanego nie jest uczenie „na pamięć”, lecz **generalizacja**.

Uczenie nienadzorowane

Uczenie wyłącznie na podstawie sygnału wejściowego $\mathbf{X}$, brak pożądanego sygnału wyjściowego. Przykłady:

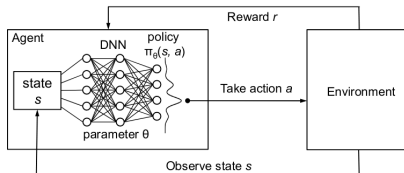
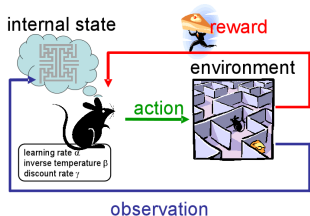
- **klasteryzacja** (analiza skupień), automatyczne wykrywanie grup o podobnych właściwościach



- wykrywanie cech i regularności, tworzenie przydatnych reprezentacji danych, kompresja (kodowanie) sygnału, redukcja wymiarowości, modelowanie sygnału, redukcja szumu

Uczenie z krytykiem

- Sygnał wyjściowy jest zazwyczaj akcją (sekwencją akcji) podejmowanych przez agenta
- Nagroda (lub kara) jest dostarczana z opóźnieniem (np. przegrana/wygrana partia w szachy)

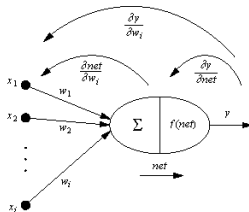


Rysunek: Hongzi Mao, „Resource Management with Deep Reinforcement Learning”
Aneek Das, The very basics of Reinforcement Learning

Główne aspekty modeli neuronowych I

1. Sposób modelowania pojedynczego neuronu

- stan aktywacji (wzbudzenia) poszczególnych neuronów, sposób aktywacji tych neuronów, funkcja opisująca sygnał wyjściowy elementu,
- neurony: progowe, liniowe, nieliniowe (np. sigmoidalne)



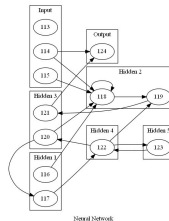
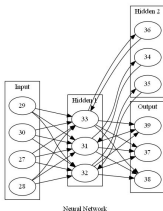
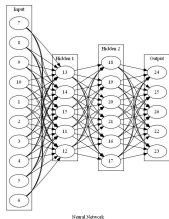
Główne aspekty modeli neuronowych II

2. Sposób propagacji sygnałów przez sieć neuronową

- sieci jednokierunkowe (feedforward)
- sieci ze sprzężeniami zwrotnymi, sieci rekurencyjne (recurrent), sieci dynamiczne

3. Topologia połączeń elementów (architektura sieci):

- budowa warstwowa (warstwa wejściowa, wyjściowa, warstwy ukryte), struktura hierarchiczna
- sieci jednowarstwowe, wielowarstwowe, płytkie, głębokie
- warstwy w pełni połączone, rzadkie połączenia, splot



Główne aspekty modeli neuronowych III

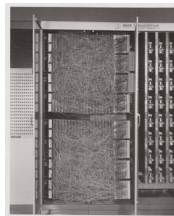
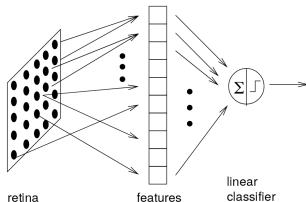
4. Reguły uczenia - reguły modyfikacji parametrów opisujących neurony i połączenia pomiędzy nimi, algorytm optymalizacji, funkcja kosztu, regularyzacja
5. Otoczenie, w którym działa sieć neuronowa: sposób prezentacji danych, rodzaj danych wejściowych, reprezentacja sygnału wyjściowego (np. kodowanie etykiet klas)
6. Realizacja techniczna: język programowania, wykorzystanie GPU, zrównoleglenie obliczeń

Perceptron

Reguły uczenia

Perceptron Rosenblatta (1958)

Klasyfikator neuronowy Mark I do rozpoznawania znaków alfanumerycznych, wzorowany na biologicznej percepcji (układ wzrokowy)



Trzy warstwy:

- **S-units:** wejście, siatkówka oka, 400 fotokomórki (20x20)
- **A-units:** asocjacyjne, zbierające dane z większych obszarów (obliczanie cech), 512 jednostek
- **R-units:** liniowy klasyfikator uczący, 8 wyjść

Rys: wikipedia.org

Martin Riedmiller, Machine Learning (lectures)

- $S_i = \{-1, +1\}$ sygnał docierający do elementów sensorycznych
- połączenia $c_{ij} = \{-1, 0, +1\}$ elementów S_j i A_i , przypadkowo rozrzucone w pewnym obszarze, nie ulegają zmianom, realizują wstępne przetwarzanie (ekstrakcja cech)

$$A_i = \begin{cases} +1, & \text{dla } \sum_j c_{ij} S_j \geq \Theta_i \\ -1, & \text{dla } \sum_j c_{ij} S_j < \Theta_i \end{cases}$$

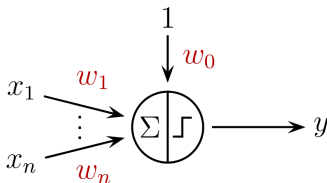
- sygnał wyjściowy

$$R_i = \begin{cases} +1 & \sum_j w_{ij} A_j > N\kappa \\ -1 & \sum_j w_{ij} A_j < -N\kappa \\ 0 & \text{w pozostałych przypadkach} \end{cases}$$

- wagi w_{ij} realizowane przez potencjometry
- proces uczenia - mechaniczna regulacja wartości potencjometrów za pomocą silników elektrycznych

Perceptron współczesny

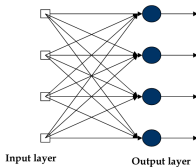
- Zwykle przez „perceptron” rozumie się teraz pojedynczy neuron z wieloma wejściami (bez jednostek S, bo tu nie ma adaptacji).
- **Perceptron prosty**



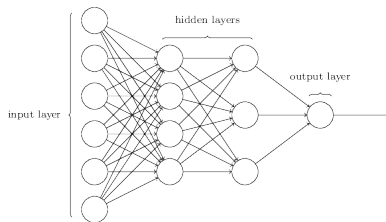
$$y = f \left(\sum_i w_i x_i + w_0 \right)$$

Rys: Martin Riedmiller, Machine Learning (lectures)

Single layer perceptron



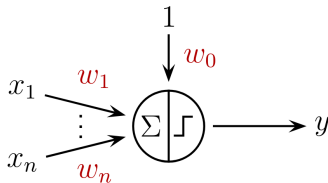
Multilayer perceptron (MLP)



Neuron binarny

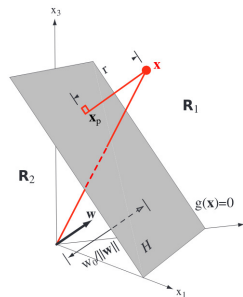
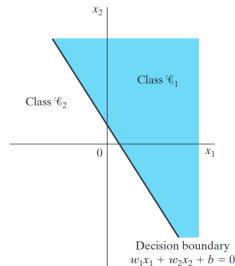
Perceptron prosty (binarny)

$$y = \begin{cases} 1, & \text{gdy } \sum_i w_i x_i + w_0 \geq 0 \\ 0, & \text{gdy } \sum_i w_i x_i + w_0 < 0 \end{cases}$$



Interpretacja geometryczna

- $x_1, \dots, x_n$ to współrzędne punktu w n -wymiarowej przestrzeni ($\mathbf{x} \in \mathbb{R}^n$)
- równanie $\sum x_i w_i + w_0 = 0$ definiuje płaszczyznę (hiperpłaszczyznę) rozdzielającą przestrzeń wejściową na dwie części R_1 i R_2
- perceptron prosty dzieli zbiór wektorów $\mathbf{x}_i$ na dwa podzbiory: te leżące poniżej płaszczyzny decyzyjnej ($y(\mathbf{x}_i) = 0$) oraz leżące powyżej płaszczyzny ($y(\mathbf{x}_i) = 1$)



Rys: Duda and Hart, Pattern Classification

Zadanie klasyfikacji binarnej

- **Klasyfikacja binarna** - przypisanie obiektów do jednej z 2 klas
- **Dane treningowe**: zbiór n przypadków $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, każdy przypisany do jednego z dwóch zbiorów $\mathbb{P}$ lub $\mathbb{N}$
- **Uczenie**: procedura doboru wag $\mathbf{w}$ i wartości progowej w_0 tak aby perceptron zwracał wartość 1 dla wszystkich $\mathbf{x}_i$ ze zbioru $\mathbb{P}$, zaś wartość 0 dla wszystkich $\mathbf{x}_i \in \mathbb{N}$
- założenie: dane są spójne, tzn. $\mathbb{P} \cap \mathbb{N} = \emptyset$

Uczenie perceptronu - idea

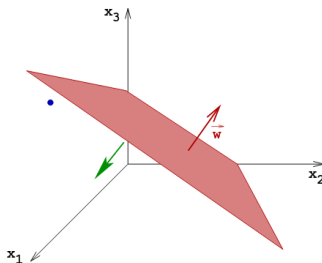
Niech $\mathbf{x}$ należy do zbioru $\mathbb{P}$. Jeżeli perceptron popełnia błąd to

$$\sum w_i x_i + w_0 < 0$$

Jak zmienić $\mathbf{w}$ i w_0 aby zniwelować błąd?

- zwiększyć w_0

👉 Training demo



przesunięcie płaszczyzny ze zwiększeniem w_0

Rys: Riedmiller, Machine Learning (lectures)

Uczenie perceptronu - idea

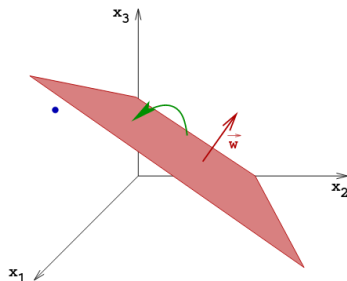
Niech $\mathbf{x}$ należy do zbioru $\mathbb{P}$. Jeżeli perceptron popełnia błąd to

$$\sum w_i x_i + w_0 < 0$$

Jak zmienić $\mathbf{w}$ i w_0 aby zniwelować błąd?

- zwiększyć w_0
- jeśli $x_i > 0$ to zwiększyć w_i
- jeżeli $x_i < 0$ to zmniejszyć w_i

👉 Training demo



obrót płaszczyzny ze zmianą $\mathbf{w}$

Rys: Riedmiller, Machine Learning (lectures)

Algorytm uczenia perceptronu

Algorithm Algorytm uczenia perceptronu

Input: zbiór wektorów należących do jednego z dwóch zbiorów $\mathbb{P}$ i $\mathbb{N}$

Output: perceptron klasyfikujący wszystkie przypadki (jeżeli istnieje)

- 1: zainicjuj wagi $\mathbf{w}$ oraz wartość progową w_0 (np. małe losowe wartości w okolicy 0)
 - 2: **while** istnieje błędnie klasyfikowany $\mathbf{x}$ **do**
 - 3: **if** $\mathbf{x} \in \mathbb{P}$ **then**
 - 4: $\mathbf{w} \leftarrow \mathbf{w} + \mathbf{x}$
 - 5: $w_0 \leftarrow w_0 + 1$
 - 6: **else**
 - 7: $\mathbf{w} \leftarrow \mathbf{w} - \mathbf{x}$
 - 8: $w_0 \leftarrow w_0 - 1$
 - 9: **return** $\mathbf{w}, w_0$
-

Algorytm uczenia perceptronu

- Tw. o zbieżności perceptronu: jeżeli zbiór wektorów $\mathbf{x}_i$ jest liniowo separowalny to algorytm uczenia perceptronu odnajdzie rozwiązanie w skończonej liczbie kroków
- Rozwiązanie nie jest unikatowe
- Możliwe cykle, gdy wektor wag się powtórzy w sekwencji uczenia to problem jest nierozwiązywalny
- Liczba błędów nie maleje monotonicznie - kolejna modyfikacja może popsuć klasyfikację poprzednio nauczonych przypadków
- Jak znaleźć najlepsze możliwe rozwiązanie nawet gdy problem nie jest liniowo separowalny?
- **Algorytm kieszeniowy** - uczenie perceptronu z zapamiętaniem wag dla których popełniono najmniej błędów

Algorithm Algorytm kieszonkowy

Input: zbiór wektorów należących do jednego z dwóch zbiorów $\mathbb{P}$ i $\mathbb{N}$

Output: perceptron klasyfikujący przypadki do dwóch klas

- 1: zainicjuj losowo wagi $\mathbf{w}$ i w_0
 - 2: $t \leftarrow 0$, $t' \leftarrow t$, $\mathbf{w}' \leftarrow \mathbf{w}$, $w'_0 \leftarrow w_0$
 - 3: **for** losowo wybranego $\mathbf{x} \in \mathbb{P} \cup \mathbb{N}$ **do**
 - 4: **if** $\mathbf{x}$ poprawnie klasyfikowany **then**
 - 5: $t \leftarrow t + 1$
 - 6: **else**
 - 7: **if** $t > t'$ **then**
 - 8: $t' \leftarrow t$, $\mathbf{w}' \leftarrow \mathbf{w}$, $w'_0 \leftarrow w_0$
 - 9: $t \leftarrow 0$
 - 10: wykonaj aktualizację wag perceptronu
 - 11: **return** $\mathbf{w}'$, w'_0
-

Algorytm kieszonkowy

- Algorytm kieszonkowy ma za zadanie znaleźć najlepsze możliwe rozwiązanie nawet w przypadku problemów, które nie są liniowo separowalne
- Zapamiętuje tylko wagi ostatniego poprawnego przypadku, więc istnieje możliwość zignorowania wcześniejszego, lepszego rozwiązania. Przeciwdziała temu alg. z zapadką, jednak wymaga on więcej nakładów obliczeniowych.
- **Algorytm kieszeniowy z zapadką** - modyfikacja algorytmu kieszeniowego, gdzie zapamiętywany jest zwycięzca tylko wtedy, gdy klasyfikuje poprawnie więcej przypadków treningowych

Algorithm Algorytm kieszonkowy z zapadką

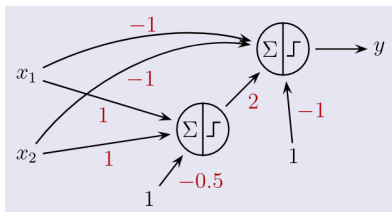
Input: zbiór wektorów należących do jednego z dwóch zbiorów $\mathbb{P}$ i $\mathbb{N}$

Output: perceptron klasyfikujący przypadki do dwóch klas

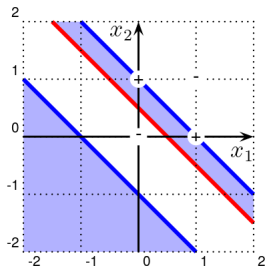
- 1: zainicjuj losowo wagi $\mathbf{w}$ i w_0
- 2: $t \leftarrow 0$, $t' \leftarrow t$, $\mathbf{w}' \leftarrow \mathbf{w}$, $w'_0 \leftarrow w_0$
- 3: **for** losowo wybranego $\mathbf{x} \in \mathbb{P} \cup \mathbb{N}$ **do**
- 4: **if** $\mathbf{x}$ poprawnie klasyfikowany **then**
- 5: $t \leftarrow t + 1$
- 6: **else**
- 7: **if** $t > t'$ **then**
- 8: **if** $\mathbf{w}$ i w_0 klasyfikują poprawnie więcej przypadków niż $\mathbf{w}'$ i w'_0 **then**
- 9: $t' \leftarrow t$, $\mathbf{w}' \leftarrow \mathbf{w}$, $w'_0 \leftarrow w_0$
- 10: $t \leftarrow 0$
- 11: wykonaj aktualizację wag perceptronu
- 12: **return** $\mathbf{w}'$, w'_0

Czego może się nauczyć perceptron?

- Perceptron binarny potrafi rozwiązać wyłącznie problemy liniowo separowalne
- Przykład XOR - nie jest liniowo separowalny
- Sieci połączonych neuronów mają większe możliwości
- Problem XOR można rozwiązać za pomocą 2 perceptronów prostych



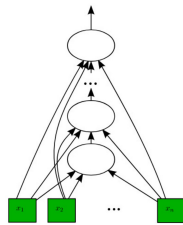
Rys: Riedmiller, Machine Learning (lectures)



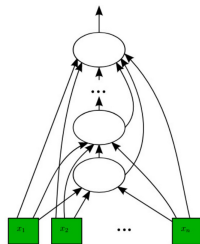
Algorytm wieżowy i piramidalny

Przykład algorytmów rozrastających się

1. Wytrenuj pojedynczy perceptron (np. algorytmem kieszonkowym) na danych treningowych
2. Jeżeli nie uzyskano zadowalającego rezultatu to dodaj perceptron, na którego wejście podany jest sygnał
 - z poprzedniego perceptronu (algorytm **wieżowy**)
 - ze wszystkich poprzednich perceptronów (algorytm **piramidalny**)
3. Wróć do punktu 1.



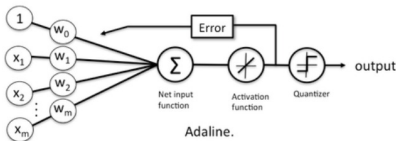
(a) Sieć wieżowa.



(b) Sieć piramidalna.

Adaline - Adaptive Linear Element

Adeline (Widrow, 1959) - jednowarstwowa sieć składająca się z perceptronów prostych. Realizacja sprzętowa z użyciem memistorów.



Reguła uczenia Widrowa-Hoffa

$$\Delta w_k = \lambda \left(y - \sum_i w_i x_i \right) x_k$$

gdzie y jest oczekiwanym sygnałem wyjściowym

Rys: <https://sebastianraschka.com/faq/docs/diff-perceptron-adaline-neuralnet.html>

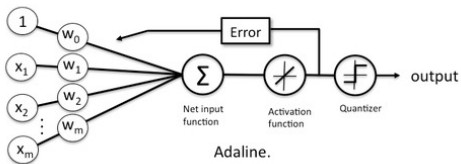
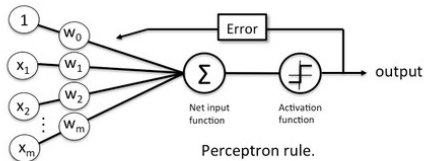
Reguła uczenia

- Reguła uczenia Widrowa-Hoffa dąży do minimalizacji błędu kwadratowego aktywacji neuronu $\mathbf{w}^T \mathbf{x}$ względem pożądanego sygnału wyjściowego y

$$E_{MSE}(\mathbf{x}, y) = \frac{1}{2} \left(y - \sum_i w_i x_i \right)^2$$

- funkcja kosztu E jest ciągła, więc możliwy jest trening metodą spadku gradientu
- dla neuronów liniowych reguła jest równoważna regule delta a Adeline rozwiązuje problem aproksymacji, gdzie $y_i \in \mathbb{R}$
- dla neuronów z wyjściem progowym Adeline staje się klasyfikatorem, gdzie $y_i = \{-1, +1\}$
- Madeline - wielowarstwowa wersja Adeline

Adeline vs. Perceptron rule



Rys: <https://sebastianraschka.com/faq/docs/diff-perceptron-adaline-neuralnet.html>

Reguła delta

- **Reguła delta** - uogólniona reguła uczenia perceptronu dla ciągłych (różniczkowalnych) funkcji aktywacji $f(x)$

$$\Delta w_i = \lambda (y - f(\mathbf{x})) f'(\mathbf{x}) x_i$$

$\lambda > 0$ współczynnik uczenia

y to pożądany sygnał dla wejścia $\mathbf{x}$

$y \in \mathbf{R}$ dla problemu aproksymacji,

$y = \{-1, +1\}$ lub $y = \{0, 1\}$ dla klasyfikacji binarnej

- minimalizacja błędu kwadratowego E w kierunku największego spadku gradientu

$$\Delta w = -\lambda \nabla_w E$$

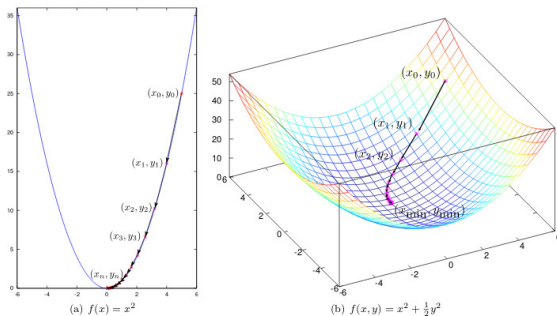
Dla błędu MSE otrzymujemy

$$\Delta w_i = -\lambda \frac{\partial}{\partial w_i} E_{MSE} = -\lambda \frac{\partial}{\partial w_i} \frac{1}{2} (y - f(\mathbf{x}))^2 = \lambda (y - f(\mathbf{x})) f'(\mathbf{x}) x_i$$

Algorytm spadku gradientu

1. ustaw punkt startu $\mathbf{w}_0$ oraz stałą uczenia $\lambda > 0$
2. dopóki nie jest spełnione kryterium stopu wykonuj

$$\mathbf{w}_{i+1} = \mathbf{w}_i - \lambda \nabla E(\mathbf{w})$$



Rys: Maja Czoków, Jarosław Piersa, Tomasz Schreiber, Wstęp do Sieci Neuronowych

Aspekty uczenia spadkiem gradientu

- Jeżeli istnieją minima lokalne to istnieje ryzyko utknięcia algorytmu w tym minimum
- Powtórzenie kilkukrotne uczenia z różnymi punktami startowymi może pozwolić uniknąć minimów lokalnych
- Trajektoria zbieżności zależy od punktu startowego
- Kryterium stopu algorytmu:
 - ilość kroków uczenia lub inne ograniczenie czasowe
 - osiągnięcie zadowalającego poziomu dokładności $E < \epsilon$
 - niewielkie zmiany $||\Delta w|| < \epsilon$
- Dobór stałej uczenia λ . Gdy za duża to rozwiązanie może być pominięte (przeskoczone) a nawet proces może być rozbieżny, gdy za mała to uczenie będzie powolne.
Nie musi być wartością stałą, np. może być zmniejszana w czasie treningu lub dobierana zależnie od wypukłości funkcji

Funkcje aktywacji

- **identyczność**

$$f(x) = x \quad f'(x) = 1$$

- funkcja **liniowa** - nieograniczona $f(x) \in \mathbb{R}$

$$f(x) = ax + b \quad f'(x) = a$$

- **progowa** unipolarna - nieciągła, nieróżniczkowalna

$$f(x) = \begin{cases} 1, & \text{gdy } x \geq a \\ 0, & \text{gdy } x < a \end{cases}$$

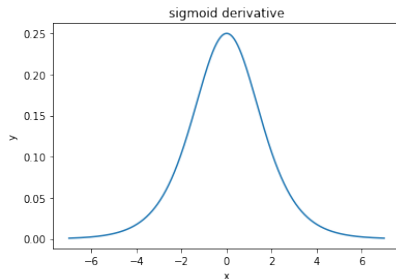
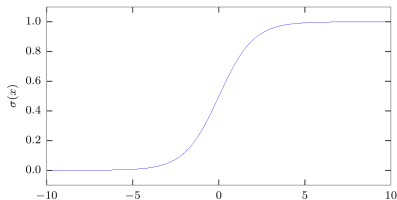
- progowa bipolarna

$$f(x) = \begin{cases} 1, & \text{gdy } x \geq a \\ -1, & \text{gdy } x < a \end{cases}$$

Funkcje aktywacji

- **sigmoidalna** (unipolarna) - ograniczona $f(x) \in (0, 1)$

$$f(x) = \frac{1}{1 + e^{-x}} \quad f'(x) = f(x)(1 - f(x))$$

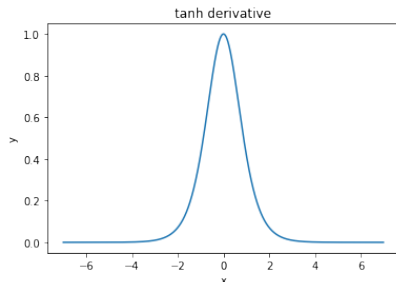
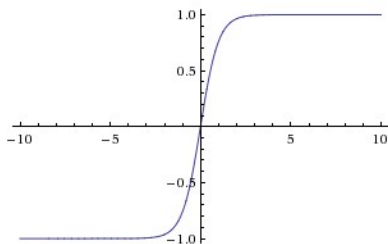


Rys: <http://cs231n.github.io/>

Funkcje aktywacji

- **tangens hiperboliczny** (bipolarna) - ograniczona
 $f(x) \in (-1, +1)$

$$f(x) = \tanh x = \frac{1 - e^{-x}}{1 + e^{-x}} \quad f'(x) = 1 - f^2(x)$$



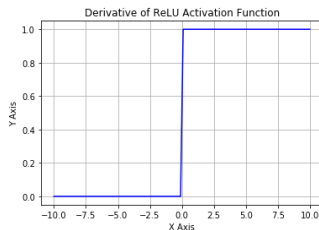
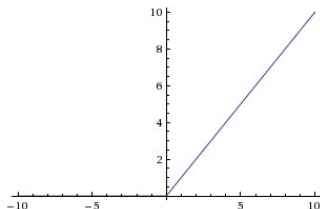
Rys: <http://cs231n.github.io/>,
<https://www.kaggle.com/moriano/a-showcase-of-how-relus-can-speed-up-the-learning>

Funkcje aktywacji

- **ReLU** (Rectified Linear Unit) - fragmentami ciągła, brak pochodnej w 0

$$f(x) = \max(0, x) = \begin{cases} x, & \text{gdy } x \geq 0 \\ 0, & \text{gdy } x < 0 \end{cases}$$

$$f'(x) = \begin{cases} 1, & \text{gdy } x > 0 \\ 0, & \text{gdy } x < 0 \end{cases}$$



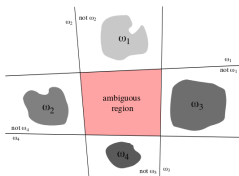
Rys: <http://cs231n.github.io/>,
<https://learnopencv.com/understanding-activation-functions-in-deep-learning/>

Klasyfikacja wielu klas

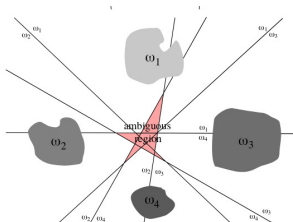
W przypadku wieloklasowym, gdy $y_i = 1, 2, 3, \dots, k$

Model składający się z wielu klasyfikatorów binarnych:

one vs. rest



one vs. one



*error
correction
codes*

#	Code				
0	0	0	0	0	0
1	0	0	1	1	0
2	0	1	0	1	1
3	0	1	1	0	1
4	1	0	0	1	1
5	1	0	1	0	0
6	1	1	0	0	0
7	1	1	1	1	1

8 classes, code-length = 5

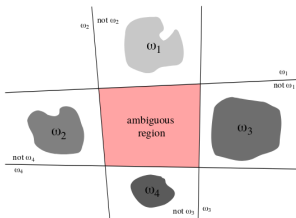
Rys: Duda and Hart, Pattern Recognition
Vivek Srikumar, Structured Prediction, Spring 2021

one vs. rest

Trening: k perceptronów prostych, po jednym na każdą klasę.
Pojedynczy perceptron separuje przypadki z pojedynczej klasy od pozostałych przypadków.

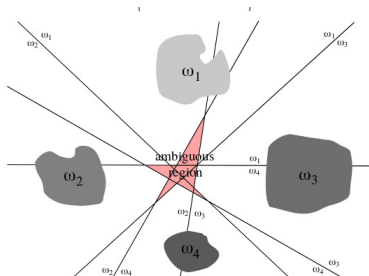
Predykcja: zwycięzca bierze wszystko

$$\arg \max_i \mathbf{w}_i^T \mathbf{x}$$



one vs. one

- Trening:** $k(k-1)/2$ perceptronów dla każdej pary klas
- Predykcja:** głosowanie większościowe, każda klasa otrzymuje $k-1$ głosów. Wybieramy klasę o największej liczbie pozytywnych odpowiedzi.



error correction codes

- każdą etykietę klasy kodujemy (losowym) ciągiem bitowym o długości $n > \log k$
- dodatkowe bity ponad $\log k$ działają jak kody korekcyjne błędy, im większe n tym lepsza korekcja kodów

Trening: trenujemy $n > \log k$ perceptronów, po jednym dla każdego bitu.

Predykcja: wybieramy klasę o najmniejszej odległości Hamminga między odpowiedzią n perceptronów a ciągiem kodującym klasę

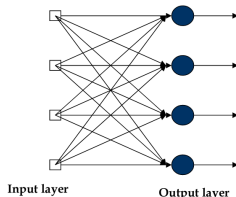
#	Code				
0	0	0	0	0	0
1	0	0	1	1	0
2	0	1	0	1	1
3	0	1	1	0	1
4	1	0	0	1	1
5	1	0	1	0	0
6	1	1	0	0	0
7	1	1	1	1	1

8 classes, code-length = 5

Maszyna liniowa

Jednowarstwowa sieć liniowa klasyfikująca do k grup

- sieć jednowarstwowa - liczba wyjść równa liczbie klas



- funkcja dyskryminująca

$$f_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

- klasyfikacja: przypisanie wektorowi $\mathbf{x}$ klasy i odpowiadającej wyjściu f_i o największej wartości

$$f_i(\mathbf{x}) > f_k(\mathbf{x}) \quad \text{dla każdego} \quad i \neq k$$

Algorithm Uczenie maszyny liniowej

Input: zbiór wektorów należących do jednej z k klas

Output: klasyfikator separujący wektory treningowe

- 1: zainicjuj losowo wagi $\mathbf{w}_i$
 - 2: **for** losowo wybranego $\mathbf{x}$ **do**
 - 3: **if** $\mathbf{x}$ należący do klasy i jest niepoprawnie przypisany do klasy j **then**
 - 4: $\mathbf{w}_i \leftarrow \mathbf{w}_i + \mathbf{x}$
 - 5: $\mathbf{w}_j \leftarrow \mathbf{w}_j - \mathbf{x}$
 - 6: **return** $\mathbf{w}, w_0$
-

Jednokierunkowe sieci
wielowarstwowe

Multilayer Perceptron (MLP)

- Perceptrony jednowarstwowe potrafią rozwiązywać problemy separowalne liniowo
- Reguła delta dla sieci jednowarstwowej:

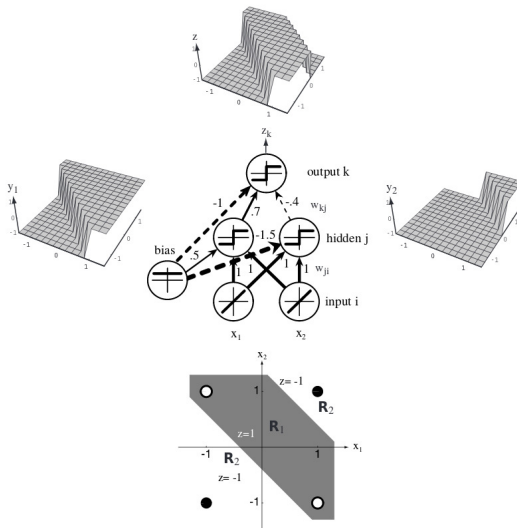
$$\Delta w_{ij} = \eta (y_j - f_j(\mathbf{x})) f_j'(\mathbf{x}) x_i$$

gdzie w_{ij} waga pomiędzy i -tym wejściem a j -tym wyjściem

- Problemy bardziej złożone (np. XOR) wymagają dodania kolejnej warstwy (nieliniowej) przetwarzania
- Każda kolejna warstwa tworzy nowy obraz danych, który może być dalej analizowany przez kolejne warstwy

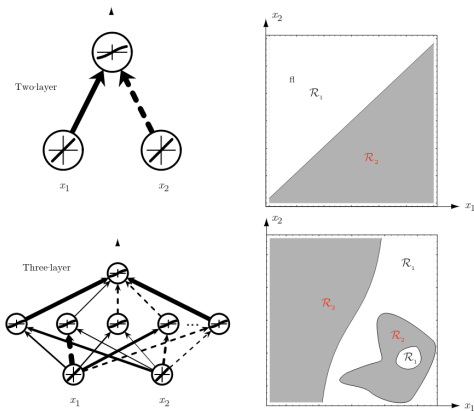
$$\mathbf{y} = f(\mathbf{x}) = f^{(M)} \left(f^{(M-1)} \left(\dots f^{(1)}(\mathbf{x}) \right) \right)$$

Jeszcze raz XOR



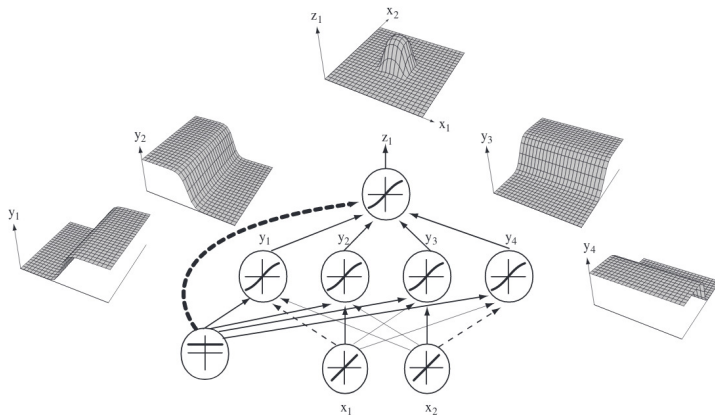
Rys: Duda, Hart, Pattern Classification

Warstwa ukryta a granice decyzji



Rys: Duda, Hart, Pattern Classification

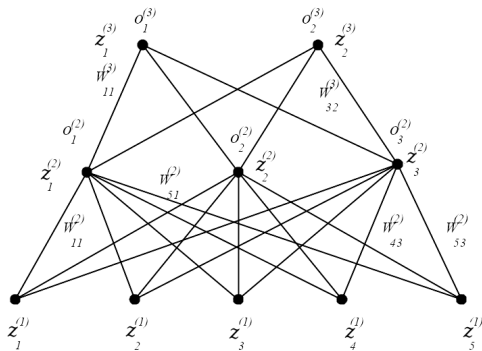
Przykład: sieć MLP 2-4-2



Rys: Duda, Hart, *Pattern Classification*

Perceptron 3 warstwowy

- Warstwy: wejściowa, ukryte, wyjściowa



- Warstwy w pełni połączone (wagi tworzą macierz)

$$\mathbf{o}^{(K)} = \sigma \left(\mathbf{W}^{(K)} \mathbf{o}^{(K-1)} \right)$$

- Waga $w_{ij} = 0$ może być interpretowana jako brak połączenia

MLP - oznaczenia

M - liczba warstw ($M = 3$)

$w_{ij}^{(l)}$ - waga łącząca element i należący do warstwy $l - 1$ oraz element j z warstwy l

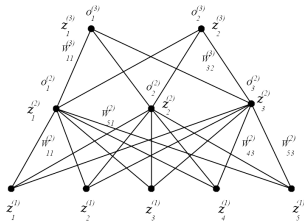
$z_j^{(l)}$ - aktywacja neuronu j w warstwie l

$$z_j^{(l)} = \sum_i w_{ij}^{(l)} o_i^{(l-1)}$$

$o_j^{(l)}$ - sygnał wychodzący z elementu j należącego do warstwy l

$$o_j^{(l)} = \sigma(z_j^{(l)}) = \sigma \left(\sum_i w_{ij}^{(l)} o_i^{(l-1)} \right)$$

$f_i(\mathbf{x}) = o_i^{(M)}$ funkcja realizowana przez MLP (i -te wyjście)



Algorytm wstecznej propagacji błędu

(1974, 1986)

Miara błędu sieci dla pojedynczego wzorca wejściowego $\mathbf{x}$ i pożądanej odpowiedzi $\mathbf{y} = [y_1, \dots, y_n]$

$$E(\mathbf{x}; \mathbf{W}) = \frac{1}{2} \sum_{i=1}^n (y_i - f_i(\mathbf{x}; \mathbf{W}))^2$$

gdzie n to liczba wyjść

Minimalizacja błędu metodą spadku gradientu:

$$\Delta w = -\eta \frac{\partial E(\mathbf{x}; \mathbf{W})}{\partial w} = \eta \sum_{i=1}^n (y_i - f_i(\mathbf{x}; \mathbf{W})) \frac{\partial f_i(\mathbf{x}; \mathbf{W})}{\partial w}$$

Zmiana wag dla warstwy wyjściowej

Dla wagi $w_{jk}^{(M)}$ łączącej wyjście $o_j^{(M-1)}$ z k -tym wyjściem sieci

$$\begin{aligned}\Delta w_{jk}^{(M)} &= \eta \sum_{i=1}^n (y_i - f_i(\mathbf{x}; \mathbf{W})) \frac{\partial f_i(\mathbf{x}; \mathbf{W})}{\partial w_{jk}^{(M)}} \\&= \eta (y_k - f_k(\mathbf{x}; \mathbf{W})) \frac{\partial f_k(\mathbf{x}; \mathbf{W})}{\partial w_{jk}^{(M)}} \\&= \eta (y_k - f_k(\mathbf{x}; \mathbf{W})) \frac{\partial \sigma(z_k^{(M)})}{\partial z_k^{(M)}} \frac{\partial z_k^{(M)}}{\partial w_{jk}^{(M)}} \\&= \eta (y_k - f_k(\mathbf{x}; \mathbf{W})) \sigma'(z_k^{(M)}) o_j^{(M-1)}\end{aligned}$$

$$\Delta w_{jk}^{(M)} = \eta (y_k - f_k(\mathbf{x}; \mathbf{W})) \sigma'(z_k^{(M)}) o_j^{(M-1)}$$

Niech $\delta_i^{(M)}$ oznacza błąd „lokalny” dla warstwy M

$$\delta_k^{(M)} = \sigma'(z_k^{(M)}) (y_k - f_k(\mathbf{x}; \mathbf{W}))$$

Zmiana wag dla warstwy wyjściowej:

$$\Delta w_{jk}^{(M)} = \eta \delta_k^{(M)} o_j^{(M-1)}$$

$$\Delta w_{0k}^{(M)} = \eta \delta_k^{(M)}$$

Zmiana wag dla warstwy ukrytej ($M - 1$)

Dla wagi $w_{jk}^{(M-1)}$ z warstwy ukrytej $M - 1$:

$$\Delta w_{jk}^{(M-1)} = -\eta \frac{\partial E(\mathbf{W})}{\partial w_{jk}^{(M-1)}} = \eta \sum_{i=1}^n (y_i - f_i(\mathbf{x}; \mathbf{W})) \frac{\partial f_i(\mathbf{x}; \mathbf{W})}{\partial w_{jk}^{(M-1)}}$$

gradient

$$\begin{aligned} \frac{\partial f_i(\mathbf{x}; \mathbf{W})}{\partial w_{jk}^{(M-1)}} &= \frac{\partial \sigma(z_i^{(M)})}{\partial z_i^{(M)}} \frac{\partial z_i^{(M)}}{\partial w_{jk}^{(M-1)}} = \sigma' \left(z_i^{(M)} \right) \frac{\partial \left(\sum_l w_{li}^{(M)} o_l^{(M-1)} \right)}{\partial w_{jk}^{(M-1)}} \\ &= \sigma' \left(z_i^{(M)} \right) w_{ki}^{(M)} \frac{\partial o_k^{(M-1)}}{\partial w_{jk}^{(M-1)}} \\ &= \sigma' \left(z_i^{(M)} \right) w_{ki}^{(M)} \sigma' \left(z_k^{(M-1)} \right) o_j^{(M-2)} \end{aligned}$$

$$\frac{\partial f_i(\mathbf{x}; \mathbf{W})}{\partial w_{jk}^{(M-1)}} = \sigma' \left(z_i^{(M)} \right) w_{ki}^{(M)} \sigma' \left(z_k^{(M-1)} \right) o_j^{(M-2)}$$

zmiana wag dla warstwy $M - 1$

$$\begin{aligned} \Delta w_{jk}^{(M-1)} &= \eta \sum_{i=1}^n (y_i - f_i(\mathbf{x}; \mathbf{W})) \frac{\partial f_i(\mathbf{x}; \mathbf{W})}{\partial w_{jk}^{(M-1)}} \\ &= \eta \sum_{i=1}^n (y_i - f_i(\mathbf{x}; \mathbf{W})) \sigma' \left(z_i^{(M)} \right) w_{ki}^{(M)} \sigma' \left(z_k^{(M-1)} \right) o_j^{(M-2)} \\ &= \eta \sigma' \left(z_k^{(M-1)} \right) \sum_{i=1}^n w_{ki}^{(M)} \delta_i^{(M)} o_j^{(M-2)} \\ &= \eta \delta_k^{(M-1)} o_j^{(M-2)} \end{aligned}$$

$$\Delta w_{0k}^{(M-1)} = \eta \delta_k^{(M-1)} \quad \text{gdzie} \quad \delta_k^{(M-1)} = \sigma' \left(z_k^{(M-1)} \right) \sum_{i=1}^n w_{ki}^{(M)} \delta_i^{(M)}$$

Struktura wzoru w kolejnych warstwach jest taka sama

Wsteczna propagacja błędu - podsumowanie

Funkcja realizowana przez sieć wielowarstwową

$$f_i(\mathbf{x}; \mathbf{W}) = \sigma \left(\sum_j w_{ji}^{(M)} \sigma \left(\sum_k w_{kj}^{(M-1)} \sigma \left(\sum_l w_{lk}^{(M-2)} \dots \sigma \left(\sum_n w_{nm}^{(2)} x_n \right) \right) \right) \dots \right)$$

Aktualizacja wag w dowolnej warstwie K - **uogólniona reguła delta**

$$\Delta w_{ij}^{(K)} = \eta \delta_j^{(K)} o_i^{(K-1)}$$

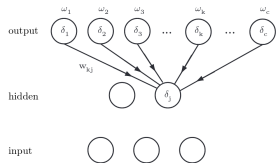
gdzie sygnał wejściowy $o_i^{(1)} = x_i$

Sygnał błędu w warstwie K

$$\delta_i^{(K)} = \sigma' \left(z_i^{(K)} \right) \sum_j \delta_j^{(K+1)} w_{ij}^{(K+1)}$$

sygnał błędu w warstwie wyjściowej:

$$\delta_i^{(M)} = \sigma' \left(z_i^{(M)} \right) (y_i - f_i(\mathbf{x}; \mathbf{W}))$$



Algorytm wstecznej propagacji

1. Zainicjuj wagi **W** małymi losowymi wartościami
2. Dopóki nie spełniony warunek stopu wykonuj
 - 2.1 dla pary uczącej (**x**, **y**) oblicz sygnał od wejścia do wyjścia

$$o_i^{(K)} = \sigma \left(\sum_j w_{ji}^{(K)} o_j^{(K-1)} \right)$$

- 2.2 oblicz sygnał błędu od warstwy wyjściowej

$$\delta_i^{(M)} = \sigma'(z_i^{(M)}) (y_i - f_i(\mathbf{x}))$$

- 2.3 oblicz sygnał błędu od warstwy wyjściowej do wejściowej

$$\delta_i^{(K)} = \sigma'(z_i^{(K)}) \sum_j \delta_j^{(K+1)} w_{ij}^{(K+1)}$$

- 2.4 aktualizuj wagi

$$\Delta w_{ij}^{(K)} = \eta \delta_j^{(K)} o_i^{(K-1)}$$

Algorytm wstecznej propagacji można w łatwy sposób rozszerzyć na:

- sieci o innej strukturze połączeń: np. dodatkowe połączenia pomiędzy wejściem lub wyjściem, lub innymi warstwami ukrytymi. Dla sieci rekurencyjnych odpowiednikiem jest BPTT (wsteczna propagacja w czasie)
- dowolną liczbę warstw
- różne nieliniowości w warstwach
- różne funkcje aktywacji dla poszczególnych neuronów
- różne stałe uczenia dla każdej warstwy lub węzła
- inne funkcje kosztu, nie tylko MSE, np. Cross Entropy

Warianty treningu

Dla zbioru zawierającego N przypadków i sieci zawierającej M wyjść

$$\hat{E} = \frac{1}{N} \sum_{i=1}^N E(\mathbf{x}_i) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M (y_{ij} - f_j(\mathbf{x}_i))^2$$

$$\Delta \hat{w} = \sum_{i=1}^N \Delta w_i$$

Batch gradient descent

- uśredniony błąd całego zbioru treningowego
- niepraktyczne dla dużych danych
- stabilny trening
- nie zależy od kolejności prezentacji wzorców

Warianty treningu

SGD - stochastic gradient descent (on-line)

- aktualizacja dla pojedynczego przypadku treningowego ($N = 1$)
- duża szybkość działania ale i duża wariancja,
- istotna jest kolejność prezentacji przypadków

MiniBatch gradient descent

- średnia z n losowych przypadków (*mini-batch*), trening on-line,
- często utożsamiany z SGD
- mniejsza wariancja niż SGD
- możliwe operacje na macierzach, łatwiejsze zrównoleglenie GPU/CPU

Własności MLP

- **Uniwersalny aproksymator:** sieć MLP z jedną warstwą ukrytą jest w stanie aproksymować dowolną funkcję ciągłą z dowolną dokładnością.
Dwie warstwy ukryte rozszerzają możliwości na funkcje nieciągłe.
Klasyfikator potrafi zrealizować dowolne granice decyzji (obszary nie muszą być wypukłe ani połączone)
- Neurony ukryte: transformacja nieliniowa do przestrzeni odwzorowań, tworząca nowe cechy za pomocą nieliniowych kombinacji
- Wiele zastosowań: klasyfikacja, wielowymiarowa regresja

Problemy I

- Optymalizacja nieliniowych funkcji zawsze sprawia problemy, tu mamy złożenie (czasem wielu) funkcji nieliniowych
- Dobór architektury sieci: Ile węzłów w warstwie ? Jakie funkcje aktywacji? Sieci ontogeniczne (rozzrastające się) dostosowujące rozmiar do złożoności problemu
- Przeuczenie i generalizacja - zbyt duża liczba optymalizowanych parametrów powoduje przeuczenie, zbyt mała - może generować zbyt proste rozwiązania
- Regularyzacja - metody ograniczenia zjawiska przeuczenia, np. modyfikacje funkcji kosztu, ograniczenie liczby parametrów
- Inicjalizacja parametrów - szybkość treningu i wynik zależy od punktu startowego
- Minima lokalne i plateau, wąskie „rynny” - np. wielokrotny start
- Wpływ nowych wzorców na już nauczone – zapominanie

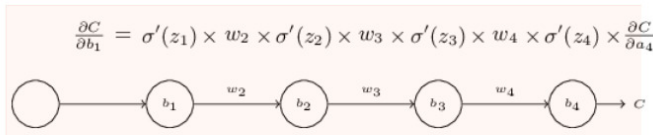
Problemy II

- Dobór stałej uczenia
- Znikający gradient, eksplodujący gradient
- Przygotowanie danych uczących: normalizacja, standaryzacja, kodowanie wyjść, ...
- Ocena modelu: zbiory walidacyjne, testowe, krosvalidacja

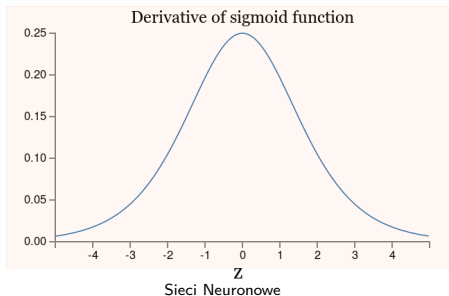
Znikający gradient

Szybkość uczenia (zmiany wag) drastycznie maleje w głębszych warstwach - sygnał błędu zanika

Przykład: sieć $1 \times 1 \times 1 \times 1$



gdy $|w_j| < 1$ wówczas $|w_j \sigma'(z_j)| < \frac{1}{4}$



Problem niestabilnego gradientu

- gradient w niższych warstwach zależy od wartości wag, aktywacji i gradientów w warstwach wyższych
- **eksplodujący gradient**
gdy $w_i \gg 1$ oraz $z_i \approx 0$ wówczas $|w_i \sigma'(z_i)| > 1$, może to owocować bardzo dużymi gradientami w początkowych warstwach
- znikający i eksplodujący gradient to przypadki szczególne **problemu niestabilnego gradientu** -> wartości gradientów w poszczególnych warstwach mogą znacznie się różnić, warstwy uczą się z różnym tempem

Specyfikacja sieci i metody treningu

- Specyfikacja modelu:
 - specyfikacja architektury
 - typ neuronów ukrytych: liniowe, sigmoidalne, ReLU, ...
 - typ neuronów wyjściowych: liniowe, sigmoidalne, softmax, ...
- Funkcja kosztu:
 - mean squared error (MSE),
 - cross-entropy (CE),
- Optymalizacja metodami spadku gradientu lub innymi:
 - SGD, RPROP, Quackprop,
 - Adam, AdaGrad, ...
- Metody regularyzacji:
 - momentum, L_1 , L_2 (weight decay), ...

Funkcja kosztu MSE

Błąd średniokwadratowy (*Mean Squared Error*)

$$J(\Theta) = \frac{1}{N} \sum (\hat{y} - y)^2$$

Gradient funkcji kosztu:

$$\frac{\partial J}{\partial \Theta} = \frac{2}{N} \sum (\hat{y} - y) \frac{\partial \hat{y}}{\partial \Theta}$$

gdzie

$\hat{y}$ - wyjścia sieci,

Θ - parametry sieci,

y - oczekiwane wyjścia

Funkcja kosztu Cross Entropy

Cross entropy

$$J(\Theta) = - \sum y \log \hat{y}$$

- miara odległości między rozkładami y i $\hat{y} \in (0, 1)$
- gdy $y \approx \hat{y}$ to funkcja kosztu bliska zeru
- $J > 0$

Gradient funkcji kosztu:

$$\frac{\partial J}{\partial \Theta} = - \sum y \frac{1}{\hat{y}} \frac{\partial \hat{y}}{\partial \Theta}$$

👉 Wizualizacja funkcji CE

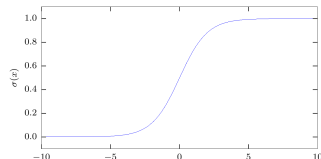
Funkcja liniowa w warstwie wyjściowej

$$z = \mathbf{w}^T \mathbf{h} + b$$

- nieograniczone wartości $(-\infty, +\infty)$
- jednostki nie nasycają się ale bardzo duże wartości aktywacji również mogą przysporzyć problemów w trakcie uczenia
- stabilniejsze w optymalizacji metodami gradientowymi od innych typowych funkcji nieliniowych stosowanych w sieciach
- zastosowanie wraz funkcją kosztu MSE do problemów regresji (dane wyjściowe ciągłe), modelowanie rozkładów gusowskich

Funkcja logistyczna w warstwie wyjściowej

$$\sigma(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$



- wartości ograniczone $(0, 1)$
- nasycają się, aktywne uczenie tylko w okolicach $x \approx 0$
- zastosowanie: klasyfikacja 2 klas, modelowanie rozkładu Bernoulliego, binarne wartości wyjściowe (*multi-label classification*)
- **binarna funkcja kosztu Cross-Entropy**, na każdym wyjściu oczekujemy stanu 0 lub 1

$$J_{CE} = - \sum [y \ln \hat{y} + (1 - y) \ln(1 - \hat{y})]$$

Funkcja sigmoidalna w warstwie wyjściowej

Podczas minimalizacji MSE błąd zanika, gdy funkcja σ się nasycza (także dla niepoprawnej odpowiedzi, gdy błąd $|\hat{y} - y|$ jest duży)

$$\frac{\partial \hat{y}}{\partial \Theta} = \frac{\partial \sigma(z)}{\partial \Theta} = \sigma'(z) \frac{\partial z}{\partial \Theta} \approx 0$$

$$\frac{\partial J_{MSE}}{\partial \Theta} = \frac{2}{N} \sum (\hat{y} - y) \frac{\partial \hat{y}}{\partial \Theta}$$

Problem zanika przy zastosowaniu kosztu Cross Entropy (CE)

$$\frac{\partial J_{CE}}{\partial \Theta} = - \sum \left(\frac{y}{\hat{y}} \frac{\partial \hat{y}}{\partial \Theta} + \frac{1-y}{1-\hat{y}} \frac{\partial (1-\hat{y})}{\partial \Theta} \right) = \sum (\hat{y} - y) \frac{\partial z}{\partial \Theta}$$

gdzie z jest liniową aktywacją (*logit*) neuronu wyjściowego
 $\hat{y} = \sigma(z)$, pochodna f. sigmoidalnej $\sigma'(z) = \sigma(z)(1 - \sigma(z))$

Softmax w warstwie wyjściowej

$$\text{softmax}(\mathbf{z})_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

- wartości $(0, 1)$ unormowane

$$\sum_i \text{softmax}(\mathbf{z})_i = 1$$

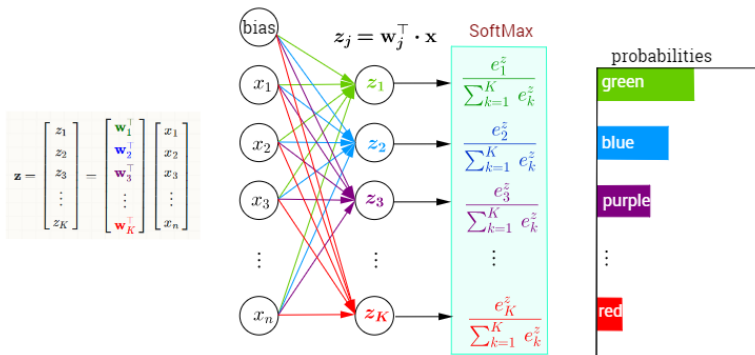
- Zastosowanie: klasyfikacja n rozłącznych klas (*multi-class classification*),
każde wyjście sieci związane jest z jedną klasą, wartości kodowane w postaci wektora binarnego (kodowanie *one hot*)

$$[0, \dots, 0, 1, 0, \dots, 0]$$

- wartości wyjściowe sieci reprezentują rozkład prawdopodobieństw zmiennej wyjściowej

$$\hat{y}_i = P(y = i | X), \quad i = 1, \dots, n$$

Multi-Class Classification with NN and SoftMax Function



Źródło: stats.stackexchange.com/questions/265905/derivative-of-softmax-with-respect-to-weights

- gradient

$$\frac{\partial \text{softmax}(\mathbf{z})_i}{\partial z_j} = \text{softmax}(\mathbf{z})_i (\delta_{ij} - \text{softmax}(\mathbf{z})_i)$$

- gradient funkcji kosztu CE gdy $\hat{y}_i = \text{softmax}(\mathbf{z})_i$

$$\frac{\partial J_{CE}}{\partial \Theta} = \sum (\hat{y} - y) \frac{\partial z}{\partial \Theta}$$

- obliczenia się upraszczają i znika problem zanikania błędu dla nasyconych wyjść, który występował dla funkcji kosztu MSE i wyjść sigmoidalnych
- softmax „dobrze współgra” z funkcją kosztu entropii krzyżowej CE, logarytm niweluje exp, upraszczając obliczenia, błąd nie znika nawet dla dużych wartości aktywacji z_i

$$\log \text{softmax}(\mathbf{z})_i = z_i - \log \sum_k e^{z_k}$$

- ważne są różnice wartości pomiędzy wyjściami a nie amplitudy poszczególnych wyjść

$$\text{softmax}(x) = \text{softmax}(x + c)$$

- minimalizacja kosztu CE powoduje zwiększenie wartości na jednym z wyjść z_i i zmniejszenie wszystkich pozostałych $-\log \sum e^{z_k}$
- ponieważ

$$\log \sum_k e^{z_k} \approx \max_i z_i$$

to dla poprawnej odpowiedzi $z_i \approx \max z_i$, błąd znika. Karane są najbardziej aktywne niepoprawne predykcje

- z punktu widzenia neurobiologicznego softmax może być widziany jako realizacja mechanizmów hamowania występującego w grupach neuronów w korze, podobnie do mechanizmu zwycięzca bierze wszystko (*winner takes all*)

Dobór funkcji kosztu uzależniony jest od rozkładu wartości wyjściowych i typu neuronów wyjściowych

Output Type	Output Distribution	Output Layer	Cost Function
Binary	Bernoulli	Sigmoid	Binary cross-entropy
Discrete	Multinoulli	Softmax	Discrete cross-entropy
Continuous	Gaussian	Linear	Gaussian cross-entropy (MSE)
Continuous	Mixture of Gaussian	Mixture Density	Cross-entropy
Continuous	Arbitrary	See part III: GAN, VAE, FVBN	Various

Dobór neuronów ukrytych

- wybór funkcji realizowanych przez warstwy ukryte jest kluczowy dla przebiegu procesu uczenia
- funkcje liniowe - wiele warstw liniowych sprowadza się do pojedynczej transformacji liniowej

$$f(\mathbf{x}) = \mathbf{W}_n \mathbf{W}_{n-1} \dots \mathbf{W}_1 \mathbf{W}_1 \mathbf{x} = \mathbf{A} \mathbf{x}$$

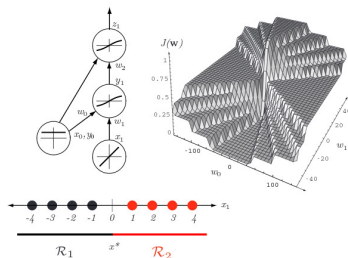
- funkcje ograniczone: sigmoidalne i tangens hiperboliczny powodują problemy z niestabilnym gradientem, uczenie może utknąć, gdy funkcja się nasyci
- obecnie najpopularniejszym typem aktywacji w sieciach jest ReLU

Demonstracje

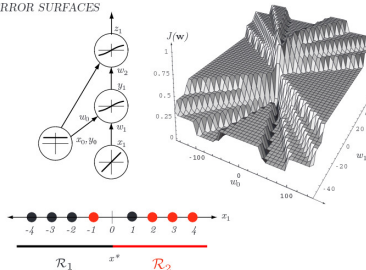
-  TensorFlow playground
-  ConvNetJS - Deep Learning in your browser

Ulepszenia treningu sieci MLP

Powierzchnia błędu



ERROR SURFACES



Minima lokalne i minimum globalne

Plateau - regiony o małej zmienności błędu względem wag

Jak omijać minima lokalne?

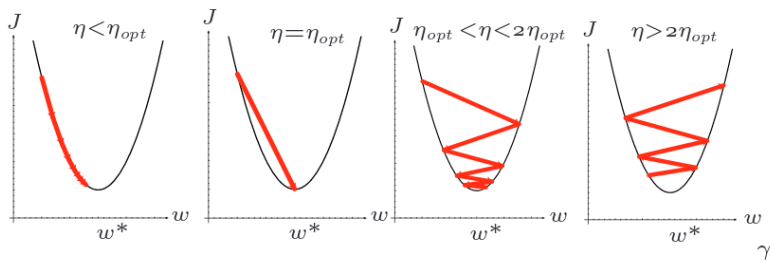
- Wielokrotny start z różnymi wartościami początkowymi - najprostsza ale skuteczna metoda
- Szum dodawany do wag lub szum dodany do danych pozwala wygładzić funkcję błędu i uciec z płytszych minimów – formalnie jest to równoważne regularyzacji, czyli dodaniu dodatkowego członu wygładzającego do funkcji błędu
- Losowa kolejność prezentowania przypadków
- Modyfikacje BP lub inne algorytmy optymalizacji

Metody globalnej minimalizacji

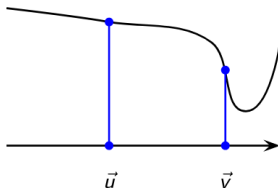
- Metody globalnej minimalizacji: wiele metod.
- Monte Carlo, symulowane wyżarzanie, metody multisympleksowe, minimalizacja Tabu, homotopia ...
- Dużo prac łączących algorytmy genetyczne z sieciami MLP
- Zalety: globalne, proste w realizacji, niektóre nie potrzebują gradientu, inne łączą zalety metod gradientowych z metodami globalnymi
- Wady: zwykle kosztowne i czasochłonne

Dobór kroku uczenia

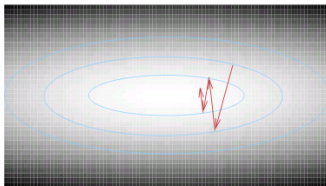
- zbyt mała wartość η - powolna zbieżność
- za duża wartość η - niestabilny trening, oscylacje
- wartość kroku może być zmieniana w czasie uczenia, różne podejścia:
 - zmniejszana w czasie treningu
 - zwiększana dla płaskich powierzchni błęd
 - dobierana niezależnie do warstwy lub pojedynczych neuronów i wag



Płaskie powierzchnie i strome doliny



Powierzchnia błędów w wielu wymiarach ma różne nachylenie



Rys: Riedmiller, Machine Learning

Trening z momentem

prędkość uczenia zależna od poprzednich wartości gradientów

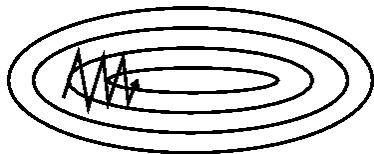
$$\Delta w(t) = -\eta \nabla E(t) + \gamma \Delta w(t-1)$$

- zwiększa krok uczenia, gdy gradient nie zmienia kierunku. Przyspieszenie na płaskich odcinkach (gdzie ∇E stałe) wynosi w przybliżeniu

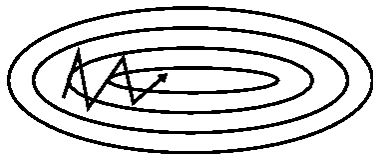
$$\Delta w(t) = -\frac{\eta}{1-\gamma} \nabla E$$

- redukuje oscylacje (spowalnia uczenie), gdy gradient zmienia kierunek
- gdy $\nabla E = 0$ wagi są nadal modyfikowane (bezwładność), może to ułatwić opuszczenie strefy „przyciągania” do minimum lokalnego
- $0 < \gamma < 1$, typowo współczynnik $\gamma = 0.9$

Trening z momentem



SGD bez momentu



SGD z momentem

Inicjalizacja wag

- Losowe wartości z rozkładu jednostajnego w okolicach 0
- Za duże wagi powodują duże wartości aktywacji i ryzyko nasycenia funkcji sigmoidalnych
- Dla d wejść można wybrać wartości z zakresu

$$-\frac{1}{\sqrt{d}} < w_{ij} < \frac{1}{\sqrt{d}}$$

co przy standaryzacji danych daje średnio aktywację neuronów w zakresie liniowym $[-1, 1]$ sigmoidy

- Analogicznie dla kolejnych warstw

Skalowanie wartości wejściowych

- Różne skale mierzonych cech x_i , „dzikie” rozkłady dalekie od rozkładu normalnego, wartości odstające
- Dane wejściowe $\mathbf{x}_i$ powinny posiadać zbliżone zakresy wartości
- Standaryzacja

$$x_s = \frac{x - \mu}{\sigma}$$

μ - wartość średnia, σ - odchylenie standardowe

- Normalizacja w zakresie $[-1, +1]$

$$x_n = 2 \frac{x - x_{min}}{x_{max} - x_{min}} - 1$$

Resilient BP (Riedmiller, Braun, 1992)

- radzi sobie z problemem znikających (oraz za dużych) gradientów
- wyłącznie znak gradientu (nie amplituda) uwzględniany w obliczeniach
- stała uczenia dobierana dla każdej wagi niezależnie

$$\Delta w_{ij}(t) = -\eta_{ij}(t) \operatorname{sgn} \left(\frac{\partial E(\mathbf{w}(t))}{\partial w_{ij}} \right)$$

RPROP

- wymaga informacji z 2 ostatnich kroków uczenia
- η_{ij} rośnie, gdy brak zmiany kierunku gradientu
- η_{ij} maleje, gdy następuje zmiana kierunku

$$\Delta w_{ij}(t) = -\eta_{ij}(t) \operatorname{sgn} \left(\frac{\partial E(\mathbf{w}(t))}{\partial w_{ij}} \right)$$

gdzie

$$\eta_{ij}(t) = \begin{cases} \min(a \cdot \eta_{ij}(t-1), \eta_{\max}) & \text{dla } \frac{\partial E(\mathbf{w}(t))}{\partial w_{ij}} \frac{\partial E(\mathbf{w}(t-1))}{\partial w_{ij}} > 0 \\ \max(b \cdot \eta_{ij}(t-1), \eta_{\min}) & \text{dla } \frac{\partial E(\mathbf{w}(t))}{\partial w_{ij}} \frac{\partial E(\mathbf{w}(t-1))}{\partial w_{ij}} < 0 \\ \eta_{ij}(t-1) & \text{w pozostałych przypadkach} \end{cases}$$

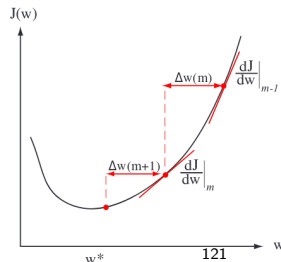
$$a = 1.2, \quad b = 0.5, \quad \eta_{\min} = 10^{-6}, \quad \eta_{\max} = 50$$

Quickprop (Fahlman, 1988)

- założenie: wagi są niezależne a funkcja błędu w okolicach minimum może być przybliżona parabolą
- przybliżenie za pomocą wartości i gradientów funkcji w 2 punktach $w(m)$ i $w(m-1)$

$$\Delta w(m+1) = \frac{\nabla_{ij} E(m)}{\nabla_{ij} E(m-1) - \nabla_{ij} E(m)} \Delta w(m)$$

- wagi mają niezależną zbieżność uczenia
- zbieżność kwadratowa
- trening może być niestabilny



Metody 2 rzędu

Rozwinięcie w szereg Taylora:

$$E(\mathbf{w} + \Delta\mathbf{w}) \approx E(\mathbf{w}) + \nabla E(\mathbf{w})^T \Delta\mathbf{w} + \frac{1}{2} \Delta\mathbf{w}^T \nabla^2 E(\mathbf{w}) \Delta\mathbf{w}$$

gdzie $\nabla^2 E(\mathbf{w}) = \mathbf{H}$ jest macierzą $n \times n$ pochodnych 2 rzędu (Hessian)

$$H_{ij} = \frac{\partial^2 E(\mathbf{x}; \mathbf{w})}{\partial w_i \partial w_j}$$

Gradient funkcji kosztu:

$$\nabla E(\mathbf{w} + \Delta\mathbf{w})^T \approx \nabla E(\mathbf{w})^T + \Delta\mathbf{w}^T \mathbf{H}$$

Minimum osiągnęte dla $\nabla E(\mathbf{w} + \Delta\mathbf{w}) = 0$, stąd

$$\Delta\mathbf{w} = -\mathbf{H}^{-1} \nabla E(\mathbf{w})$$

rozwiązanie w jednym kroku, jeżeli potrafimy obliczyć $\mathbf{H}^{-1}$

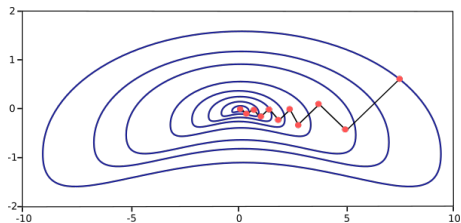
Metoda Newtona

Iteracyjna metoda 2 rzędu:

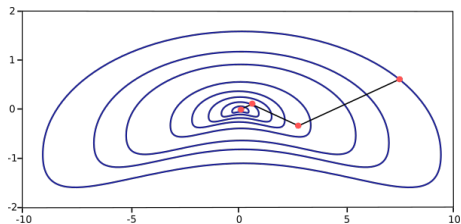
$$\mathbf{w}(t+1) = \mathbf{w}(t) - \mathbf{H}^{-1} \nabla E(\mathbf{w}(t))$$

- metoda kosztowna czasowo $O(n^3)$ (odwracanie macierzy w każdej iteracji)
- zbieżność (kwadratowa) po kilku iteracjach
- metoda kosztowna pamięciowo, Hessian $O(n^2)$, gdzie n liczba wag
- praktyczne zastosowania tylko dla małych sieci
- $\mathbf{H}$ nie zawsze jest dodatnio określona - stosuje się aproksymacje

Spadek gradientu



Metoda Newtona



Rys: https://www.neuraldesigner.com/blog/5_algorithms_to_train_a_neural_network

Metody quasi-Newtona

Przybliżenia do Hesjanu

- zaniechanie pozadiagonalnych elementów

$$\Delta w_i = -\frac{\partial E}{\partial w_i} \bigg/ \frac{\partial^2 E}{\partial w_i^2}$$

- metoda zmiennej metryki - przybliżenie do H^{-1} oraz iteracyjna metoda Newtona, kwadratowo zbieżna
 - Davidon-Fletcher-Power (DFP)
 - Broyden-Fletcher-Goldfarb-Shanno (BFGS).
- Metoda Levenberg-Marquardta oparta jest na przybliżeniu Gaussa-Newtona

Metoda Levenberg-Marquardta

Metoda LM znajduje rozwiązanie zadań optymalizacji, które daje się sprowadzić do postaci

$$E = \sum_{i=1}^p e_i^2$$

W przypadku danych zawierających n przypadków i sieci o m wyjściach sumowanie obejmuje $n \cdot m$ wartości dla funkcji błędu MSE:

$$E = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m (f_j(\mathbf{x}_i) - y_{ij})^2$$

Niech $\mathbf{e} = [e_1, e_2, \dots, e_p]$ tworzy **wektor rezydualny** zaś $\mathbf{J}$ macierz **Jakobianu**

$$J_{ij} = \frac{\partial e_i}{\partial w_j}, \quad i = 1, \dots, p \quad j = 1, \dots, k$$

gdzie k to ilość wszystkich wag.

Metoda Levenberg-Marquardta

Gradient funkcji błędu można wyrazić w postaci:

$$\nabla E = 2\mathbf{J}^T \mathbf{e}$$

Metoda LM zastępuje macierz Hessego w metodzie Newtona przybliżeniem opartym na wartości gradientu wraz z czynnikiem regularyzacyjnym μ

$$\mathbf{H} \approx 2\mathbf{J}^T \mathbf{J} + \mu \mathbf{I}$$

Współczynnik tłumienia μ zapewnia dodatnią określoność $\mathbf{H}$.

Reguła aktualizacji wag:

$$\Delta \mathbf{w} = -2 (\mathbf{J}^T \mathbf{J} + \mu \mathbf{I})^{-1} \mathbf{J}^T \mathbf{e}$$

Metoda Levenberg-Marquardta

$$\Delta \mathbf{w} = -2 (\mathbf{J}^T \mathbf{J} + \mu \mathbf{I})^{-1} \mathbf{J}^T \mathbf{e}$$

- dla dużych μ czynnik kwadratowy zanika i uzyskujemy metodę największego spadku
- dla $\mu = 0$ mamy metodę Newtona o kwadratowej zbieżności
- LM startuje z dużą wartością μ a następnie w trakcie uczenia w miarę zbliżania się do rozwiązania μ jest zmniejszane
- bardzo szybko zbieżna metoda dla funkcji, które są sumą kwadratów (MSE)
- nie nadaje się do zastosowań z funkcją Cross Entropy
- nie praktyczne dla dużych danych, duży koszt pamięci, Jakobian ma wymiar $p \times k$, gdzie $p = n \cdot m$ - ilość wektorów razy ilość wyjść, k - ilość parametrów (wag)

Metoda Levenberg-Marquardta

Algorytm doboru współczynnika μ w i -tym kroku, dla danego $\mathbf{w}(i)$:

1. oblicz wartość $\mathbf{w}(i+1)$ zgodnie z regułą uczenia LM
2. oblicz wartość błędu w punkcie $\mathbf{w}(i+1)$
3. jeśli błąd wzrósł, wróć do poprzedniej wartości $\mathbf{w}(i)$ i zwiększ wartość μ k -krotnie i wróć do kroku 1 (przybliżenie liniowe minimalizowanej funkcji w otoczeniu $\mathbf{w}(i)$ okazało się nie dość ścisłe, więc zwiększamy „wpływ” metody największego spadku)
4. jeśli błąd zmalał, zaakceptuj ten krok i zmniejsz wartość μ k -krotnie (założenie o liniowości minimalizowanej funkcji w otoczeniu $\mathbf{w}(i)$ okazało się wystarczająco ścisłe, więc zwiększamy „wpływ” metody Gaussa-Newtona).

Typowo $k = 10$.

Metoda gradientów sprzężonych

Metoda **gradientów sprzężonych** (*conjugated gradients*) poszukuje minimum wzdłuż kierunku $\Delta \mathbf{w}(m)$ ortogonalnego i sprzężonego do wszystkich kierunków z poprzednich iteracji.

Kierunki $\Delta \mathbf{w}(m)$ i $\Delta \mathbf{w}(m-1)$ są sprzężone gdy:

$$\Delta \mathbf{w}^T(m-1) \mathbf{H} \Delta \mathbf{w}(m) = 0$$

Kierunek wyszukiwania minimum w iteracji m wyznaczamy z reguły

$$\Delta \mathbf{w}(m) = -\nabla E(m) + \beta(m) \Delta \mathbf{w}(m-1)$$

gdzie **współczynnik sprzężenia** β kumuluje informacje o kierunkach z poprzednich iteracji.

Metoda gradientów sprzężonych

Przykładowe postaci współczynnika sprzężenia:

- reguła Fletchera-Reevesa

$$\beta(m) = \frac{\nabla E(m)^T \nabla E(m)}{\nabla E(m-1)^T \nabla E(m-1)}$$

- reguła Polaka-Ribiera

$$\beta(m) = \frac{\nabla E(m)^T (\nabla E(m) - \nabla E(m-1))}{\nabla E(m-1)^T \nabla E(m-1)}$$

- reguła Hestenesa-Stiefela

$$\beta(m) = \frac{\nabla E(m)^T (\nabla E(m) - \nabla E(m-1))}{-\nabla E(m-1)^T (\nabla E(m) - \nabla E(m-1))}$$

- reguła Dai-Yuan

$$\beta(m) = \frac{\nabla E(m)^T \nabla E(m)}{\nabla E(m-1)^T \nabla E(m-1)}$$

Algorytm metody gradientów sprzężonych

Pierwszy kierunek

1. zainicjuj $\mathbf{w}(0)$ i znajdź pierwszy kierunek $\Delta\mathbf{w}(0)$ metodą największego spadku $\Delta\mathbf{w}(0) = -\nabla E(0)$
2. wykonaj liniowe przeszukiwanie
 $\alpha = \arg \min_{\alpha} E(\mathbf{w}(m) + \alpha\Delta\mathbf{w}(m))$
3. zaktualizuj wagi $\mathbf{w}(1) = \mathbf{w}(0) + \alpha\Delta\mathbf{w}(0)$

Kolejne kierunki

1. wyznacz kierunek $\Delta\mathbf{w}(m) = -\nabla E(m)$ metodą największego spadku
2. wyznacz współczynnik sprzężenia $\beta(m)$
3. zaktualizuj kierunek sprzężony
 $\Delta\mathbf{w}(m) = -\nabla E(m) + \beta(m)\Delta\mathbf{w}(m-1)$
4. wykonaj liniowe przeszukiwanie
 $\alpha = \arg \min_{\alpha} E(\mathbf{w}(m) + \alpha\Delta\mathbf{w}(m))$
5. zaktualizuj wagi $\mathbf{w}(m+1) = \mathbf{w}(m) + \alpha\Delta\mathbf{w}(m)$

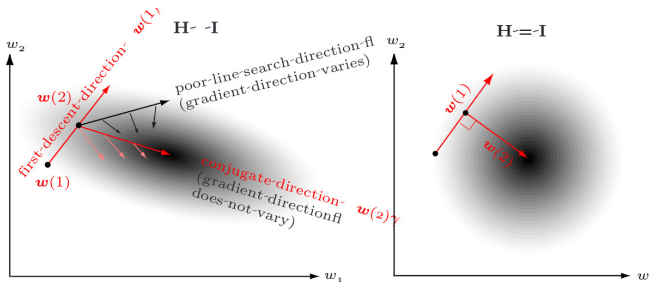
Metoda gradientów sprzężonych

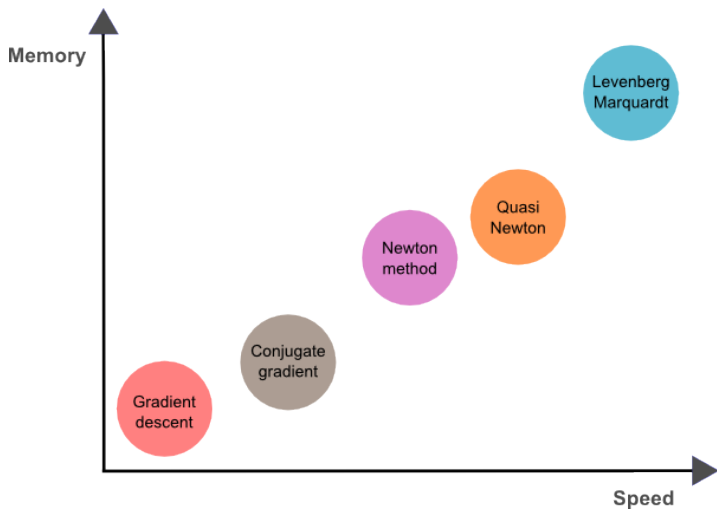
- ponieważ na skutek przybliżeń w kolejnych iteracjach kierunki mogą zatracić ortogonalność po n krokach (gdzie n jest liczbą optymalizowanych parametrów) metodę inicjujemy ponownie w ostatnio znalezionym punkcie
- dla kwadratowej funkcji kosztu w n wymiarach metoda CG osiąga minimum w n krokach
- zbieżność liniowa lecz znacznie szybsza od GD bez konieczności wyznaczania odwrotności hesjanu $\mathbf{H}^{-1}$
- małe zapotrzebowanie pamięciowe, stosunkowo mała złożoność obliczeniowa pozwala stosować do bardzo złożonych problemów optymalizacyjnych

https://en.wikipedia.org/wiki/Nonlinear_conjugate_gradient_method

Minimalizacja CG

- kolejny kierunek sprzężony nie wpływa ujemnie na wartość błędu wzdłuż poprzednio wyszukanych kierunków
- dla $\mathbf{H}$ diagonalnej kolejne kierunki są ortogonalne względem siebie



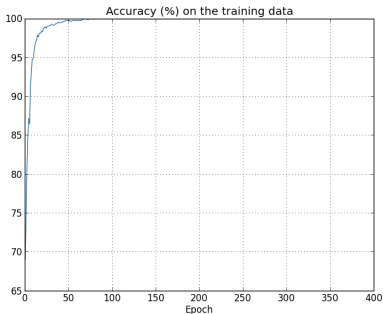


Rys: https://www.neuraldesigner.com/blog/5_algorithms_to_train_a_neural_network

Generalizacja i regularyzacja sieci MLP

Genralizacja

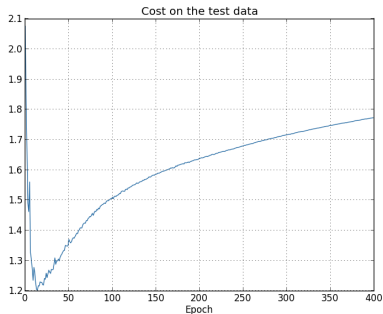
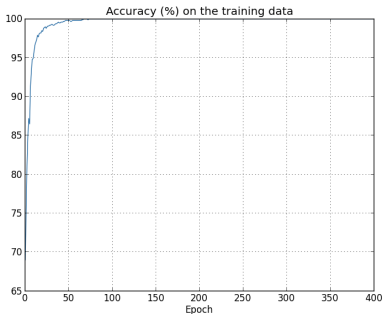
Wartość funkcji błędu dla zbioru treningowego może osiągnąć 0.
Czy to znaczy, że mamy najlepszy model?



Rys: <http://neuralnetworksanddeeplearning.com/chap3.htm>

Genralizacja

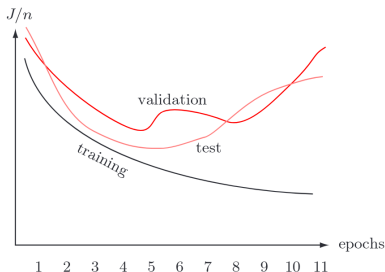
Wartość funkcji błędu dla zbioru treningowego może osiągnąć 0.
Czy to znaczy, że mamy najlepszy model?



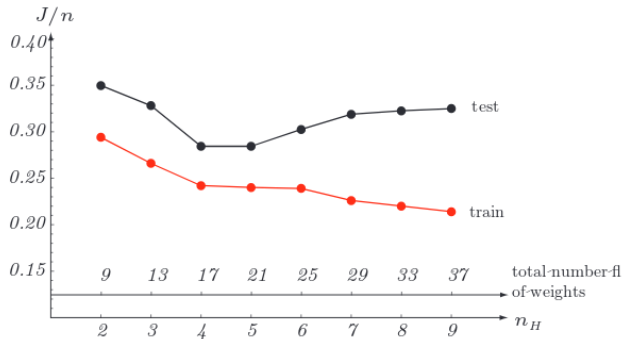
Rys: <http://neuralnetworksanddeeplearning.com/chap3.htm>

Generalizacja

- **Generalizacja** jest celem uczenia, zdolność modelu do uogólniania reguły, która wytworzyła dane.
W praktyce dążymy do minimalizacji błędu na danych, które nie zostały użyte w treningu (tzw. zbiór testowy).
- Zbiór **walidacyjny** to wyodrębniony fragment danych (np. ze zbioru uczącego) pozwala na ocenę błędu generalizacji w czasie uczenia.
Jest przydatny przy określaniu punktu zatrzymania treningu.



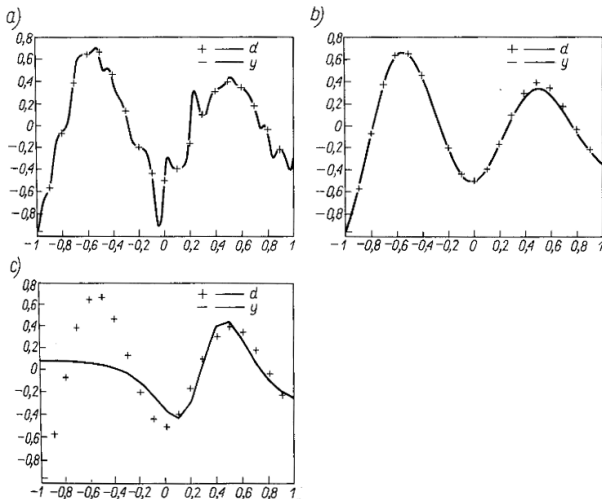
Rozmiar sieci



- ilość wag określa ilość stopni swobody - nie powinna przekraczać liczby przypadków uczących (np. $n/10$)
- za dużo wag - model uczy się na pamięć (przeuczenie)
- za mało wag - model zbyt prosty (niedouczony)

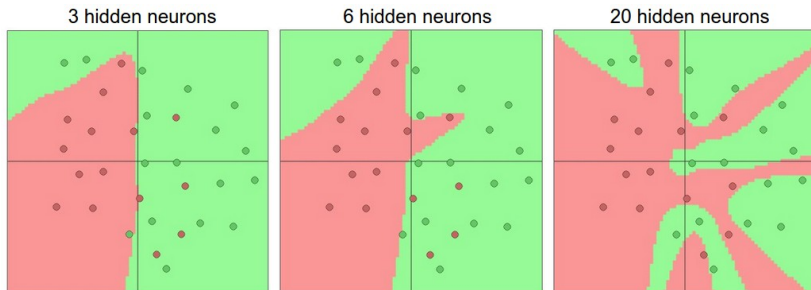
Rozmiar warstwy ukrytej

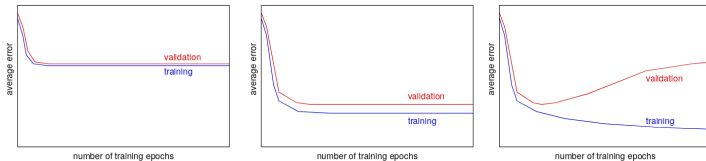
Zdolność uogólniania w problemie aproksymacji



Rozmiar warstwy ukrytej

Zdolność uogólniania w problemie klasyfikacji





- niedouczenie - błąd treningowy i walidacyjny pozostają duże
- przeuczenie - błąd walidacyjny rośnie a błąd treningowy maleje

Metody regularyzacji sieci neuronowych

Regularyzacja to metody, których celem jest poprawienie generalizacji

- dobór wielkości modelu (ilości neuronów i wag), preferowane są najprostsze rozwiązania (brzytwa Ockhama)
 - spośród wielu modeli wybierz ten o najmniejszym błędzie walidacyjnym i najmniejszej liczbie parametrów (neuronów)
 - **pruning** - usuwanie nadmiarowych połączeń lub neuronów z wytrenowanej sieci
 - metody **konstrukcyjne** (rosnące) - zacznij od najprostszego modelu i zwiększaj jego złożoność w kolejnych iteracjach
 - sieci **ontogeniczne** - rozrastające się i kurczące się w czasie treningu
- małe wartości wag są preferowane
 - wagi zerowe oznaczają brak połączenia
 - neurony sigmoidalne z małymi wagami są w przybliżeniu liniowe

Regularyzacja

- modyfikacje funkcji kosztu wymuszające odpowiednią strukturę sieci, np. wymuszające minimalizację wartości wag

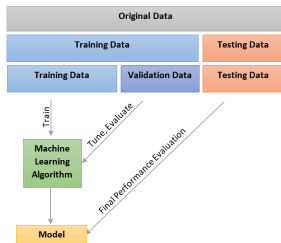
$$\bar{E}(\mathbf{w}) = E(\mathbf{w}) + \lambda \Omega(\mathbf{w})$$

gdzie $\lambda > 0$ steruje wielkością regularyzacji

- **wczesne zatrzymanie** treningu, zanim błąd walidacyjny wzrośnie
- zwiększenie liczby danych
- przetransformowanie danych do takiej postaci, która zwiększy szansę znalezienia optymalnego rozwiązania
- dodanie szumu do wag lub danych treningowych
- selekcja cech - wybór tylko istotnych zmiennych do treningu
- uśrednianie wyników z wielu modeli, np.: boosting, bagging

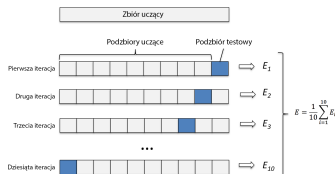
Trening i walidacja

- uczenie sieci minimalizuje błąd treningowy a nie błąd generalizacji
- **zbiór walidacyjny** pozwala oszacować błąd generalizacji w trakcie treningu. Wydzielany ze zbioru treningowego, więc powoduje zmniejszenie liczby próbek uczących.
- **zbiór testowy** używany wyłącznie do ewaluacji nauczonego modelu

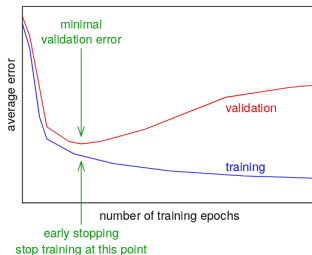


Walidacja krzyżowa (kroswalidacja)

- **Kroswalidacja** - metoda pozwalająca na oszacowanie błędu generalizacji i wariancji modelu poprzez powtarzanie treningu i ewaluacji na kolejnych podziałach zbioru uczącego
- k krotna kroswalidacja - dzieli zbiór na k części, gdzie $k - 1$ jest używanych do treningu a reszta do testu
- kroswalidacja **leave-one-out** (LOO) - gdy N równe liczbie wektorów uczących, tj. trening odbywa się na $N - 1$ przypadkach. Stosowane dla bardzo małych danych.
- średni błąd wyznaczany jest z k treningów



Wczesne zatrzymanie



Strategie:

- przerwij trening w momencie, gdy błąd walidacyjny zacznie rosnąć (np. gdy brak poprawy przez k epok). Często wystarczy tylko kilka iteracji.
- trenuj przez N epok i zapamiętaj model o najmniejszym błędzie walidacyjnym wyznaczonym po każdej epoce

Rys: Riedmiller, *Machine Learning*

Dobór liczby parametrów

Błąd generalizacji $E_G(h)$ hipotezy h pochodzącej z klasy $\mathbb{H}$ o skończonym wymiarze VC wytrenowany na N przypadkach jest ograniczony przez:

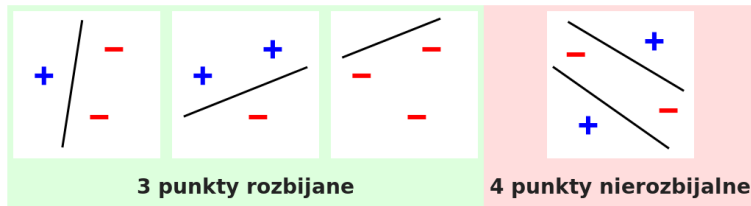
$$E_G(h) \leq E_T(h) + \sqrt{\frac{VC_{dim}(\mathbb{H}) \cdot \left(1 + \log\left(\frac{2N}{VC_{dim}(\mathbb{H})}\right)\right) - \log\left(\frac{\delta}{4}\right)}{N}}$$

z prawdopodobieństwem $1 - \delta$ jeżeli $VC_{dim}(\mathbb{H}) \ll N$.

VC_{dim} - miara złożoności Vapnika-Chervonenkisa (VC dimension), zależy od złożoności sieci (ilości wag).

Wymiar Vapnika-Chervonenkisa

Wymiar VC klasyfikatora to największa liczba punktów, którą klasyfikator jest w stanie *rozbić*, tj. podzielić przy dowolnej kombinacji etykiet (2^N możliwości w przypadku 2 klas)



Dla perceptronu z 2 wejściami $y = w_1x_1 + w_2x_2 + b$

$$VC_{dim}(\text{Perceptron}(\mathbb{R}^2)) = 3$$

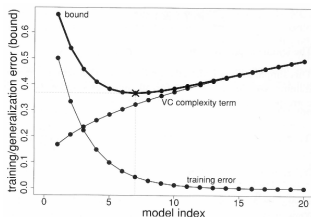
W przestrzeni $\mathbb{R}^d$ mamy $(d + 1)$ parametrów.

$$VC_{dim}(\text{Perceptron}(\mathbb{R}^d)) = d + 1$$

Złożoność VC_{dim} dowolnego modelu jest proporcjonalna do liczby **wolnych parametrów** (wag). Regularyzacja zmniejsza VC_{dim} .

Minimalizacja ryzyka strukturalnego

Structural risk minimization (SRM) (Vapnik) zasada równoważenia złożoności modelu względem jego zdolności do dopasowywania się do danych



- błąd generalizacji rośnie wraz z VC_{dim} , podczas gdy błąd treningowy spada
- jeśli znamy VC_{dim} klasyfikatora to możemy oszacować ile próbek uczących jest potrzebnych aby uzyskać satysfakcjonujący błąd generalizacji
- im bardziej złożony model tym więcej potrzeba danych uczących

Dobór złożoności sieci

Dla sieci MLP o wagach $\mathbf{w}$ w skrócie

$$E_G(\mathbf{w}) \leq E_T(\mathbf{w}) + \epsilon \left(\frac{N}{VC_{dim}}, E_T \right)$$

gdzie $\epsilon()$ określa przedział ufności.

Rule of thumb

$$E_G(\mathbf{w}) \approx \frac{VC_{dim}}{N}$$

- prawdopodobieństwo przeuczenia rośnie, gdy $VC_{dim} > N$

Wymiar VC dla MLP

Oszacowanie wymiaru VC dla sieci MLP z binarną funkcją aktywacji

$$2 \left\lceil \frac{N_h}{2} \right\rceil d \leq VC_{dim} \leq 2N_w(1 + \log N_n)$$

N_h - ilość neuronów ukrytych

d - ilość cech (wymiar przestrzeni wejściowej)

N_w - liczba wag

N_n - liczba wszystkich neuronów

- w praktyce $VC_{dim} \approx N_w$,
dla funkcji sigmoidalnych $VC_{dim} \approx 2N_w$
- oszacowanie wymiaru VC pozwala określić minimalny wymiar zbioru uczącego
- liczba próbek powinna być kilkukrotnie (np. 10 razy) większa od wymiaru VC

Regularyzacja L_2

Regularyzacja L_2 (weight decay, regularyzacja Tichonova) - dąży do zmniejszenia wartości wag

$$\hat{E}(\mathbf{w}) = E(\mathbf{w}) + \frac{1}{2}\lambda \sum_{ij} w_{ij}^2$$

gdzie $\lambda > 0$ decyduje o sile regularyzacji (typowo $\lambda = 0.01$ lub 0.001)

Zmniejszane są wszystkie wagi w trakcie uczenia. Korekta wag $\hat{w}_{ij}$ w stosunku do wersji nieregularyzowanej w_{ij}

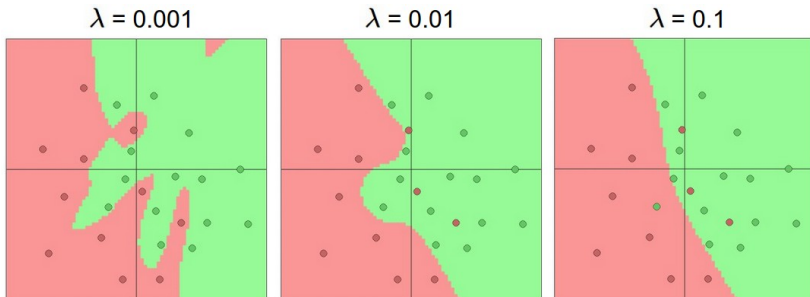
$$\hat{w}_{ij} = (1 - \lambda\eta)w_{ij}$$

$$\mathbf{w}(n+1) = \hat{\mathbf{w}}(n) - \eta \nabla E(\mathbf{w})$$

Możliwe modyfikacje:

- wagi, które zmalały poniżej pewnego progu mogą być usunięte
- usuwamy neurony dla których $\sum |w_i|$ jest bliskie zera

Regularyzacja



Rys: <http://cs231n.github.io/neural-networks-1/>

Zmodyfikowana regularyzacja L_2

W przypadku regularyzacji L_2 wartości wszystkich wag maleją.
Poniższa modyfikacja zachowuje wagi o dużych wartościach

$$\hat{E}(\mathbf{w}) = E(\mathbf{w}) + \frac{1}{2}\lambda \sum_{ij} \frac{w_{ij}^2}{1 + w_{ik}^2}$$

Korekta wag $\hat{w}_{ij}$ w stosunku do wersji nieregularyzowanej w_{ij}

$$\hat{w}_{ij} = \left(1 - \lambda \eta \frac{1}{(1 + w_{ik}^2)^2} \right) w_{ij}$$

- dla małych wag $w_{ij} \ll 1$ równoważne z L_2
- dla dużych wag $w_{ij} \gg 1$ regularyzacja zanika

Zmodyfikowana regularyzacja L_2

Zmodyfikowany człon kary wykorzystywany do eliminacji niepotrzebnych neuronów w warstwie:

$$\hat{E}(\mathbf{w}) = E(\mathbf{w}) + \frac{1}{2}\lambda \sum_{ij} \frac{w_{ij}^2}{1 + \sum_k w_{ik}^2}$$

Minimalizacja E powoduje:

- eliminację połączeń (wagi zanikają)
- eliminację neuronów dla których $\sum_k |w_{ik}|$ jest bliskie zera

Regularyzowana zmiana wag zależna jest od wartości w_{ij}

$$\hat{w}_{ij} = \left(1 - \lambda \eta \frac{1 + 2 \sum_{k \neq j} w_{ik}^2}{\left(1 + \sum_k w_{ik}^2 \right)^2} \right) w_{ij}$$

- małe wartości wag w_{ik} prowadzące do i -tego neuronu zanikają
- duże wartości wag $\sum_k |w_{ik}|$ neuronu niwelują wpływ czynnika korekcyjnego

Regularyzacja L_1

$$\hat{E}(\mathbf{w}) = E(\mathbf{w}) + \lambda \sum_{ij} |w_{ij}|$$

$$\nabla \hat{E}(\mathbf{w}) = \nabla E(\mathbf{w}) + \lambda \operatorname{sgn}(\mathbf{w})$$

regularyzacja wpływa na gradient ze stałą wartością (zależy tylko od znaku w_{ij})

$$\hat{w}_{ij} = \begin{cases} w_{ij} - \lambda \eta & \text{dla } w_{ij} > 0 \\ w_{ij} + \lambda \eta & \text{dla } w_{ij} < 0 \end{cases}$$

Regularyzacja L_1 daje rzadsze rozwiązanie od L_2 , tzn. dąży do rozwiązania o mniejszej liczbie niezerowych wag

Metody wrażliwościowe redukcji sieci

- Eliminacja wag na podstawie jej wpływu na wartość błędu
- Wagi o małej wartości $|w_i|$ są mniej istotne od dużych wag
- Wrażliwość połączenia

$$S_{ij} = E - E(w_{ij} = 0)$$

- Połączenia o najmniejszej wrażliwości usuwa się po czym sieć jest douczana
- Usuwamy neuron dla którego wszystkie dochodzące (lub wychodzące) połączenia są wyzerowane

Optimal Brain Damage (LeCun)

Założenie: hesjan $\mathbf{H}$ jest diagonalnie dominujący, więc uwzględniamy tylko składową diagonalną.

Z rozwinięcia funkcji błędu E w szereg Taylora i przyjmując $\nabla E = 0$ (punkt zbieżności dla wytrenowanej sieci)

$$\Delta E = \frac{1}{2} \Delta \mathbf{w}^T \mathbf{H} \Delta \mathbf{w} = \frac{1}{2} \sum_i H_{ii} \Delta w_i^2$$

Miara ważności wagi

$$S_i = \frac{1}{2} \frac{\partial^2 E}{\partial^2 w_i} w_i^2$$

Wagi o najmniejszym S_i mogą być usunięte z wytrenowanej sieci.

Optimal Brain Damage

Algorytm treningu OBD

1. Wytrenuj sieć dowolnym algorytmem uczenia
2. Wyznacz diagonalne elementy hesjanu oraz wartości wrażliwości S_i dla każdej wagi w_i
3. Posortuj wagi zgodnie z rosnącymi wartościami S_i i obetnij te o najmniejszych wartościach.
4. Wróć do punktu 1

Redukcja neuronów o małej aktywności

- Neurony o niewielkiej aktywności (których sygnał wyjściowy nie zmienia się dla całego zbioru treningowego) można usunąć bez szkody dla generalizacji
- duża aktywność neuronu świadczy o jego przydatności
- modyfikacja funkcji kosztu z czynnikiem kary za zbyt małą aktywność neuronów (Y. Chaurin)

$$\hat{E} = E + \lambda \sum_{i=1}^k \sum_{j=1}^n e(\Delta_{ij}^2)$$

gdzie Δ_{ij} oznacza zmianę sygnału wyjściowego i -tego neuronu dla j -tego wektora treningowego

- przykładowy czynnik korelacyjny

$$e = \frac{1}{1 + \Delta_{ij}^2}$$

Architektury konstruktywne (rozrastające się)

- automatycznie dobierają złożoność modelu do złożoności problemu, rozrastając się od prostych do złożonych modeli
- strategia zachłanna - dodanie neuronu maksymalizuje zysk w pojedynczym kroku rozbudowy
- przedwczesne przerwanie uczenia nie musi być katastrofą
- (zazwyczaj) niski koszt obliczeniowy, np. w każdym cyklu douczamy tylko dodatkowy neuron, reszta połączeń jest „zamrożona”
- możliwość budowania sieci o różnorodnych funkcjach aktywacji
- przykład: algorytm wieżowy i piramidalny, korelacja kaskadowa

Korelacja kaskadowa (Fahlman i Labiere, 1989)

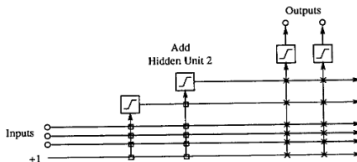
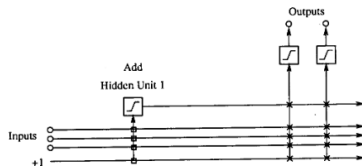
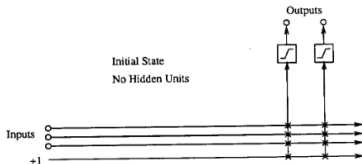
Sieć rozrastająca się:

- zaczynamy od treningu sieci bez warstwy ukrytej
- w kolejnych krokach dodawany jest neuron do warstwy ukrytej, wejścia neuronu połączone są ze wszystkimi wejściami sieci i wyjściami poprzednio dodanych neuronów ukrytych
- wagi neuronu dobierane są w procesie maksymalizacji korelacji nowego neuronu k z błędem wykazywanym przez neurony wyjściowe

$$C = \sum_{j=1}^M \left| \sum_{i=1}^N (o(\mathbf{x}_i) - \bar{o}) ((y_{ij} - f_j(\mathbf{x}_i)) - (\bar{y}_{ij} - \bar{f}_j(\mathbf{x}_i))) \right|$$

gdzie N - liczba wzorców, M - liczba wyjść sieci

Korelacja kaskadowa



Kaskadowa korelacja

- wagi kandydata maksymalizujące korelację są „zamrażane” - douczaniu podlegają wyłącznie wagi wyjściowe sieci
- jeśli wyjścia neuronu kandydata są skorelowane dodatnio z błędem to w wyniku treningu wykształci wagi, które mają szansę zniwelować ten błąd
- w każdym kroku rozpatruje się zbiór kandydatów (od 5 do 10), mogą posiadać różne funkcje aktywacji (sigmoidalna, gaussowska, itp.),
- kandydaci trenowani są równoległe konkurując ze sobą, wybiera się najlepszego (o największej korelacji)

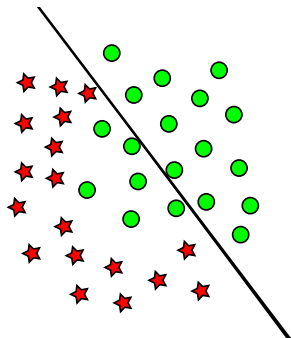
Klasyfikator cząstkowy

Niech $\mathbb{Q}^+ \cup \mathbb{Q}^- \subset \mathbb{X} \quad \wedge \quad \mathbb{Q}^+ \cap \mathbb{Q}^- = \emptyset$

Klasyfikator cząstkowy f_c

daje odpowiedź $+1$ dla co najmniej jednego obiektu ze zbioru $\mathbb{Q}^+$ oraz odpowiedź -1 dla wszystkich elementów ze zbioru $\mathbb{Q}^-$

Niech $\mathbb{R} = \{(\mathbf{x}_i, y_i) : f_c(\mathbf{x}_i) = +1\} \subset \mathbb{Q}^+$



Dla funkcji logicznych (wejścia binarne) zawsze istnieje klasyfikator cząstkowy w postaci perceptronu

Ogólna konstruktywistyczna metoda sekwencyjna

General Sequential Constructive Method (GSCM, Musseli, 1998)

Algorithm GSCM problem dwuklasowy

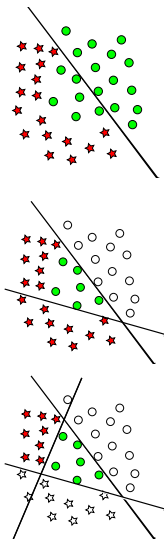
Input: Zbiór treningowy $\mathbb{D}$

Output: Sieć jednowarstwowa

- 1: $h \leftarrow 0$ ▷ pusta warstwa ukryta
 - 2: **while** zbiór $\mathbb{D}$ zawiera przypadki z przeciwnych klas **do**
 - 3: $h \leftarrow h + 1$
 - 4: wybierz etykietę $d_h = \{-1, +1\}$
 - 5: wytrenuj klasyfikator cząstkowy f_h dla klasy d_h
 - 6: $\mathbb{D} \leftarrow \mathbb{D} \setminus \mathbb{R}_h$
 - 7: Utwórz wynikową sieć zawierającą klasyfikatory cząstkowe w warstwie ukrytej
-

Przykładowe wagi połączeń do warstwy wyjściowej

$$u_0 = \sum_{j=1}^h u_j + d_{h+1}, \quad u_j = d_j 2^{h-j} \quad \text{dla } j = 1, \dots, h$$



Uczenie klasyfikatora cząstkowego

Realizacja klasyfikatora cząstkowego za pomocą perceptronu, przykłady:

- algorytm nieregularnego podziału (Irregular Partitioning IPA, Marchand, Golea, 1993)
- algorytm zamiany etykiet (Target Switch TSA, Campbell, Vicente, 1995)

Algorytm nieregularnego podziału

1. startuje z pustego zbioru odseparowanych $\mathbb{R} = \emptyset$
2. dla kolejnych wektorów $\mathbf{x} \in \mathbb{Q}^+ \setminus \mathbb{R}$
trenuje perceptron w poszukiwaniu hiperpłaszczyzny oddzielającej $\{\mathbf{x}\} \cup \mathbb{R}$ od $\mathbb{Q}^-$
jeżeli istnieje separacja to $\mathbb{R} \leftarrow \mathbb{R} \cup \{\mathbf{x}\}$
3. zwróć ostatnio uzyskany perceptron

Uczenie klasyfikatora cząstkowego

Algorytm zamiany etykiet

1. wytrenuj perceptron dowolną metodą
2. jeżeli nie realizuje on klasyfikatora cząstkowego to znajdź najdalej oddalony od płaszczyzny decyzyjnej, błędnie klasyfikowany wektor z Q^+ i zmień jego etykietę (przenieś do Q^-)
3. powtarzaj aż do uzyskania perceptronu realizującego klasyfikator cząstkowy

Algorytmy rozrastające się - podsumowanie

- Problem z przeuczeniem, kiedy zaprzestać rozrost?
- Za dużą sieć można przyciąć ograniczając wpływ przeuczenia
- Uczenie zachłanne - dokładany pojedynczy neuron, koszt obliczeń zależy od czasu uczenia nowego neuronu, złożone struktury danych mogą nie zostać wykryte przez dodanie pojedynczego neuronu, zachłanna strategia nie zawsze jest najlepsza
- Algorytmy rosnące nie gwarantują najprostszyc sieci
- Niektóre tworzą specyficzne architektury
- Sieci ontogeniczne - rosną i kurczą się w trakcie treningu
 - wyszukiwanie najlepszej architektury, stosuje się np. alg. genetyczne, ewolucyjne
 - zazwyczaj wymagające obliczeniowe

Rozmiar zbioru treningowego

- większa liczba danych zmniejsza ryzyko przeuczenia
- „There's is no data like more data”
- dodanie szumu do danych uodparnia sieć na drobne zmiany w przestrzeni wejściowej
- sztuczne zwielokrotnianie danych (*data augmentation*), często stosowane z dobrymi rezultatami
 - w klasyfikacji obrazów: zmiany kontrastów i nasycenia barw, przesunięcia, obroty, itp.
 - w analizie mowy: pogłos, szum otoczenia, zmiana głośności, przyspieszenie i zwolnienie mowy, itp..

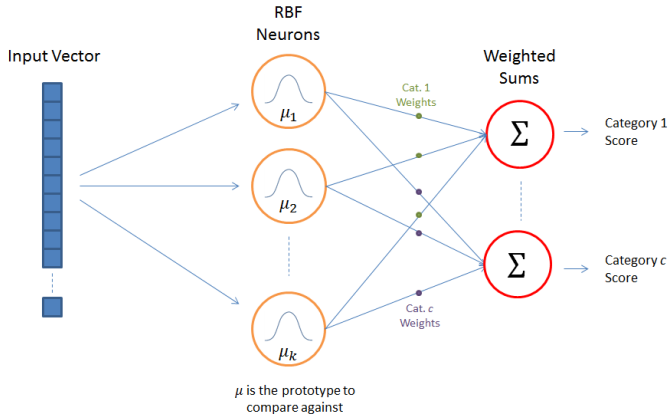
Selekcja i ekstrakcja cech

- zmienne, które nie zawierają istotnych informacji mogą wpływać negatywnie na trening
- metody selekcji cech usuwają nieistotne zmienne zmniejszając tym samym rozmiar zbioru treningowego
- dodanie cech, które niosą wartościową informację poprawia generalizację
- istnieje wiele metod, najczęściej dobierane są do specyfiki problemu
- niezbalansowane klasy - częsty problem w klasyfikacji, standardowe funkcje błędu identycznie traktują błędy popełniane dla różnych klas. W ekstremalnym przypadku wszystkie przypadki z klasy mniejszościowej zostaną błędnie przypisane do klasy większościowej.

RBF Radial Basis Functions

Sieci neuronowe z radialnymi
funkcjami bazowymi

Radial Basis Function Network



Rys: Chris McCormick, Radial Basis Function Network (RBFN) Tutorial

Sieć RBF

Funkcja realizowana przez sieć RBF z jednym wyjściem:

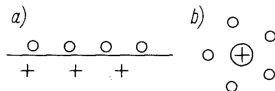
$$f(\mathbf{x}) = \sum_{i=1}^K w_i \varphi(\|\mathbf{x} - \mathbf{c}_i\|) + w_0$$

gdzie K - liczba funkcji bazowych z centrami w $\mathbf{c}_i$

- jedna warstwa ukryta, neurony z radialnymi funkcjami
- wagi połączeń neuronów ukrytych definiują centrum $\mathbf{c}_i$ funkcji radialnej
- funkcja bazowa (radialna) φ określa podobieństwo przypadku $\mathbf{x}$ do centrum (prototypu) $\mathbf{c}_i$
- wyjścia liniowe - ważona suma pobudzeń neuronów radialnych

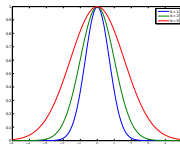
Funkcje radialne

- funkcje radialne mapują relacje lokalne, gdzie perceptrony (np. neurony sigmoidalne) mapują relacje globalne



- argumentem funkcji radialnej jest odległość od pewnego centrum (od prototypu) $\|\mathbf{x} - \mathbf{c}\|$. Zazwyczaj stosuje się metrykę euklidesową.
- najczęściej stosowane funkcje radialne $\varphi(x)$ posiadają maksimum w $x = 0$ i zanikają w obszarach dalekich od $x = 0$, np. funkcja Gaussa

$$\varphi(x) = e^{-\frac{x^2}{\sigma}}$$



- współczynnik dyspersji σ decyduje o zasięgu funkcji bazowej

Przykłady funkcji radialnych

Niech $r = \|\mathbf{x} - \mathbf{c}_i\|$

$h(r) = r$ liniowa

$h(r) = e^{-(\frac{r}{\sigma})^2}$ gaussowska

$h(r) = (\sigma^2 + r^2)^\beta, \quad 1 > \beta > 0$ multiquadratic (wielokwadratowa)

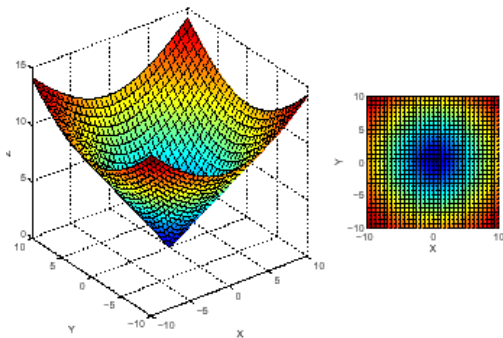
$h(r) = (\sigma^2 + r^2)^{-\alpha}, \quad \alpha > 0$ inverse multiquadratic

$h(r) = (\sigma r^2) \ln(\sigma r)$ thin-plate spline (cienkiej płytki)

$h(r) = \sigma^2 + r^2$ wielomianowa

Funkcja liniowa współrzędnej radialnej

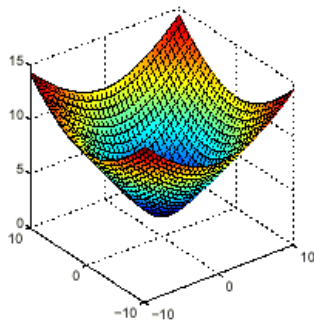
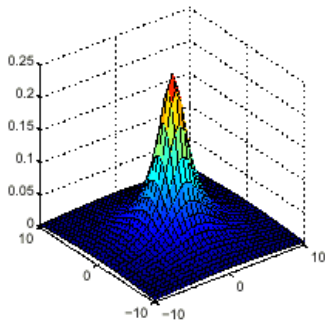
$$h_i(r) = r = \|\mathbf{x} - \mathbf{c}_i\|$$



Funkcje wielokwadratowe

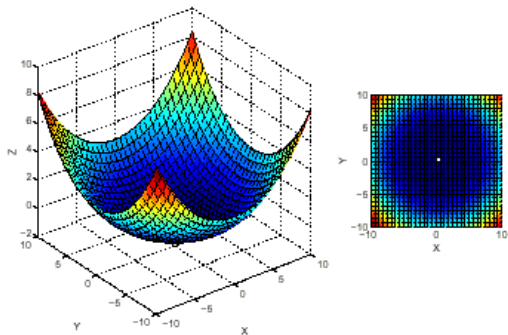
$$h(r) = (\sigma^2 + r^2)^{-\alpha}, \quad \alpha = 1$$

$$h(r) = (\sigma^2 + r^2)^{\beta}, \quad \beta = 0.5$$



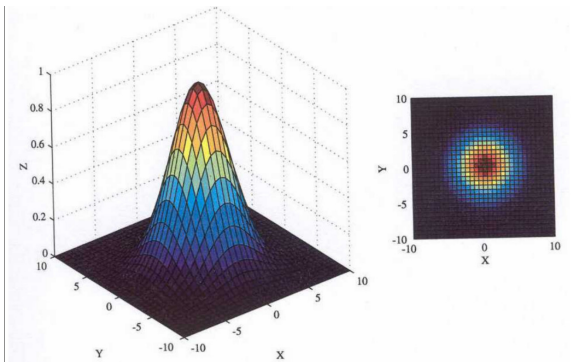
Funkcja cienkiej płytki

$$h(r) = (\sigma r^2) \ln(\sigma r)$$



Funkcja Gaussa

$$h(r) = e^{-\frac{r^2}{2\sigma^2}}$$

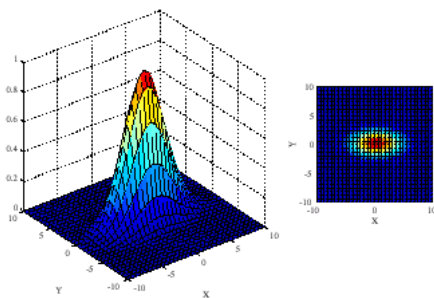


Najczęściej używana w roli funkcji bazowej RBF

Funkcja Gaussa z macierzą dyspersji

$$h(\mathbf{r}) = e^{-\frac{r^2}{2\sigma^2}} = e^{-\frac{1}{2}(\mathbf{x}-\mathbf{c})^T \Sigma^{-1}(\mathbf{x}-\mathbf{c})}$$

Funkcja Gaussa wielu zmiennych



Macierz Σ określa rozmycia i rotacje funkcji w wielowymiarowej przestrzeni

Uczenie RBF jako problem aproksymacji

Dla N punktów $\mathbf{x}_i, y_i$ znajdź funkcję spełniającą

$$f(\mathbf{x}_i) = y_i \quad i = 1, \dots, N$$

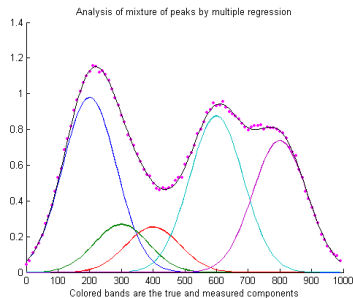
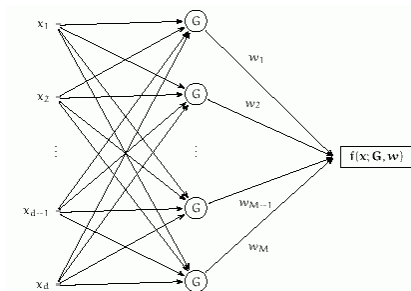
funkcja RBF

$$f(\mathbf{x}_i) = \sum_{j=1}^K w_j \varphi(\|\mathbf{x}_i - \mathbf{c}_j\|)$$

funkcja błędu MSE

$$E = \sum_{i=1}^N (y_i - f(\mathbf{x}_i))^2$$

RBF z jedną warstwą ukrytą i linowym wyjściem jest uniwersalnym aproksymatorem



Rys: <https://terpconnect.umd.edu/~toh/spectrum/CurveFittingB.html>

Rozwiązanie RBF

Dla $K = N$, funkcje bazowe o ustalonych centrach w punktach treningowych $\mathbf{x}_i = \mathbf{c}_i$

$$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,N} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N,1} & h_{N,2} & \cdots & h_{N,N} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

$$h_{ij} = \varphi(\|\mathbf{x}_i - \mathbf{x}_j\|)$$

$$\mathbf{H}\mathbf{w} = \mathbf{y}$$

dla $\mathbf{x}_1 \neq \mathbf{x}_2 \neq \dots \neq \mathbf{x}_N$ macierz $\mathbf{H}$ jest nieosobliwa i dodatnio określona, więc istnieje

$$\mathbf{w} = \mathbf{H}^{-1}\mathbf{y}$$

RBF jako problem aproksymacji

- dla radialnych f. bazowych macierz interpolacji $\mathbf{H}$ dodatnio określona (Light 1992)
- dla wąskich f. Gaussowskich (gdy $\mathbf{H} \approx \mathbf{I}$) istnieje rozwiązanie $\mathbf{w} = \mathbf{y}$ o niskiej generalizacji
- problem źle określony, przewymiarowanie
- hiperpłaszczyzna interpolacji nie jest gładka
- większe dyspersje σ i mniej funkcji ($K < N$)
→ lepsza generalizacja

Aproksymacja RBF

Dla ustalonych $\mathbf{c}_i$ oraz σ_i gdy $K < N$

$$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,K} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N,1} & h_{N,2} & \cdots & h_{N,K} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_K \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{pmatrix}$$

$$h_{ij} = \varphi(\|\mathbf{x}_i - \mathbf{c}_j\|)$$

$$\mathbf{H}\mathbf{w} = \mathbf{y}$$

rozwiązanie układu równań przez pseudoinwersję

$$\mathbf{w} = \mathbf{H}^+ \mathbf{y}, \quad \text{gdzie} \quad \mathbf{H}^+ = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T$$

W praktyce pseudoinwersję realizuje się za pomocą dekompozycji SVD

Rozwiązanie RBF

$$\mathbf{w} = \mathbf{H}^+ \mathbf{y}$$

dokonując rozkładu

$$\mathbf{H}^+ = \mathbf{USV}^T$$

gdzie $\mathbf{U}$ i $\mathbf{V}$ są ortogonalne o wymiarach $N \times N$ oraz $K \times K$
Macierz $\mathbf{S}$ o rozmiarach $N \times K$ jest pseudodiagonalna, gdzie

$$s_1 \geq s_2 \geq \dots \geq s_k \geq 0$$

Redukując liczbę kolumn do r naistotniejszych wartości

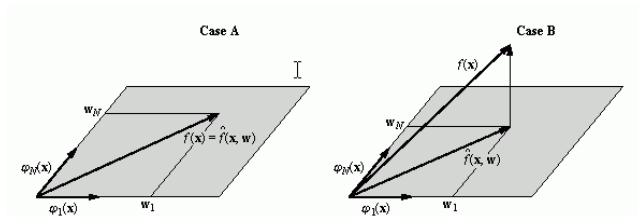
$$\mathbf{w} \approx \mathbf{V}_r \mathbf{S}_r^{-1} \mathbf{U}_r^T \mathbf{y}, \quad \text{gdzie} \quad \mathbf{S}_r^{-1} = \left[\frac{1}{s_1}, \frac{1}{s_2}, \dots, \frac{1}{s_r} \right]$$

- dobór wag w jednym kroku dla ustalonych funkcji radialnych
- macierze $\mathbf{V}_r$ i $\mathbf{U}_r$ są ortogonalne, więc problem jest dobrze uwarunkowany
- głównym problemem jest dobór centrów i rozmyć funkcji φ

Interpretacja geometryczna

Jeśli prawdziwa aproksymowana funkcja $f(\mathbf{x})$ leży w przestrzeni rozpiętej przez wektory bazowe $\Phi(\mathbf{x})$ to możliwe jest rozwiązanie bez błędu, w przeciwnym razie aproksymowana jest projekcja ortogonalna (błąd jest ortogonalny do p-ni bazowej).

$$\hat{f}(\mathbf{x}; \mathbf{w}) = \sum_i w_i \varphi_i(\mathbf{x})$$



Rozwiązanie z regularyzacją

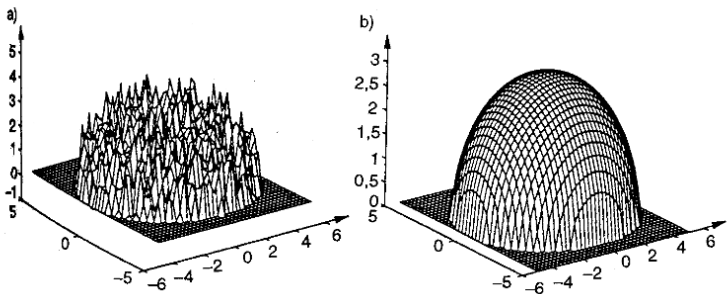
Minimalizacja funkcji błędu z członem regularyzacyjnym

$$E = \sum_{i=1}^N (y_i - f(\mathbf{x}_i))^2 + \lambda \|\hat{P}f\|^2$$

gdzie operator $\hat{P}$ wymusza np. gładkość funkcji aproksymującej

$$\|\hat{P}F_{\mathbf{w}}\|^2 = \int_{\mathbf{R}^N} \left\| \frac{\partial^2 F_{\mathbf{w}}(\mathbf{x})}{\partial x_i \partial x_j} \right\| d\mathbf{x}$$

Wpływ regularyzacji



Nadmiarowa liczba funkcji bazowych o małej dyspersji bez regularyzacji i po regularyzacji (Ossowski 1996)

Separacja w wielowymiarowej przestrzeni

Twierdzenie (Cover 1965):

Jeśli przekształcić wzorce $\mathbb{X} = \{\mathbf{x}_i\}, i = 1, \dots, N$, nieliniową funkcją na wektory $\Phi(\mathbf{x}_i) = [\varphi_1(\mathbf{x}_i), \varphi_2(\mathbf{x}_i), \dots, \varphi_K(\mathbf{x}_i)]^T$ gdzie $K > N$ to rośnie prawdopodobieństwo liniowej separacji, tj. istnieje płaszczyzna

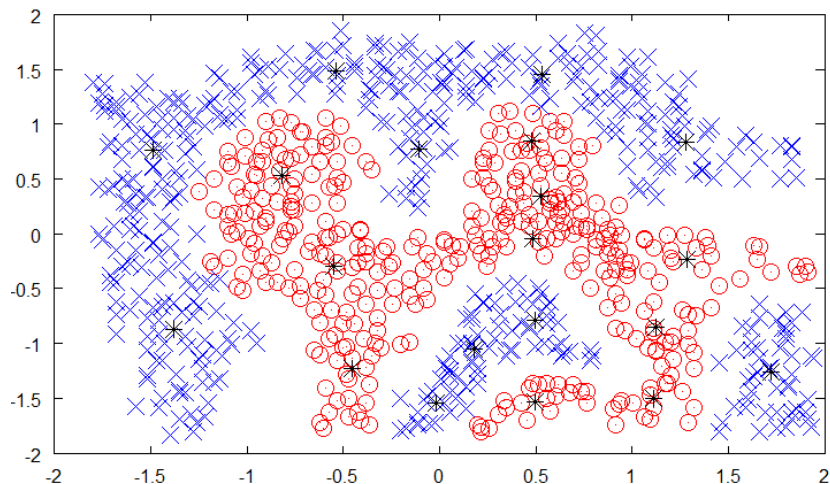
$$\begin{aligned}\mathbf{w}^T \Phi(\mathbf{x}_i) &\geq 0 & \text{dla } \mathbf{x}_i \in \mathbb{C}_1 \\ \mathbf{w}^T \Phi(\mathbf{x}_i) &< 0 & \text{dla } \mathbf{x}_i \in \mathbb{C}_2\end{aligned}$$

granica decyzyjna zdefiniowana jest równaniem

$$\mathbf{w}^T \Phi(\mathbf{x}) = 0$$

Przy dostatecznie dużej wymiarowości przestrzeni do której rzutujemy wzorce prawdopodobieństwo separowalności liniowej rośnie do 1.

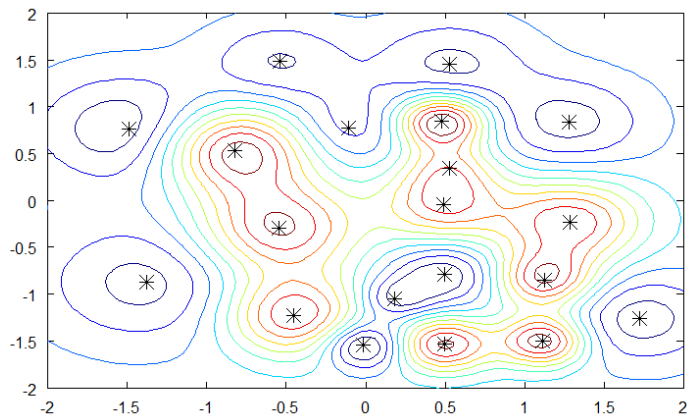
Przykład: 2 klasy



Rys: Chris McCormick, Radial Basis Function Network (RBFN) Tutorial

Przykład: 2 klasy

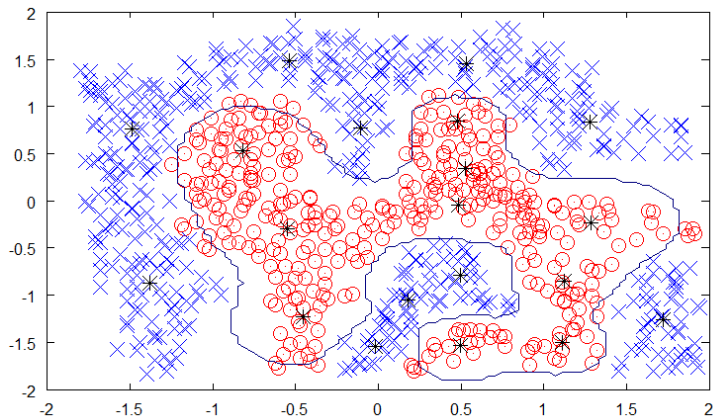
Wartości dla klasy 1, najczęściej wagi powiązane z funkcjami radialnymi dla klasy 1 mają wartości dodatnie a dla klasy 2 - ujemne



Rys: Chris McCormick, Radial Basis Function Network (RBFN) Tutorial

Przykład: 2 klasy

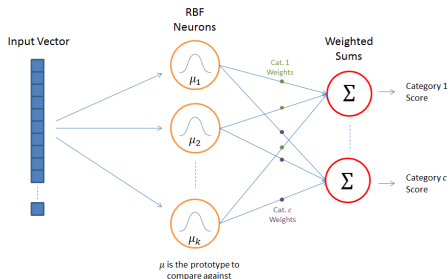
Granica decyzyjna



Rys: Chris McCormick, Radial Basis Function Network (RBFN) Tutorial

Sieć RBF

- jedna warstwa ukryta z radialnymi funkcjami + wyjścia liniowe
- parametry: wagi $\mathbf{w}$, centra $\mathbf{c}_i$, dyspersja σ_i (lub pełna macierz Σ), liczba funkcji radialnych K



Rys: Chris McCormick, Radial Basis Function Network (RBFN) Tutorial

Sieć RBF

- Sieć **GRBF** (Generalized RBF) gdy $K < N$ liczba funkcji radialnych jest mniejsza niż liczba wzorców
- Sieć **HRBF** (Hyper RBF) zamiast normy euklidesowej używa normy ważonej $\|\mathbf{x}\|_Q^2$ z pełną macierzą **Q** obrotów i skalowania **Q**

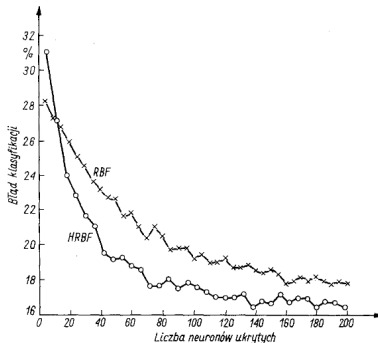
$$\begin{aligned}\|\mathbf{x} - \mathbf{c}\|_Q^2 &= (\mathbf{x} - \mathbf{c})^T \mathbf{Q}^T \mathbf{Q} (\mathbf{x} - \mathbf{c}) \\ &= \sum_{ij} s_{ij} (x_i - c_i)(x_j - c_j)\end{aligned}$$

posiada o wiele więcej parametrów do uczenia ale zwykle wymaga mniejszej liczby funkcji bazowych.

Gdy **Q** jest diagonalna to $\|\mathbf{x} - \mathbf{c}\|_Q^2 = \sum_i s_{ii} (x_i - c_i)^2$
różne skalowanie (dyspersja) w każdym wymiarze

Gdy **Q** = **1** to $\|\mathbf{x}\|_Q^2 = \|\mathbf{x}\|^2$

Dobór liczby parametrów sieci RBF



HRBF uzyskuje podobną poprawność do RBF przy mniejszej liczbie neuronów

Trening RBF

Parametry do treningu: centra, dyspersje; + wagi.

Różne strategie uczenia:

- Trening 2 etapowy:
 - najpierw ustawienie funkcji bazowych
 - potem rozwiązanie układu równań (pseudoinwersja)
- Możliwe jednoczesne uczenie centrów, dyspersji i wag, np. za pomocą metod gradientowych (wsteczna propagacja)
- Inicjalizacja początkowych centrów: losowa, klasteryzacja, samoorganizacja, probabilistyczna
- Inicjalizacja dyspersji: wartość stała, zależna od gęstości danych w różnych rejonach, średnie odległości od wektorów z innych klas

Ustalenie centrów - różne strategie

- losowe równomierne pokrycie przestrzeni wejściowej - może nie oddawać specyfiki problemu
- w problemie aproksymacji ustawiamy centra w miejscach minimum i maksimum a następnie usuwamy wzorce z otoczenia tych centrów a resztę rozmieszczemy równomiernie wśród pozostałych wzorców
- dla klasyfikacji ustalamy w pobliżu granic decyzyjnych (bliskie przypadki z różnych klas) lub w centrach skupisk wektorów z danej klasy
- losowy podzbiór K wzorców treningowych - prosta ale skuteczna metoda
- douczanie pozycji centrów metodami nadzorowanymi lub nienadzorowanymi: klasteryzacja, samoorganizacja, LVQ, ...

Ustalenie dyspersji - różne strategie

- wymagane gładkie odwzorowanie, rozmycie pełni rolę regularyzacji
- „pola recepcyjne” wszystkich funkcji bazowych powinny pokrywać cały obszar danych wejściowych
- pola recepcyjne sąsiadujących funkcji bazowych powinny nakrywać się w niewielkim stopniu
- rozmycie σ jednakowe dla wszystkich funkcji np. (Haykin, 1994):

$$\sigma = \frac{d}{\sqrt{2K}}$$

gdzie K ilość funkcji bazowych, d maksymalna odległość pomiędzy centrami. Funkcja gaussowska przybiera tu postać:

$$\varphi(\|\mathbf{x} - \mathbf{c}_i\|) = e^{-\frac{K\|\mathbf{x} - \mathbf{c}_i\|^2}{d^2}}$$

- rozmycie równe średniej odległości od sąsiadujących centrów

Ustawienie dyspersji - różne strategie

- σ_i dla każdej funkcji radialnej równe odległości euklidesowej od najbliższego sąsiedniego centrum klastra (Tarasenko, 1994)
- uwzględniając odległość od P najbliższych sąsiadów (gdzie zwykle $P \leq 3$) (Moody, 1989)

$$\sigma_i = \sqrt{\frac{1}{P} \sum_{i=1}^P \|\mathbf{c}_i - \mathbf{c}_k\|^2}$$

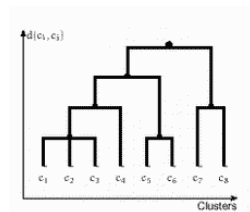
Incjalizacja RBF: klasteryzacja

Inicjalizacja centrów funkcji radialnych w centrach skupisk (klastrow) uzyskanych za pomocą algorytmów klasteryzacji (analizy skupień).

Przykład: metoda dendrogramów, algorytm k -średnich

Metoda dendrogramów:

1. przypisz każdy x_i do odrębnego klastra
2. połącz najbliższą parę klastrów x_i i x_j w jeden klaster
3. powtarzaj punkt 2 aż do uzyskania zadowalającej liczby klastrów lub gdy najmniejsza odległość między klastrami przekroczy ustalony poziom



Odległość między klastrami może być definiowana na wiele sposobów, np. minimalna odległość między obserwacjami z różnych klastrów

Incjalizacja RBF: klasteryzacja

Algorytm k -średnich

1. Rozmieść k centrów $\mathbf{c}_i$ równomiernie w przestrzeni wejściowej lub losowo wybierając k obserwacji ze zbioru treningowego
2. Dla każdego centrum $\mathbf{c}_i$ znajdź zbiór wszystkich punktów $\mathbf{x}_j$ leżących najbliżej tego centrum
3. Aktualizuj położenie centrum $\mathbf{c}_i$ ustawiając je w punkcie wyznaczonym przez średnią położenia punktów przypisanych do danego centrum

$$\mathbf{c}_i(k+1) = \frac{1}{N_i} \sum_{j=1}^{N_i} \bar{\mathbf{x}}_j(k)$$

4. Powtarzaj dwa ostatnie kroki aż do uzyskania zbieżności

👉 Clustering Demo

Inicjalizacja RBF: samoorganizacja

Algorytm inicjalizacji centrów przez samoorganizację.

Metoda on-line k -średnich:

1. Wybierz losowo k wektorów $\mathbf{c}_i$
2. Dla każdego $\mathbf{x}$ ze zbioru treningowego znajdź najbliższe centrum $\mathbf{c}_j$
3. Zmień położenie centrum $\mathbf{c}_j$ zwycięzcy przesuwając je w kierunku punktu $\mathbf{x}$

$$\mathbf{c}_i(t+1) = \mathbf{c}_i(t) + \eta_t(\mathbf{x} - \mathbf{c}_i(t))$$

4. Powtarzaj kroki 2 i 3 określoną liczbę razy lub do uzyskania zbieżności

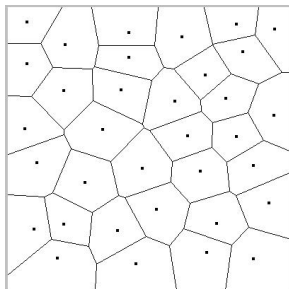
Stała uczenia η_t powinna zanikać ze wzrostem liczby iteracji t

$$\eta_t = \frac{\eta_0}{1 + \frac{t}{T}}$$

gdzie T to maksymalna liczba iteracji

Diagram Voronoi

W drodze samoorganizacji przestrzeń dzielona jest na obszary Voronoi z centrum funkcji radialnej w każdej komórce definiującymi teselację

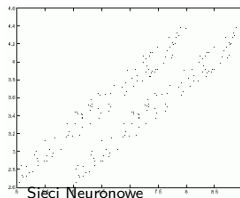


Trening RBF - obroty i rozmycia

- Optymalizacja pełnej macierzy transformacji $\mathbf{Q}$ dla danych wysoko-wymiarowych jest kosztowana obliczeniowo
- Obroty funkcji zlokalizowanych są przydatne, wystarczy zachować elementy diagonalne $Q_{ii} \neq 0$ i przydiagonalne $Q_{ii+1} \neq 0$, to wystarczy do realizacji dowolnych obrotów
- Inna metoda: iloczyn funkcji radialnej o skalarnym wsp. rozmycia σ z funkcją bi-centralną, której kierunek zanikania określony jest przez pewien wektor $\mathbf{w}$

$$\varphi(\mathbf{x}) \cdot (\sigma(\mathbf{w}\mathbf{x} + w_0) - \sigma(\mathbf{w}\mathbf{x} + w_0))$$

- Uproszczenie: kąt obrotu funkcji radialnej ustalony po inicjacji położeń



Trening RBF: metoda probabilistyczna

Zał: rozkład równomierny danych w zbiorze treningowym,
diagonalna macierz skalująca

Funkcja radialna $\varphi_i(\mathbf{x})$ reprezentuje prawdopodobieństwo warunkowe, że $\mathbf{x}$ należy do klastra w centrum $\mathbf{c}_i$

$$\varphi_i(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mathbf{c}_i)^T \Sigma^{-1}(\mathbf{x}-\mathbf{c}_i)}$$

gdzie $\Sigma^{-1} = \mathbf{Q}^T \mathbf{Q}$.

Centra $\mathbf{c}_i$ oraz rozmycia Σ_i optymalizowane są równocześnie

$$\mathbf{c}_i(k+1) = \frac{\mathbf{c}_i(k) + \eta_k(\mathbf{x}\varphi_i(\mathbf{x}) - \mathbf{c}_i(k))}{(1 - \eta_k) + \eta_k\varphi_i(\mathbf{x})}$$

$$\Sigma_i(k+1) = \frac{\Sigma_i(k) + \eta_k(\varphi_i(\mathbf{x})\|\mathbf{x} - \mathbf{c}_i(k)\|^2 - \Sigma_i(k))}{(1 - \eta_k) + \eta_k\varphi_i(\mathbf{x})}$$

Współczynnik uczenia maleje w czasie $\eta_k = \frac{\eta_0}{k}$, gdzie $\eta_0 \in [0, 1]$

Trening wag $\mathbf{w}_i$ warstwy wyjściowej odbywa się po ustaleniu parametrów funkcji radialnych (np. rozkład SVD, minimalizacja MSE)

Uczenie metodami gradientowymi

Minimalizacja funkcji błędu

$$E = \sum_{i=1}^N (y_i - f(\mathbf{x}_i))^2 = \sum_{i=1}^N \left(y_i - \sum_{k=0}^K w_k \varphi_k(\|\mathbf{x}_i - \mathbf{c}_k\|) \right)^2$$

dla funkcji gusowskiej i metryki euklidesowej

$$\Delta w_i = -\eta \frac{\partial E}{\partial w_i} = -\eta (y - f(\mathbf{x})) \varphi_i(\mathbf{x})$$

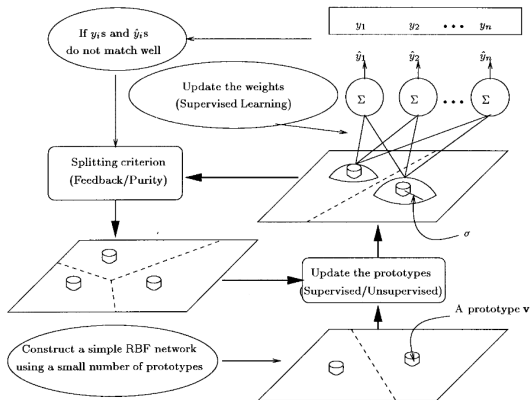
$$\Delta c_{ji} = -\eta \frac{\partial E}{\partial c_{ji}} = -\eta (y - f(\mathbf{x})) w_j \varphi_j(\mathbf{x}) \frac{(x_i - c_{ji})}{\sigma_j^2}$$

$$\Delta \sigma_j = -\eta \frac{\partial E}{\partial \sigma_j} = \eta (y - f(\mathbf{x})) w_j \varphi_j(\mathbf{x}) \frac{\|\mathbf{x} - \mathbf{c}_j\|^2}{\sigma_j^3}$$

Konstruktywny RBF

- GAL (Growing and Learning)
- GrRBF (Growing Radial Basis Function)
- FEN (Function Estimation Networks)
- RAN (Resource Allocation Networks)
- Klasyfikatory Gaussowskie, sieci probabilistyczne ...

Growing RBF



Nicolaos, *Growing Radial Basis Neural Networks: Merging Supervised and Unsupervised Learning with Network Growth Techniques*, 1997

Supervised Growing Cell Structures (Fritzke)

1. Zainicjuj małą sieć RBF
2. Wytrenuj sieć dowolną metodą
3. Wstaw nową funkcję radialną w punkcie o największym błędzie

$$\mathbf{c}_{new} = \frac{1}{2}(\mathbf{c}_m - \mathbf{c}_n)$$

gdzie $\mathbf{c}_m$ centrum odpowiadające największemu błędowi, $\mathbf{c}_n$ - najbliższe centrum względem $\mathbf{c}_m$

4. Powtarzaj dopóki błąd nie zmaleje do pożądanej wartości

Resource-Allocating Network (RAN)

- RAN (Pratt, 1991) algorytm rozrostu sieci RBF
- Startuje od pustej warstwy ukrytej i dla danego wektora $\mathbf{x}_t$ dodawany jest nowy neuron jeżeli spełnione są kryteria „odkrywczości”

$$\|\mathbf{x}_t - \mathbf{c}_i\| \geq \epsilon(t)$$

$$\|\mathbf{e}(t)\| = \|\mathbf{y}_t - f(\mathbf{x}_t)\| > e_{min}$$

gdzie $\mathbf{c}_i$ to centrum najbliższe $\mathbf{x}_t$

- Wartość progu $\epsilon(t)$ maleje z każdym krokiem

$$\epsilon(t) = \max\{\epsilon_{max} e^{-\frac{t}{T}}, \epsilon_{min}\}$$

- zakładając, że w kroku t istnieje k neuronów, wówczas nowy neuron dodawany jest z parametrami

$$\mathbf{c}_{k+1} = \mathbf{x}_t, \quad w_{k+1,j} = e_j(t), \quad \sigma_{k+1} = \alpha \|\mathbf{x}_t - \mathbf{c}_i\|$$

- gdy kryteria „odkrywczości” nie są spełnione to błąd jest korygowany aktualizacją parametrów (spadkiem gradientu)

Porównanie MLP vs RBF

MLP	RBF
nielocalne relacje, wymagają douczania	lokalne efekty, stabilność, nie-wrażliwość na odstające przypadki
jeden rodzaj parametrów (wagi)	kilka rodzajów parametrów: wagi, centra, rozmycia
trudna inicjalizacja	łatwa inicjalizacja
trudna interpretacja	łatwa interpretacja (prototypy)
uczenie tylko pod nadzorem	możliwe uczenie bez nadzoru
zawsze wie	czasami nie wie (przypadki w obszarach nie objętych polami recepcyjnymi)
BP dość skomplikowane dla wielu warstw	uczenie łatwe bo 1 warstwa
granica decyzji perceptronu: hiperpłaszczyzna	granica decyzyjna funkcji radialnej: hypersfera
aktywacja: ważona suma sygnałów $\mathbf{w}^T \mathbf{x}$	odległość od centrum funkcji bazowej $\ \mathbf{x} - \mathbf{c}\ $

SOM Self Orgnizing Map

Samoorganizacja i uczenie konkurencyjne

Uczenie nadzorowane i nienadzorowane

Uczenie nadzorowane:

- sygnał wejściowy $\mathbf{x}_i$ posiada oczekiwaną odpowiedź $\mathbf{y}_i$
- trening dąży do minimalizacji błędu pomiędzy odpowiedzią sieci $f(\mathbf{x}_i)$ a wartością oczekiwaną $\mathbf{y}_i$
- przykłady: klasyfikacja, aproksymacja

Uczenie nienadzorowane:

- modelowanie rozkładu danych $\mathbf{x}_i$, brak sygnału zwrotnego $\mathbf{y}_i$
- samoorganizacja - prowadzi do uporządkowania sieci poprzez lokalne, niezależne oddziaływania

Mechanizmy samoorganizacji

- oparte o mechanizm konkurencji - reguła Kohonena
- oparte o regułę asocjacji Hebba (metody korelacyjne)

Aspekty samoorganizacji

- trening dostosowuje wagi neuronów (a przez to ich aktywacje) do rozkładu wartości sygnałów uczących (korelacje, podobieństwo)
- pewne grupy neuronów reagują na podobne wzorce (grupowanie)
- większe sygnały $\rightarrow$ większe wagi $\rightarrow$ większe aktywacje $\rightarrow$ większe różnice pomiędzy grupami neuronów reagujących na odmienne bodźce
- nadmiarowość danych (redundancja) - wielokrotne powtarzanie podobnych wzorców jest niezbędne w samoorganizacji

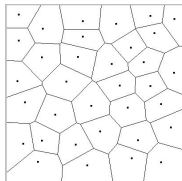
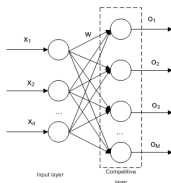
Zastosowanie samoorganizujących się sieci

- wykrywanie istotnych cech w sygnale, podobieństw i różnic
- modelowanie rozkładu danych
- **analiza skupień** (klasteryzacja) - grupowanie wektorów podobnych
- analiza składowych głównych, **redukcja wymiarowości** do najistotniejszych cech
- **prototypy** - wykrycie typowego reprezentanta pewnej grupy wektorów treningowych
- **wektorowa kwantyzacja, kodowanie, kompresja** - opisanie zbioru danych mniejszą liczbą wektorów kodujących (prototypów)
- **mapy cech** - wizualizacja organizacji danych w niskowymiarowych przestrzeniach (1D, 2D, 3D) przy zachowaniu porządku topologicznego
- klasyfikacja

Uczenie konkurencyjne

Uczenie konkurencyjne (*competitive learning*) - w trakcie uczenia neurony konkurują ze sobą o prawo do reprezentacji danych wejściowych

- **Zwycięzca bierze wszystko WTA** (*winner takes all*) nazywane również Hard Competitive Learning, tylko sygnał neuronu najbardziej pobudzonego (zwycięzcy) jest brany pod uwagę
- **Zwycięzca bierze większość WTM** (*winner take most*) - łagodniejsza forma (*soft competitive*) biorąca pod uwagę obok zwycięzcy również neurony o słabszym pobudzeniu, w pewnym otoczeniu $h(\mathbf{r}_i, \mathbf{r})$



Neuron zwycięzca k o największym pobudzeniu

$$\mathbf{w}_k^T \mathbf{x} > \mathbf{w}_j^T \mathbf{x}, \quad j = 1, \dots, N$$

Dla znormalizowanych wektorów $\mathbf{w}_i$ oraz $\mathbf{x}$

$$\hat{\mathbf{w}} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

neuron zwycięzca $\mathbf{w}_k$ ma najmniejszą odległość do $\mathbf{x}$

$$\|\mathbf{x} - \mathbf{w}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{w}\|^2 - 2\mathbf{w}^T \mathbf{x}$$

$$\min \|\mathbf{x} - \mathbf{w}\| = \max \mathbf{w}^T \mathbf{x}$$

stąd neuron zwycięzca

$$k = \arg \min_j \|\mathbf{w}_j - \mathbf{x}\| = \sqrt{\sum_{i=1}^d (x_i - w_{ij})^2}, \quad j = 1, \dots, N$$

Miary podobieństwa

- odległość Euklidesowa L_2

$$d(\mathbf{x}, \mathbf{w}_j) = \sqrt{\sum_{k=1}^d (x_k - w_{jk})^2}$$

- odległość Manhattan L_1

$$d(\mathbf{x}, \mathbf{w}_j) = \sum_{k=1}^N |x_k - w_{jk}|$$

- iloczyn skalarny (odległość kątowa)

$$d(\mathbf{x}, \mathbf{w}_j) = 1 - \mathbf{w}_j^T \mathbf{x} = 1 - \|\mathbf{x}\| \|\mathbf{w}_j\| \cos(\mathbf{x}, \mathbf{w}_j)$$

- norma L_∞

$$d(\mathbf{x}, \mathbf{w}_j) = \max_k |x_k - w_{jk}|$$

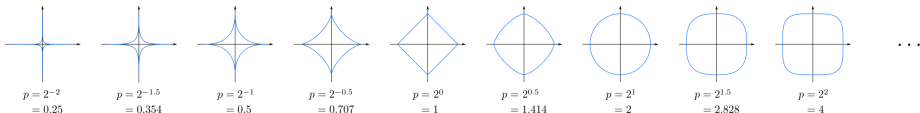
Odległość Minkowskiego

Uogólniona funkcja odległości

$$d(\mathbf{x}, \mathbf{w}) = \left(\sum_{k=1}^K (x_k - w_k)^p \right)^{\frac{1}{p}}$$

- gdy $p = 1$ odległość Manhattan
- gdy $p = 2$ odległość euklidesowa

Jednostkowy okrąg w przestrzeni R^2



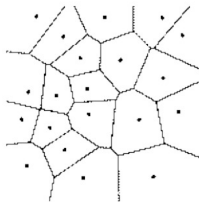
Rys: wikipedia.org

Voronoi i Delaunay

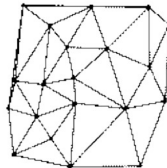
Zbiór punktów



diagram Voronoi



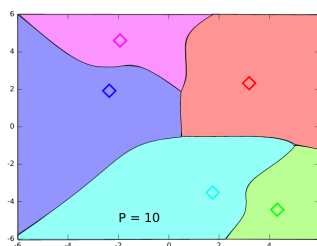
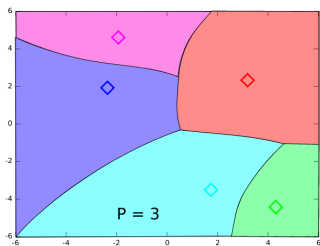
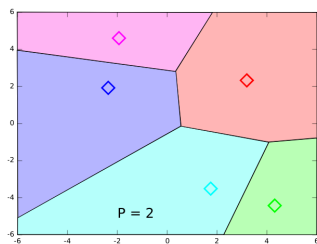
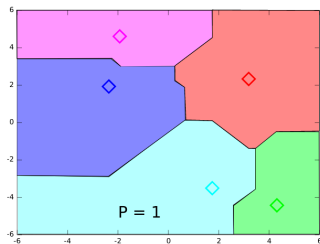
Triangulacja
Delauny



Neurony dzielą przestrzeń cech na rozdzielne obszary (obszary Voronoi), w danym obszarze tylko jeden neuron zwycięża konkurencję

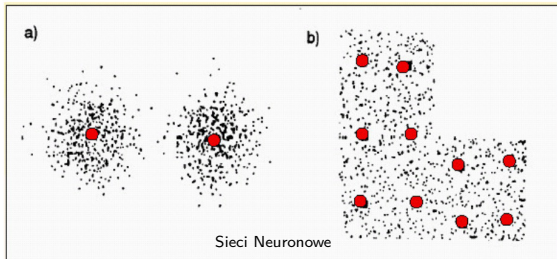
- **zbiór Voronoi** - zbiór wektorów wewnątrz obszaru Voronoi.
- łącząc węzły, których obszary Voronoia mają wspólną krawędź otrzymujemy **traingulację Delaunaya**

Strefy wpływów dla różnych metryk



Kwantowanie wektorowe

- **Kwantowanie wektorowe** - zakodowanie danych za pomocą wektorów prototypowych. Pojedynczy prototyp reprezentuje grupę wektorów wejściowych
- **Księga kodów** - zbiór prototypów kodujących (ich numery)
- Strata kompresja danych - sygnał kodowany sekwencją numerów prototypów z książki kodów, liczba prototypów k jest mniejsza od liczby wzorców n
- Do zakodowania wektora $\mathbf{x} \in \mathbb{R}^d$ wystarczy $\log_2 k$ bitów, gdzie k to rozmiar książki kodów. Cały zbiór danych o rozmiarze $d \cdot n$ wymaga $n \cdot \log_2 k + d \cdot k$ bitów



Algorytm kwantowania wektorowego WTA

1. Zainicjuj wagi k neuronów
2. Dla każdego wzorca uczącego $\mathbf{x}$ znajdź neuron zwycięzcę o największej aktywacji
3. Przesuń wagi zwycięskiego neuronu w kierunku wektora $\mathbf{x}$ (**reguła Kohonena**)

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \eta(t) \|\mathbf{x} - \mathbf{w}(t)\|$$

4. Powtarzaj 2 i 3 aż do uzyskania zbieżności

Błąd kwantowania wektorowego dla k neuronów

$$E(\mathbf{x}; \mathbf{w}, t) = \frac{1}{k} \sum_i \left\| \mathbf{x}_i - \mathbf{w}_{m(i)} \right\|^2$$

gdzie $\mathbf{w}_{m(i)}$ to wagi neuronu zwycięskiego dla sygnału $\mathbf{x}_i$

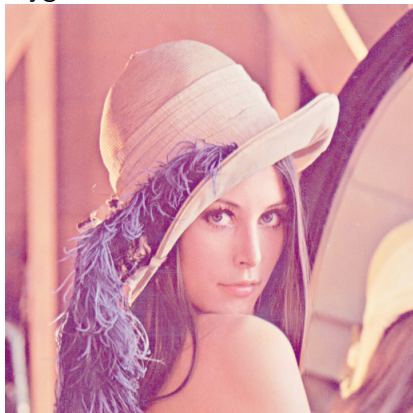
Przykład: kompresja obrazów

- obraz o wymiarach $N_x \times N_y$ dzielimy na równe ramki o wymiarach $n_x \times n_y$
- wektor $\mathbf{x}_i$ to wektor $n_x \times n_y$ pikseli (np. odcienie szarości)
Zbiór wszystkich wektorów $\mathbf{x}_i$ tworzy zbiór treningowy
- uczenie samoorganizujące dowolną metodą w celu minimalizacji błędu kwantyzacji
- podobne ramki pobudzają ten sam neuron zwycięski
- kompresja: prezentacja wszystkich ramek $\mathbf{x}_i$ pozwala ustalić kolejność neuronów zwycięskich 1, 5, 1, 30, ..., ich numery tworzą księgę kodów
- współczynnik kompresji

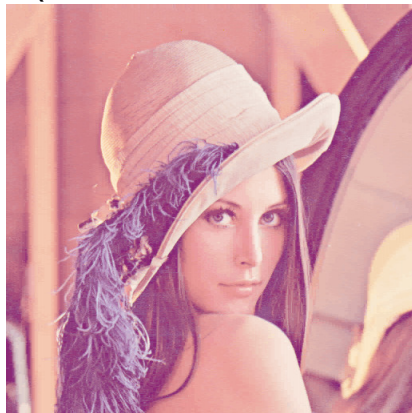
$$k_r = \frac{N n_x n_y T}{N \log_2 k + k n_x n_y t}$$

gdzie N liczba ramek, k - liczba neuronów, T liczba bitów potrzebna do reprezentacji wartości $\mathbf{x}_i$, t liczba bitów potrzebna do reprezentacji wag $\mathbf{w}_i$

oryginał



VQ 1024 codebook



kompresja: 2.6 bits/pixel

👉 Image Compression with Vector Quantization by Ivan-Assen Ivanov

Własności kwantyzacji wektorowej

- VQ nie musi odzwierciedlać gęstości rozkładu danych, przecenia rejony mało prawdopodobnych danych i nie docenia rejonów o dużym prawdopodobieństwie danych
- wynik zależy od punktu początkowego i kolejności prezentacji wzorców - każde uruchomienie może prowadzić do innej kwantyzacji
- zasada jednakowego zniekształcenia: każdy neuron powinien mieć podobny wkład do końcowego błędu kwantyzacji
- problem **martwych neuronów**: przy losowej inicjalizacji część neuronów może znaleźć się w obszarach gdzie nie będzie pobudzeń
- rozwiązanie: mechanizmy zmęczenia (np. neurony z sumieniem), odpowiadają chwilowemu zmęczeniu neuronu po wzbudzeniu w sieciach neuronowych biologicznych

Mechanizm sumienia (DeSieno)

Neurony z sumieniem (*conscience learning*) - mechanizm zmęczenia zapobiegający powstawaniu martwych neuronów

Miara częstości zwycięstw $p_i(t)$

$$p_i(t+1) = \begin{cases} p_i(t) + B \cdot (1 - p_i(t)) & \text{dla neuronu zwycięzcy} \\ p_i(t) - B \cdot p_i(t) & \text{dla pozostałych} \end{cases}$$

gdzie $0 < B \ll 1$ (np. $B = 0.0001$)

Zwycięzca spośród N neuronów

$$k = \arg \min_j \left(\|\mathbf{w}_j - \mathbf{x}\| - C \left(\frac{1}{N} - p_j(t) \right) \right) \quad j = 1, \dots, N$$

- u często wygrywających neuronów rośnie „poczucie winy”
- dla $C = 0$ nie występuje zmęczenie (standardowy algorytm Kohonena)
- potencjał p_i ograniczony do 1

Mechanizm zmęczenia ze współczynnikiem czułości

Każdy prototyp $\mathbf{w}_i$ posiada współczynnik czułości $s_i \geq 0$ początkowo równy 0.

W kolejnych krokach treningu kwantyzacji wektorowej

$$s_i(t+1) = \begin{cases} 0 & \text{dla neuronu zwycięzcy} \\ s_i(t) + \gamma & \text{dla pozostałych} \end{cases}$$

gdzie $\gamma > 0$

Zwycięzca wybierany jest spośród N neuronów

$$k = \arg \min_i (\|\mathbf{w}_i - \mathbf{x}\| - s_i) \quad i = 1, \dots, N$$

- lepsze odwzorowanie gęstości danych, nawet przy dużej zmienności gęstości w różnych obszarach

Mechanizm zmęczenia z modyfikacją odległości

Efektywna odległość zależna od częstości zwycięstw N_i

$$\hat{d}(\mathbf{x}, \mathbf{w}_i) = N_i d(\mathbf{x}, \mathbf{w}_i)$$

- kara dodawana przy określaniu zwycięzcy, częściej wygrywające neurony mają większą odległość efektywną
- w trakcie aktualizacji wag używa się oryginalnej miary odległości $d(\mathbf{x}, \mathbf{w})$
- po kilku początkowych cyklach, wszystkie neurony mają szansę wygrać przynajmniej raz

Gaz neuronowy (Schulten i Martinez 1991)

Wektorowa kwantyzacja WTM - krok aktualizacji zależny od pozycji w rankingu zwycięzców

1. Losowa inicjalizacja N neuronów
2. Powtarzaj następne kroki T razy
3. Wybierz losowy wektor $\mathbf{x}$
4. Zrób ranking neuronów posortowanych względem odległości od $\mathbf{x}$

$$\|\mathbf{w}_i - \mathbf{x}\| \leq \|\mathbf{w}_j - \mathbf{x}\|, \quad i < j$$

Niech $m(i)$ określa miejsce neuronu w rankingu, gdzie dla zwycięzcy $m(i) = 0$

5. Adaptacja

$$\Delta \mathbf{w}_i = \eta(t) h_i(t) (\mathbf{x} - \mathbf{w}_i)$$

$$h_i(t) = e^{-\frac{m(i)}{\lambda(t)}}$$

6. Zmniejsz eksponencjalnie obszar $\lambda(t)$ i stałą uczenia $\eta(t)$

Promień sąsiedztwa

Promień sąsiedztwa maleje w trakcie uczenia od λ_0 do pewnej wartości λ_{min}

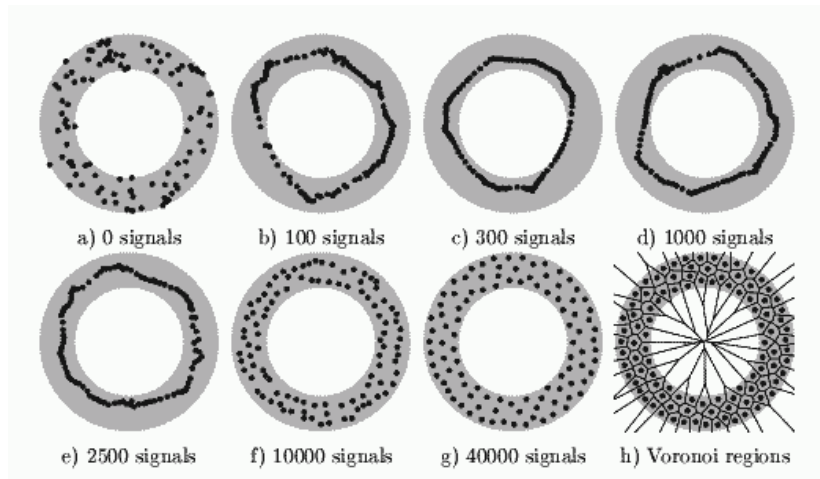
$$\lambda(t) = \lambda_0 \left(\frac{\lambda_{min}}{\lambda_0} \right)^{\frac{t}{T}}$$

- dla $\lambda = 0$ uczenie WTA
- dla $\lambda > 0$ aktualizowane są wagi wszystkich neuronów z siłą zależną od pozycji w rankingu

Współczynnik uczenia $\eta(t)$ zazwyczaj maleje wykładniczo

$$\eta(t) = \eta_0 \left(\frac{\eta_{min}}{\eta_0} \right)^{\frac{t}{T}}$$

Przykład kwantyzacji gazu neuronowego



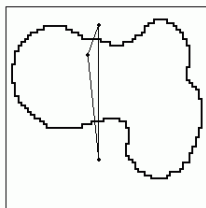
Growing Cell Structures

Rozrastająca się wektorowa kwantyzacja. Neurony połączone są krawędziami definiującymi ich sąsiedztwo.

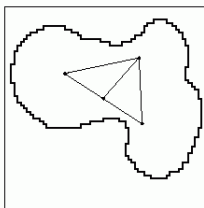
Początkowa topologia: k -wymiarowy sympleks (np. $k=1, 2, 3$).

1. Dla losowego $\mathbf{x}$ znajdź zwycięzcę c
2. Aktualizuj wagi zwycięzcy oraz połączonych sąsiadów:
$$\Delta \mathbf{w}_i = \eta_s (\mathbf{x} - \mathbf{w}_i)$$
3. Dla zwycięzcy c zwiększ sumaryczny błąd
$$E_c = E_c + ||\mathbf{x} - \mathbf{w}||^2$$
4. Po ustalonej liczbie epok T znajdź neuron o największym E_c i wstaw nowy neuron r w środku krawędzi łączącej go z najdalszym sąsiadem tworząc lokalny k -wymiarowy sympleks. Błąd sąsiadów r zmniejsz o wartość proporcjonalną do liczby tych sąsiadów. Błąd nowego neuronu ustaw na średni błąd jego sąsiadów.
5. Powtarzaj powyższe kroki aż do osiągnięcia kryterium stopu (np. określony rozmiar sieci)

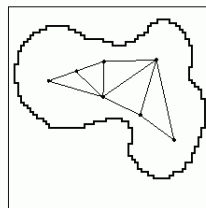
Trening GCS



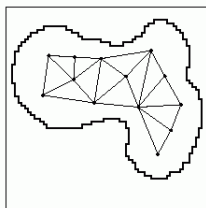
a) 0 signals



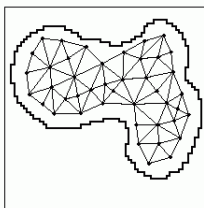
b) 100 signals



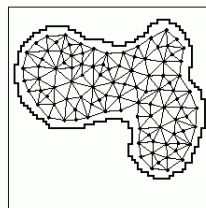
c) 400 signals



d) 1000 signals



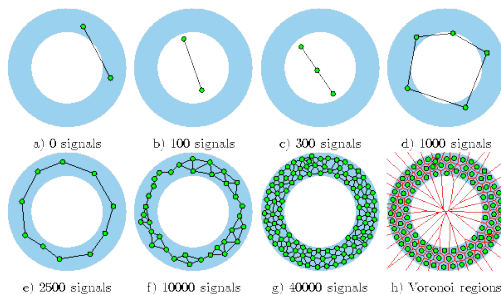
e) 4000 signals



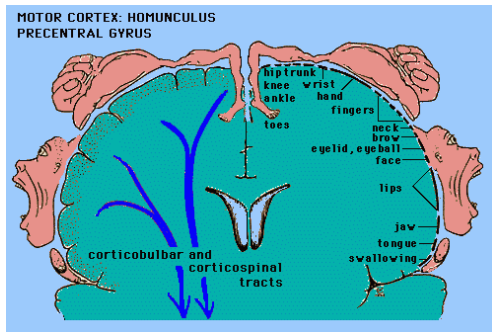
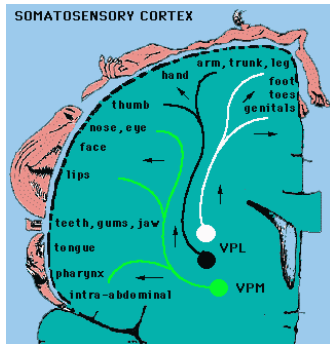
f) 10000 signals

Growing Neural Gas (Fritzke, 1994)

- GNG, rozrastający się gaz neuronowy, algorytm zbliżony do GCS, startuje od pary neuronów i wprowadza mechanizmem starzenia się połączeń
- nowe połączenie powstaje między zwycięzcą i drugim w rankingu neuronem
- połączenia, które żyją za długo są usuwane



Mapy czuciowe i motoryczne



Inspiracje biologiczne SOM

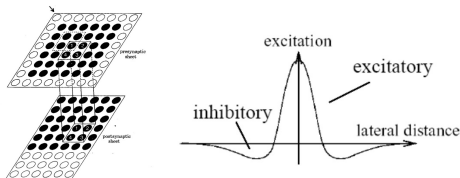
- przetwarzanie różnych danych zmysłowych i motorycznych (wzrok, słuch, dotyk, ...) odbywa się w różnych obszarach kory mózgowej
- neurony o podobnych funkcjach są obok siebie => mapy topograficzne
- istnieje topologiczna zależność w obliczeniach realizowanych przez mózg
- Przykłady
 - mapy somatosensoryczne układu czuciowego,
 - mapy motoryczne kory i mózdzku,
 - mapy tonotopiczne układu słuchowego,
 - mapy orientacji dwuocnej układu wzrokowego,
 - mapy wielomodalne układu orientacji (wzgórki czworacze górne)

Modele samoorganizacji

- Samoorganizująca się mapa cech - SOM lub SOFM (Self-Organized Feature Mapping)
- Połączenia lokalne (sąsiednie neurony): neuron silnie pobudzany przez pobliskie, słabo przez odległe, hamowany przez neurony pośrednie
- von der Malsburg i Willshaw (1976) - model układu wzrokowego
- Amari (1980) - model ciągłej tkanki neuronowej uwzględniający pobudzenie i hamowanie zależne od odległości
- Kohonen (1981) uproszczony model, bez hamowania, dwie fazy uczenia: konkurencja i kooperacja.

Model Willshaw, van der Malsburg (1976)

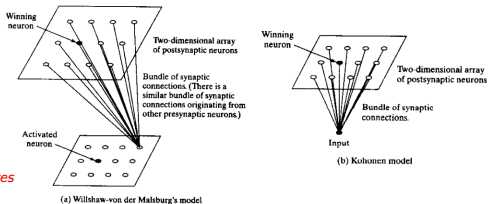
Samoorganizacja w mapowaniu sygnału z siatkówki na korę wzrokową, transformacja 2D $\rightarrow$ 2D, brak redukcji wymiaru, połączenia wewnątrzwarstwowe,



- aktywacja typu „Meksykańskiego kapelusza”
 - kooperacja - pobudzenia bliskiego zasięgu
 - konkurencja - hamowanie w dalszych regionach
- siatkówka (wejście) posiada narzuconą organizację topologiczną
- wybór zwycięskiego neuronu wynikiem dynamiki neuronalnej

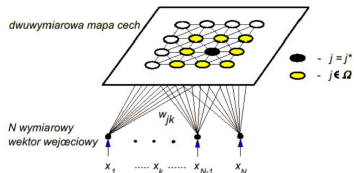
Sieć Kohonena (1981)

- Sieć Kohonena - najbardziej popularny model tworzenia map, stąd pojęcie SOM często utożsamiane z „siecią Kohonena”
- dwie fazy uczenia
 - konkurencja - najpierw wybierz globalnego zwycięzcę
 - kooperacja - uczenie neuronu zwycięzcy oraz neuronów w sąsiedztwie zwycięzcy
- sieci bez hamowania, brak połączeń wewnątrz warstwy
- wejście jest sygnałem ciągłym, nie posiada narzuconej struktury topologicznej, dowolna wymiarowość sygnału wejściowego, mapowanie nD do 1D, 2D, 3D



Rys: Marcus Kaiser, Neuroinformatics, lectures

Mapa cech Kohonena



- neurony zachowują porządek topologiczny (np. siatka 2D) tworząc mapę cech
- sąsiadujące na siatce neurony powinny reagować na podobne wzorce
- podobieństwo wzorców określa odległość $\|\mathbf{x} - \mathbf{w}_i\|$ w przestrzeni wejściowej $\mathbf{x}, \mathbf{w}_i \in \mathbb{R}^N$
- sąsiadujące neurony na siatce, odległość neuronów $\|\mathbf{r}_i - \mathbf{r}_j\|$

Rys: Małgorzata Krętowska, Sztuczne sieci neuronowe

Algorytm SOM

1. Zainicjuj wagi połączeń neuronów ukrytych
2. Dla danego sygnału $\mathbf{x}(t)$ w kroku $t = 1, 2, \dots, T$ znajdź najsilniej reagujący neuron c (np. najbliższy)

$$\|\mathbf{x} - \mathbf{w}_j\| = \sqrt{\sum_i (x_i - w_{ij})^2}, \quad c = \arg \min_j \|\mathbf{x} - \mathbf{w}_j\|$$

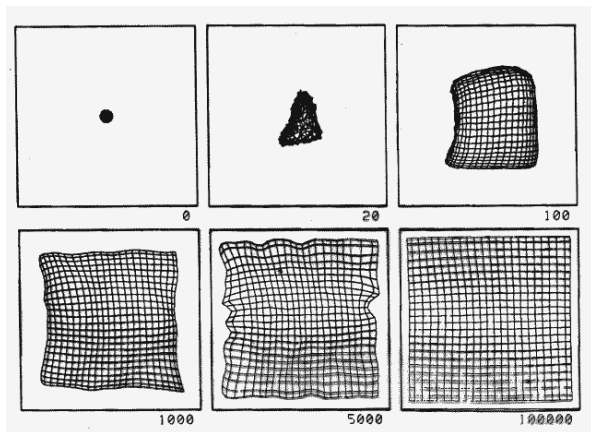
3. Zmień wagi neuronu c oraz neuronów w sąsiedztwie $O(c)$ przybliżając je w stronę wektora $\mathbf{x}$

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \eta(t) h(\mathbf{r}_i, \mathbf{r}_c, t) [\mathbf{x}(t) - \mathbf{w}_i(t)] \quad \text{dla } i \in O(c)$$

Funkcja $h(\mathbf{r}, \mathbf{r}_c, t)$ określa zasięg sąsiedztwa zwycięzcy

$$h(\mathbf{r}, \mathbf{r}_c, t) = h_0(t) \exp\left(-\frac{\|\mathbf{r} - \mathbf{r}_c\|^2}{2\sigma^2(t)}\right)$$

Przykład: mapowanie 2D $\rightarrow$ 2D

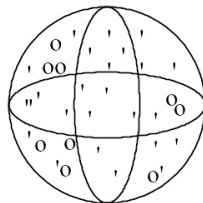


Rozkład jednostajny w kwadracie. SOM uczy się jednorodnego rozkładu. Początkowo wszystkie $\mathbf{w} \approx 0$

Sieć 2D, cechy 3D (mapowanie 3D->2D)

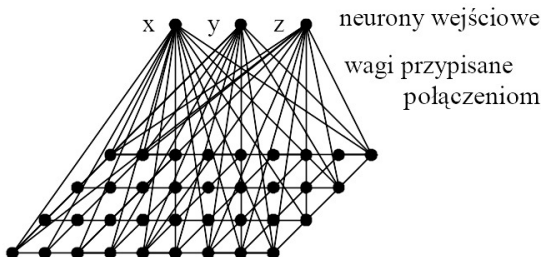
o=dane

' = wagi sieci



przestrzeń cech

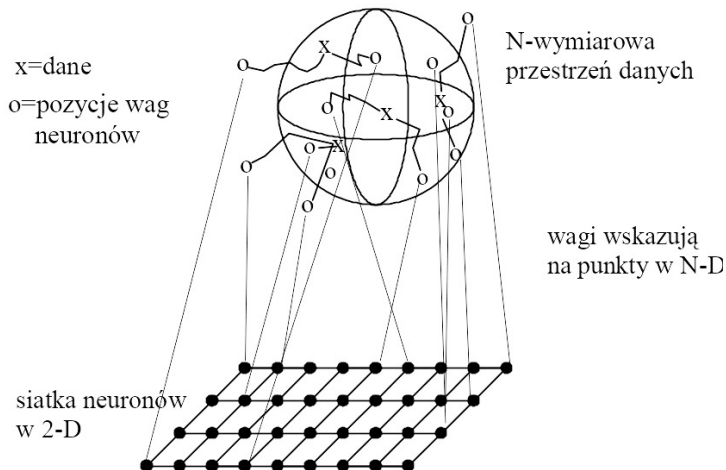
2-D siatka
neuronów



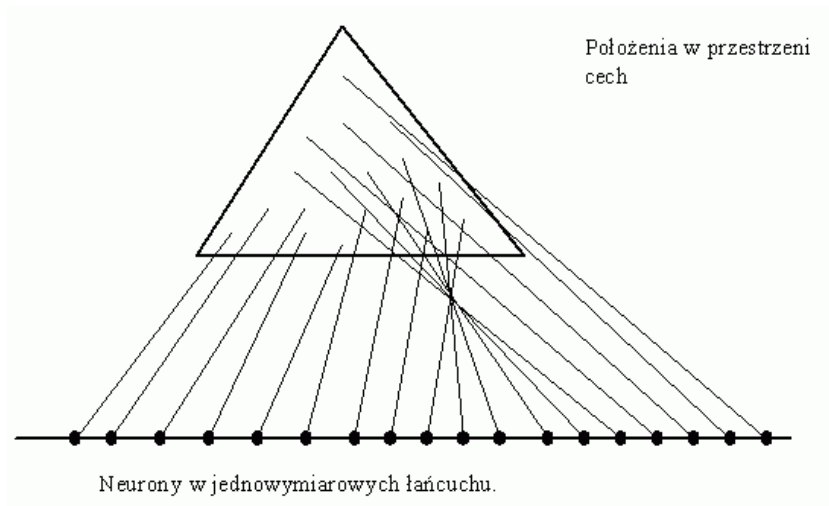
neurony wejściowe

wagi przypisane
połączeniom

Uczenie sieci 2D

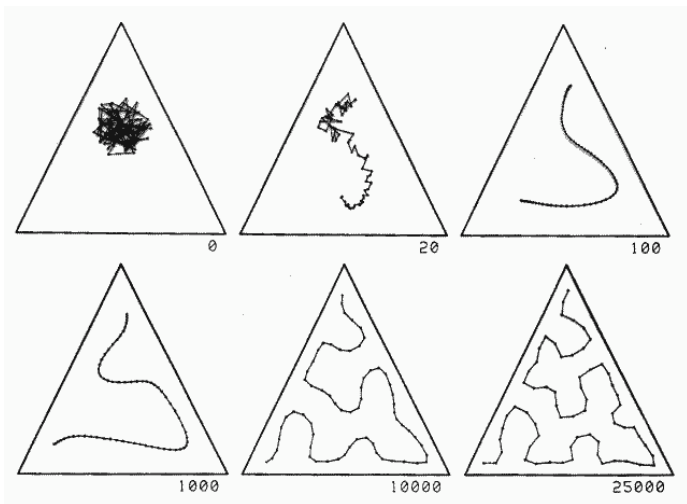


Sieć mapująca dane 2D do 1D (neurony ustawione w linii)

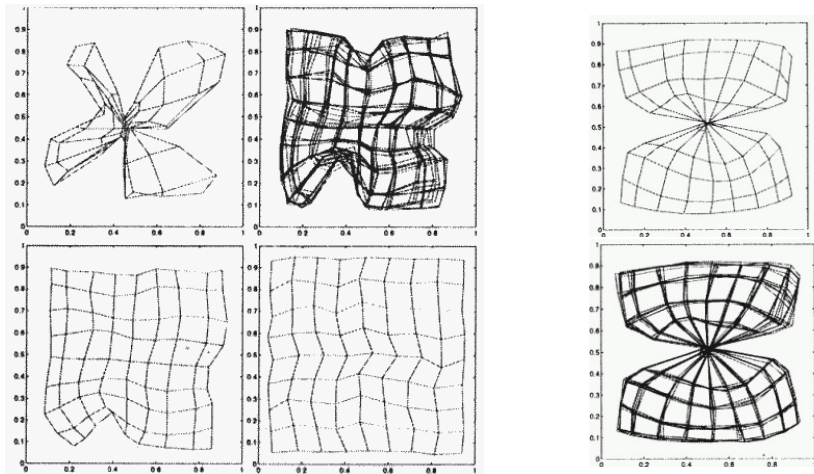


Przykład: fraktalne krzywe Peano

Krzywa Peano (lub Hilberta) - krzywa wypełniająca płaszczyznę



Zniekształcenia



Niektóre zniekształcenia nie mają szans na zniwelowanie

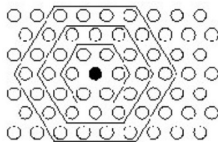
Rozmiar sąsiedztwa

Klasyczny algorytm Kohonena - sąsiedztwo prostokątne

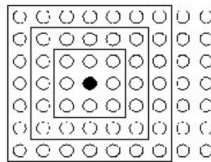
$$h(\mathbf{r}_i, \mathbf{r}, t) = \begin{cases} 1 & \text{dla } d(\mathbf{r}_i, \mathbf{r}) \leq \sigma(t) \\ 0 & \text{dla pozostałych} \end{cases}$$

Promień sąsiedztwa $\sigma(t)$ jest zmniejszany w czasie treningu, może być określony miarą odległości lub liczbą neuronów sąsiadujących

- w początkowej fazie uczenia uwzględnia duże sąsiedztwo
- w końcowej fazie tylko neuron zwycięski jest aktualizowany



(a) Hexagonal grid



(b) Rectangular grid

Funkcja sąsiedztwa SOM

Dla gausowskiej funkcji sąsiedztwa

$$h(\mathbf{r}_i, \mathbf{r}_j, t) = \exp\left(-\frac{|\mathbf{r}_i - \mathbf{r}_j|^2}{2\sigma(t)^2}\right)$$

zasięg funkcji maleje w czasie treningu

$$\sigma(t) = \sigma_0 e^{-2\sigma_0 t / t_{\max}}$$

Dla $\sigma = 0$ funkcja sąsiedztwa obejmuje tylko wektor zwycięzcę i SOM sprowadza się do metody kwantyzacji wektorowej (np. on-line k -średnich)

Funkcja błędu SOM

Próba wprowadzenia funkcji błędu (Luttrell; Heskes i Kappen)

Błąd lokalny neuronu i jest sumą po wszystkich neuronach:

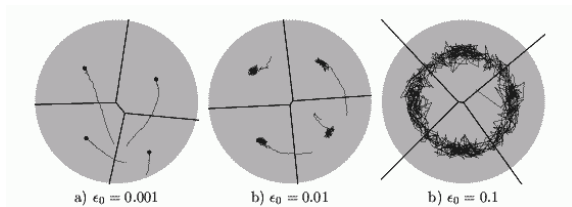
$$E_i(\mathbf{x}; \mathbf{w}, t) = \frac{1}{2} \sum_j h(\mathbf{r}_i, \mathbf{r}_j, t) \|\mathbf{x}(t) - \mathbf{w}_j(t)\|^2$$

Neuron-zwycięzca ma najmniejszy błąd lokalny:

$$c = \arg \min_i \sum_j h(|\mathbf{r}_i - \mathbf{r}_j|, t) \|\mathbf{x}(t) - \mathbf{w}_j(t)\|^2$$

Stała uczenia

Duża stała uczenia prowadzi do eksploracji znacznej części przestrzeni



Stała uczenia maleje w trakcie uczenia

$$\eta(t) = \eta_0 e^{-\frac{t}{t_{max}}}$$

Własności SOM

- sąsiednie neurony kodują sąsiednie obszary, ale sąsiednie obszary mogą być kodowane przez odległe neurony, relacje odległości nie zawsze są zachowane
- powstają „skręcone” konfiguracje przy zbyt szybkiej redukcji sąsiedztwa
- powolna zbieżność algorytmu SOM, wymaga wielu iteracji
- problemy związane z ustaleniem zbieżności i punktów stacjonarnych, wyniki analityczne znane tylko w 1D dla ciągłego czasu: wtedy wartości wag wzdłuż prostej porządkują się
- w oryginalnym SOM nie ma funkcji błędu, brakuje oceny pozwalającej na redukcję zniekształceń, brak szukanego minimum i nie można wyznaczyć gradientu
- zwykle niska jakość klasyfikacji, SOM służy głównie do wizualizacji ... ale nie ma pewności, że wizualizacja oddaje pełną naturę danych

Przykład: Włoska oliwa

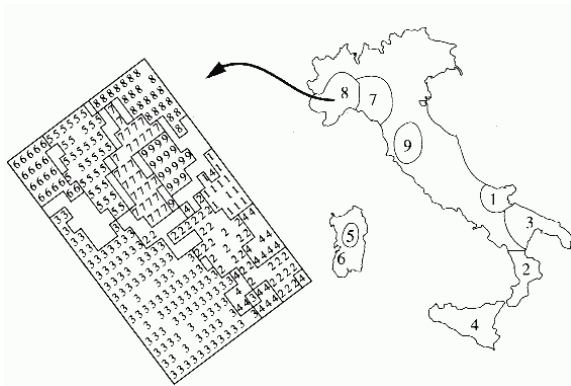
Dane: 572 próbki oliwy z 9 prowincji Włoch

Cechy: poziom 8 tłuszczu w każdej próbce.

Mapa SOM: 20 × 20, Redukcja 8D => 2D.

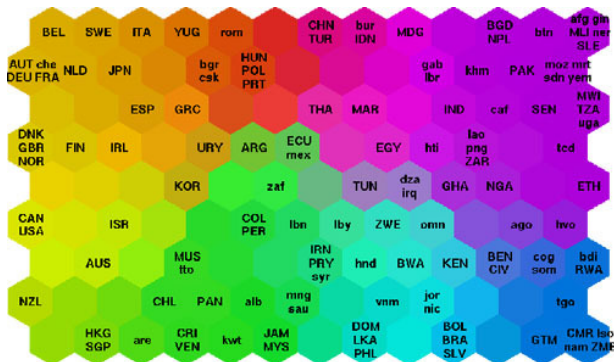
Dokładność klasyfikacji około 95-97%.

Topograficzne relacje zostały zachowane.



Przykład: World poverty map

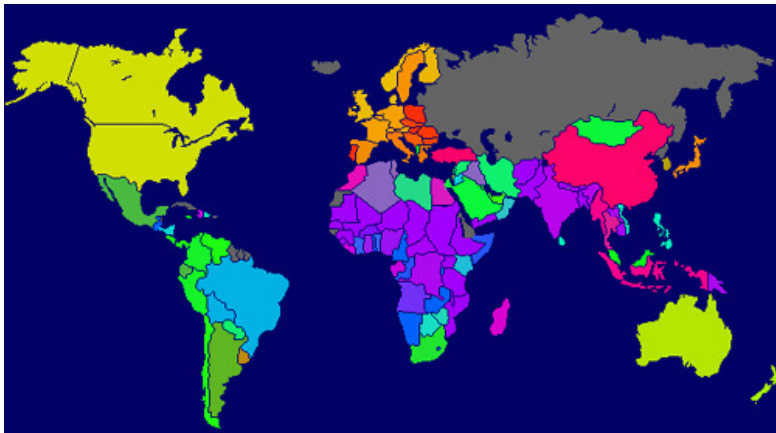
Dane: WorldBank data 1992 rok, 39 współczynników opisujących jakość życia w kraju (poziom służby zdrowia, edukacji, etc.)



Kraje o podobnej wartości poziomu życia pobudzają bliskie grupy neuronów, od krajów o najwyższym wsp. jakości życia (żółty) do najniższego (niebieski)

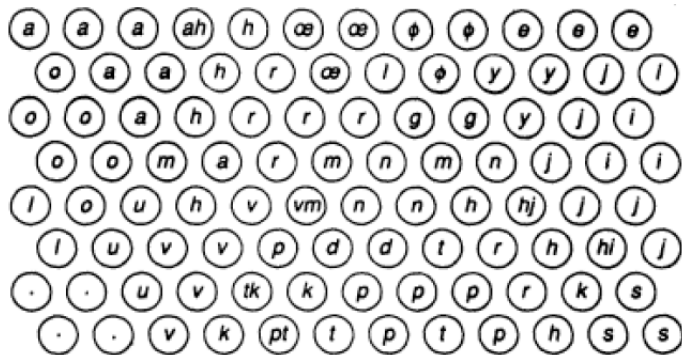
Przykład: Wodr powerty map

Po naniesieniu kolorów na mapę świata



Klawiatura fonetyczna

Dane: 10 ms próbki FFT mowy



Mapa fonetyczna dla języka Fińskiego, Kohonen 1980

Demos

- Wizualizacje do wykładu M. Czoków, J. Piersa: 🖱️ Sześcian, 🖱️ Sześcian z funkcją gaussowską, 🖱️ Kula, 🖱️ Kula z funkcją gaussowską
- 🖱️ SOM 2D dla pogrupowanych danych
- Self-organizing Maps: PyMVPA kolory na mapie 2D
- 🖱️ Demo GNG
- 🖱️ Klasyfikacja kolorów
- 🖱️ GeoSOM Environmental modelling

Przykłady zastosowania SOM

Helsinki University of Technology web site

<http://www.cis.hut.fi/research/refs/> has a list (2010) of
> 7700 papers on SOM and its applications !

- Badania mózgu: modelowanie topograficznych map w różnych obszarach kory (motoryczna, sensoryczna, wizyjna)
- Robotyka i AI: analiza danych z czujników, sterowanie ruchami robota, mapy orientacji przestrzennej
- Kategoryzacja dokumentów
- Analiza skupień genów, białek, cząsteczek chemicznych, fonemów, dźwięków zwierząt, obiektów astronomicznych, dane ekonomiczne i finansowe
- Kompresja danych (obraz, dźwięk), filtrowanie informacji
- Medical and technical diagnostics.

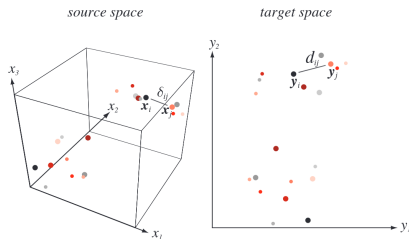
- Analiza języka naturalnego: linguistic analysis, parsing, learning languages, hyphenation patterns
- Problemy optymalizacyjne: configuration of telephone connections, VLSI design, time series prediction, scheduling algorithms.
- Przetwarzanie sygnałów: adaptive filters, real-time signal analysis, radar, sonar seismic, USG, EKG, EEG and other medical signals ...
- Rozpoznawanie i analiza obrazów: segmentation, object recognition, texture recognition ...
- Content-based retrieval: examples of WebSOM, PicSom – similarity based image retrieval, RGB and textures.
- Viscovery SOMine, komercyjny program do wizualizacji, eksploracji, klasyfikacji, prognozowania oparty na SOM.

SOM vs MDS

MDS

Multi-Dimensional Scaling (MDS) (Thorton 1954, Kruskal 1964), zwane także mapowaniem Sammona (Sammon 1964)

- redukcja wymiarowości z zachowaniem relacji odległości, zazwyczaj redukcja do 2D, 3D w celu wizualizacji
- szukamy mapowania $\mathbf{x} \in \mathbb{R}^d \rightarrow \mathbf{y} \in \mathbb{R}^k$ (gdzie $k < d$) minimalizującego **funkcję stresu** określającą miarę zniekształceń topograficznych



Rys: Duda, Hart, Pattern Classification

Algorytm MDS

Przestrzeń cech $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, 2, \dots, n$

Przestrzeń docelowa $\mathbf{y}_i \in \mathbb{R}^k$, gdzie $d < k$

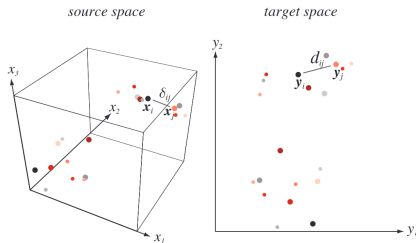
Odległość w przestrzeni cech i w przestrzeni docelowej

$$\delta_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\| \quad d_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|$$

Można użyć dowolnej metryki (najczęściej metryka Euklidesowa)

Funkcja stresu (Kruskal)

$$E = \sum_{i < j}^n (\delta_{ij} - d_{ij})^2$$



Optymalizacja $k \cdot n$ parametrów (współrzędnych $\mathbf{y}_i$)

Funkcja stresu Kruskala

$$E_1 = \sum_{i < j}^n (\delta_{ij} - d_{ij})^2 \geq 0$$

- zachowuje ogólną strukturę klastrów (skupisk)
- minimum w 0 gdy $d_{ij} = \delta_{ij}$
- mogą dominować duże wartości odległości δ_{ij} ,

Odwzorowanie Sammona

$$E_2 = \frac{1}{c} \sum_{i < j}^n \frac{(\delta_{ij} - d_{ij})^2}{\delta_{ij}} \geq 0 \quad \text{gdzie} \quad c = \frac{1}{\sum_{i < j} \delta_{ij}}$$

- wpływ dużych odległości δ_{ij} jest zredukowany

MDS i SOM

SOM

- nie gwarantuje wiarygodnej aproksymacji gęstości danych
- brak minimalizowanej funkcji, uczenie samoorganizujące się na zasadzie konkurencji
- wiele innych zastosowań, np. klasyfikacja

MDS

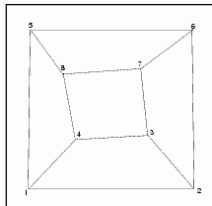
- redukcja wymiarowości z zachowaniem relacji odległości
- brak funkcji mapującej $f : \mathbf{x} \rightarrow \mathbf{y}$
- dodanie nowego punktu treningowego $\mathbf{x}_i$ wymaga ponownej optymalizacji

SOM i MDS

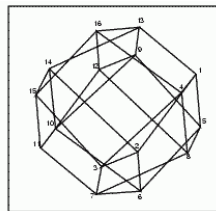
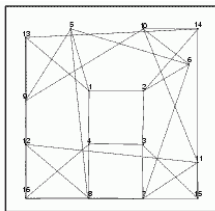
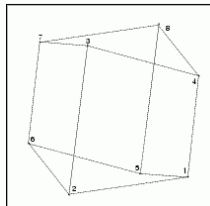
- wrażliwe na szum w danych i nieistotne cechy
- wyniki uzależnione od punktu startowego, każda optymalizacja może generować odmienne wyniki

Wizualizacja: hiperkostka 3D i 4D

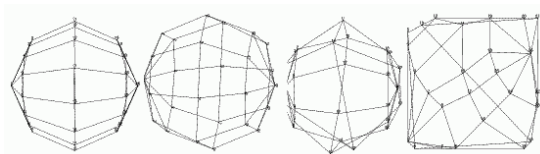
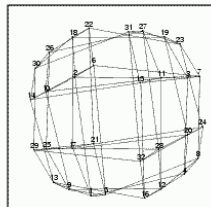
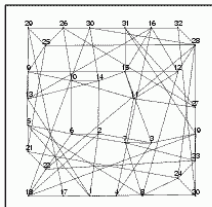
SOM warstwa 2D, 20x20



MDS



Wizualizacja: hiperkostka 5D, punkty na sferze w 3D



The two-dimensional representations of the 26 points on the sphere obtained by minimization of *S*, *E*, *A*, and by **SOFM** (left to right) with a 20 x 20 neurons map.

Mapy semantyczne

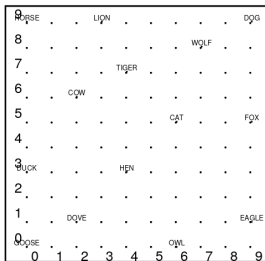
- próba uchwycenia sensu słów i pojęć poprzez własności oraz relacji semantycznych
- Przykład: 8 ptaków i 8 ssaków
dove, hen, duck, goose, owl, hawk, eagle,
fox, dog, wolf, cat, tiger, lion, horse, zebra, cow.
- każde pojęcie opisane 13 cechami binarnymi
size is: small, medium large;
has 2 legs, 4 legs,
has hair, hoofs, mane, feathers;
likes to: hunt, run, fly, swim.
- utwórz zdania opisujące zwierzęta za pomocą wybranych cech
Horse is big, has 4 legs, mane, hair, hoofs, likes to run.

Mapy semantyczne

animal		d	o	h	u	s	w	e	a	g	f	d	o	c	g	i	l	o	r	b	e	c
		o	v	e	n	k	e	l	k	e	x	g	f	t	r	n	e	a	w			
is	small	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	
	medium	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	
	big	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	
has	2 legs	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
	4 legs	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	
	hair	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	
	hooves	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	
	mane	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	1	0	0	
	feathers	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
likes to	hunt	0	0	0	0	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	
	run	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1	0	0	0	
	fly	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
	swim	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Mapy semantyczne

SOM (Ritter & Kohonen 1989)



MDS (Naud & Duch, 1996))

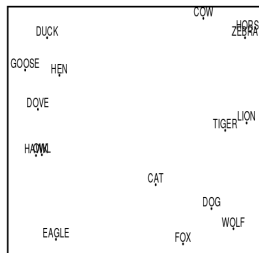


FIG. 1. The two-dimensional representations of the 13-dimensional semantic data obtained by SOM (left) with a 10 x 10 neurons map, a training of 10000 cycles, with final stress of 0.25, and the MDS (right) with final stress of 0.20 after 10 iterations.

Naturalna klasyfikacja zwierząt zachowująca ich podobieństwo

- ssaki odseparowane od ptaków
- małe zwierzęta od dużych
- drapieżniki od roślinożerców

Podobieństwo zwierząt jest powiązane z odległością na mapie

Samoorganizacja

Uczenie hebbowskie

Samoorganizacja

- **uczenie nienadzorowane**: wykrywa samoistnie (bez nadzoru) istotne cechy powiązań między sygnałami
- globalne uporządkowanie sieci jest możliwe przez działania samoorganizujące prowadzone **lokalnie** w różnych punktach sieci, niezależnie od siebie
- pewne neurony lub grupy neuronów **współpracują** lub/i **konkurują** ze sobą specjalizując w wykrywaniu określonych wzorców

Sieci samoorganizujące się

Uczenie konkurencyjne

- reguła Kohonena
- konkurencja WTA, WTM
- kwantyzacja wektorowa, SOM

Uczenie hebbowskie

- reguła Hebba, reguła Oja
- asocjacje i korelacje między sygnałami
- dekompozycja składowych głównych PCA (Principal Component Analysis)
- dekompozycja na składowe niezależne ICA (Independent Component Analysis) , separacja źródeł BSS (Bling Source Separation), nielinowe PCA, ...

Reguła Hebba (1949)

„Neurons, that fire together, wire together.”

Jeżeli neuron A jest cyklicznie pobudzany przez neuron B, to staje się on jeszcze bardziej czuły na pobudzenie tego neuronu

- synchroniczne pobudzanie neuronów A i B wzmacnia połączenie synaptyczne między nimi
- jeżeli neurony nie są pobudzane jednocześnie (asynchronicznie) to połączenie jest osłabiane

Zmiana wagi połączenia pomiędzy neuronami A i B

$$w_{AB}(k+1) = w_{AB}(k) + \eta y_A(k) y_B(k)$$

Reguła Hebba

Dla neuronu liniowego

$$y = \mathbf{w}^T \mathbf{x}$$

reguła Hebba przybiera postać

$$\Delta \mathbf{w} = \eta y \mathbf{x} = \eta \mathbf{x} \mathbf{x}^T \mathbf{w}$$

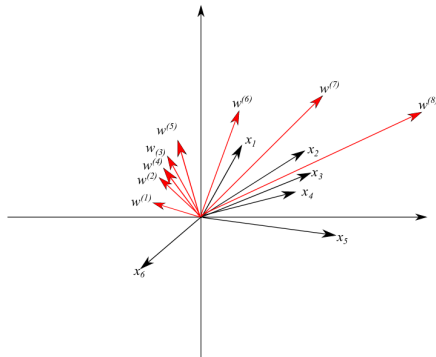
Decydujący wpływ na wartość końcową $\mathbf{w}$ będą miały sygnały zgrupowane (rozciągające się) wzdłuż znajdującego kierunku - maksymalizacja wariancji $\mathbf{w}^T \mathbf{x}$.

Przy prezentacji wielu próbek uczących wektor wag kumuluje wartości macierzy autokorelacji $\mathbf{R}$ wektorów $\mathbf{x}$

$$\Delta \mathbf{w} = \eta \mathbf{R} \mathbf{w} \quad \text{gdzie} \quad \mathbf{R} = E[\mathbf{x} \mathbf{x}^T]$$

Interpretacja geometryczna reguły Hebba

Wektor wag przemieszcza się w kierunku wyznaczonym przez środek masy wektorów treningowych (maleje kąt pomiędzy kierunkiem największej wariancji danych)



Trening prowadzi do nieograniczonego wzrostu wag (norma $\|\mathbf{w}\|$ rośnie w każdym kroku)

Modyfikacje reguły Hebba

Zapobieganie nieograniczonemu wzrostowi wag:

- sztywne granice

$$w_{min} < w_i < w_{max}$$

- miękkie granice

$$\Delta w_i = \eta x_i y (w_i - w_{min})(w_{max} - w_i)$$

- normalizacja wag po każdym kroku uczenia
- dodanie czynnika normalizującego wagi, np. reguła Oji (1982)

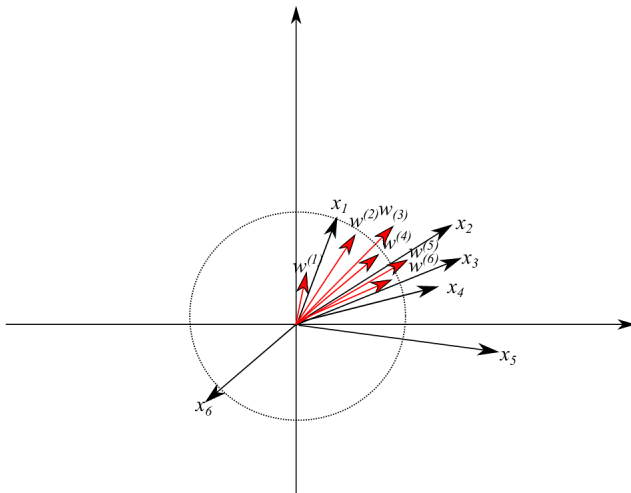
$$\Delta w_i = \eta y (x_i - y w_i)$$

Reguła Oji

$$\Delta \mathbf{w} = \eta (y\mathbf{x} - \alpha y^2 \mathbf{w})$$

- dodanie wyrazu zanikania wagi proporcjonalnego do y^2
- adaptacja prowadzi do unormowania wag $\|\mathbf{w}\| = \frac{1}{\sqrt{\alpha}}$
- wektor wag zbliża się do wektora własnego macierzy kowariancji $E[\mathbf{X}\mathbf{X}^T]$ o największej wartości własnej
- wektor wag leży wzdłuż kierunku maksymalizującego wartość oczekiwaną y^2 (wariancja wartości wyjściowej neuronu)

Interpretacja geometryczna reguły Oji



Własności reguły Oji

Dla wyuczonego wektora wag wartość oczekiwana zmian dla wszystkich wektorów uczących

$$E[\Delta \mathbf{w}] = 0$$

stąd

$$\begin{aligned} E[\Delta \mathbf{w}] &= \eta E[y(\mathbf{x} - y\mathbf{w})] = \eta E[(\mathbf{w}^T \mathbf{x})\mathbf{x} - y^2 \mathbf{w}] \\ &= \eta (E[\mathbf{x}\mathbf{x}^T]\mathbf{w} - E[y^2]\mathbf{w}) = 0 \end{aligned}$$

Macierz kowariancji $E[\mathbf{x}\mathbf{x}^T] = \mathbf{R}_x$.

Oznaczmy $\lambda = E[y^2]$

Reguła Oji prowadzi do wyznaczenia wektora własnego $\mathbf{w}$ macierzy kowariancji $\mathbf{R}_x$

$$\mathbf{R}_x \mathbf{w} = \lambda \mathbf{w}$$

Normalizacja wag

$$\lambda = E[y^2] = E[(\mathbf{w}^T \mathbf{x})(\mathbf{w}^T \mathbf{x})] = \mathbf{w}^T E[\mathbf{x}\mathbf{x}^T] \mathbf{w} = \mathbf{w}^T \mathbf{R}_x \mathbf{w}$$

ponieważ

$$\mathbf{R}_x \mathbf{w} = \lambda \mathbf{w}$$

stąd

$$\lambda = \mathbf{w}^T \lambda \mathbf{w} \Rightarrow \|\mathbf{w}\| = 1$$

Dodatkowy czynnik ogranicza wartość wag.

Analiza składowych głównych PCA

Transformacja liniowa

$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

gdzie

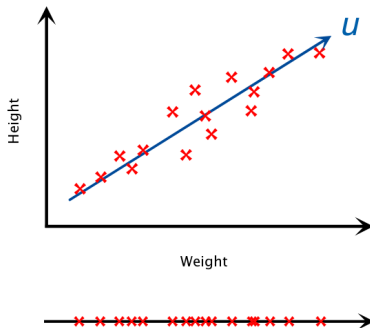
$$\mathbf{x} \in \mathbb{R}^N \quad \mathbf{y} \in \mathbb{R}^K \quad \mathbf{W} \in \mathbb{R}^{K \times N} \quad K < N$$

Redukcja wymiarowości z zachowaniem maksymalnej informacji zawartych w danych.

W PCA kierunek zawierający najwięcej informacji to kierunek o największej wariancji

Projekcja danych na wektor wag

Projekcja danych 2D na kierunek o największej wariancji



Rys: Boedecker, Machine Learning Summer 2015

PCA

Niech $\mathbf{X} \in \mathbb{R}^N$ wektor losowy o zerowej wartości oczekiwanej, wówczas wartość oczekiwana macierzy kowariancji

$$R_x = E[\mathbf{X}\mathbf{X}^T]$$

dla skończonej próbki o liczności p

$$R_x \approx \frac{1}{p} \sum_{i=1}^p \mathbf{x}_i \mathbf{x}_i^T$$

PCA

Wartości własne λ_i oraz wektory własne $\mathbf{w}_i$ macierzy kowariancji

$$\mathbf{R}_x \mathbf{w}_i = \lambda_i \mathbf{w}_i \quad i = 1, \dots, N$$

R jest symetryczna i nieujemna $\Rightarrow$ wartości własne rzeczywiste i nieujemne

Niech

$$\lambda_1 > \lambda_2 > \dots > \lambda_N \geq 0$$

Ograniczając do K pierwszych składowych

$$\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_k]^T$$

otrzymujemy transformację PCA

$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

PCA

Wariancja wzdłuż i -tego kierunku

$$\text{Var}(\mathbf{w}_i^T \mathbf{x}) = E[(\mathbf{w}_i^T \mathbf{x} - E[\mathbf{w}_i^T \mathbf{x}])^2] = E[(\mathbf{w}_i^T \mathbf{R}_x \mathbf{w}_i)] = \lambda_i$$

Celem PCA jest znalezienie wektorów $\mathbf{w}_1, \dots, \mathbf{w}_K$, które maksymalizują wariancję przy zachowaniu ortogonalności

$$\mathbf{w}_i^T \mathbf{w}_j = 0 \quad \text{dla} \quad j \geq i$$

$$\mathbf{w}_i^T \mathbf{w}_i = 1$$

Algorytm PCA

1. Wyśrodkowanie danych dla każdej cechy j

$$x_{ij} \leftarrow x_{ij} - \frac{1}{n} \sum_{j=1}^n x_{ij}$$

2. Wyznaczenie macierzy kowariancji

$$\mathbf{R}_x = \frac{1}{n} \mathbf{X} \mathbf{X}^T$$

3. Diagonalizacja za pomocą rozkładu SVD

$$\mathbf{S} = \mathbf{W}^{-1} \mathbf{R}_x \mathbf{W}$$

$\mathbf{W}$ zawiera wektory własne, $\mathbf{S}$ wartości własne na diagonalu

4. Projekcja danych na kierunki wektorów własnych odpowiadającym K największym wartościom własnym

$$\mathbf{y}_i = \mathbf{W}_K \mathbf{x}_i$$

Rekonstrukcja

Rekonstrukcja obrazu wejściowego

$$\hat{\mathbf{x}} = \mathbf{W}^T \mathbf{y}$$

PCA minimalizuje błąd rekonstrukcji

$$E_r = E[\|\mathbf{x} - \hat{\mathbf{x}}\|^2]$$

dla k składowych PCA

$$E_r = \sum_{i=k+1}^N \lambda_i$$

Minimalizacja E_r odpowiada maksymalizacji $\sum_{i=1}^K \lambda_i$

- Gdy zmienne są skorelowane to znajomość tylko części zmiennych pozwala określić pozostałe
- Względny wkład poszczególnych składowych

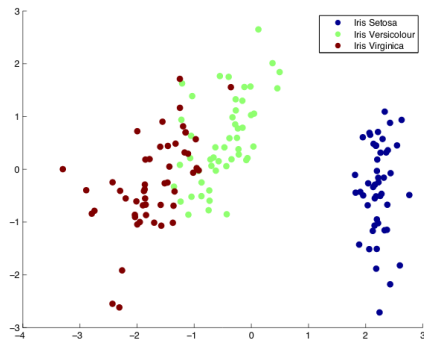
$$m_i = \frac{\lambda_i}{\sum \lambda_j}$$

- Standardowe metody wyznaczania wektorów własnych macierzy $\mathbf{R}_x$ dla dużych wymiarów są zbyt złożone

Metody adaptacyjne wyznaczania składowych głównych (np. reguła Oji)

- nie wymagają obliczania macierzy $\mathbf{R}_x$
- mogą być stosowane w metodach on-line, gdy nie mamy dostępu do wszystkich danych jednocześnie

Przykład: PCA na danych Iris

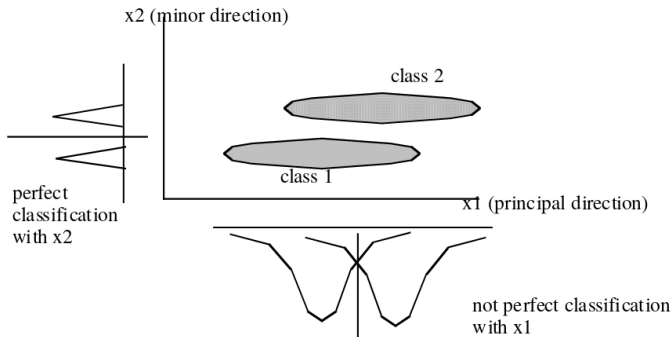


$$\lambda_1 = 2.8914 \quad \lambda_2 = 0.9151 \quad \lambda_3 = 0.1464 \quad \lambda_4 = 0.0205$$

PCA często używane do przetworzenia danych przed treningiem algorytmów uczenia maszynowego

Rys: Boedecker, Machine Learning Summer 2015

PCA i klasyfikacja



Projekcja PCA nie gwarantuje uzyskania najlepszej dyskryminacji

MDS i PCA

- MDS wynik zależny od punktu startowego, stosowany wielokrotny start
- PCA może zostać użyte jako punkt startowy dla MDS. Nieistotne składowe są usuwane i nie mają wpływu na wizualizację MDS.
- PCA jest linowe, nie zachowuje odległości pomiędzy obiektami
- PCA wymaga utworzenia macierzy korelacji $d \times d$, koszt $O(nd^2)$, diagonalizacja tej macierzy $O(d^3)$
- MDS wymaga obliczenia macierzy odległości $O(n^2d)$ i minimalizacji $2n - 3$ parametrów

Trening pierwszego kierunku

Estymator pierwszego kierunku

$$y_1 = \mathbf{w}^T \mathbf{x} = \sum_{j=0}^N w_{1j} x_j$$

za pomocą reguły Oja

$$\mathbf{w}_1(k+1) = \mathbf{w}_1(k) + \eta(k) y_1(k) (\mathbf{x}(k) - \mathbf{w}_1(k) y_1(k))$$

Dobór $\eta(k)$ jest istotny - powinien maleć z czasem, np.:

$$\eta(k) = \frac{\eta(0)}{k^\gamma} \quad \text{gdzie} \quad 0.5 \leq \gamma \leq 1$$

Sieci neuronowe uczone regułą Hebba

W warstwie składającej się z neuronów liniowych uczonych regułą Hebba wszystkie neurony będą zbiegały do kierunku maksymalizującego wariancję.

Mechanizmy konkurencji wymuszające różnorodność rozwiązań

- WTA, zwycięzca bierze wszystko, tylko neuron z najsilniejszą aktywacją podlega uczeniu
- WTM, grupa k najbardziej wzbudzonych neuronów podlega adaptacji
- wymuszenie ortogonalności między wektorami wag, uogólniona reguła Hebba (GHA)
- dodanie połączeń ze sprzężeniem zwrotnym wpływającymi na pobudzenia sąsiednich neuronów, np. APEX

Uogólniony algorytm hebbowski (GHA)

Jednowarstwowa sieć liniowa (Sanger, Oja, 1989)

$$y_i = \mathbf{w}_i^T \mathbf{x}$$

Reguła uczenia Sangera z czynnikiem ortogonalizującym Grama-Schmidta

$$\Delta \mathbf{w}_i = \eta y_i \left(\mathbf{x}_i - \sum_{k=1}^i y_k \mathbf{w}_k \right)$$

oznaczając

$$\hat{\mathbf{x}}_i(k) = \mathbf{x}_i - \sum_{k=1}^{i-1} y_k \mathbf{w}_k$$

otrzymujemy lokalną regułę Oji dla przeskalowanych wejść $\hat{\mathbf{x}}_i$

$$\Delta \mathbf{w}_i = \eta y_i (\hat{\mathbf{x}}_i - \mathbf{w}_i y_i)$$

Wartości wag $\mathbf{w}_i$ kolejnych neuronów uzyskują wartości kolejnych wartości własnych macierzy kowariancji $\mathbf{R}$ (kolejne składowe główne).

Uczenie anty-hebbowskie

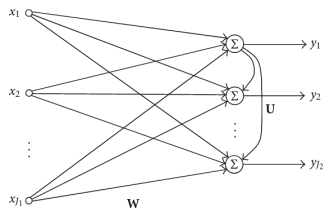
Reguła anty-Hebbowska

$$\Delta w_{ij} = -\eta x_i y_j$$

- znak minus sprawia, że trening prowadzi do znalezienia kierunku **minimalizującego** wariancję wyjścia
- uczenie jednostki liniowej prowadzi do zerowania wyjścia - trening dąży do znalezienia projekcji danych do punktu
- jeżeli nie ma kierunku, wzdłuż którego wariancja wynosi 0 to wagi zbiegają do 0

Adaptive Principal Component Extraction (Diamantaras 1990)

Dla danych wektorów $\mathbf{w}_1, \dots, \mathbf{w}_{i-1}$ realizujących kierunki PCA sieć iteracyjnie wyznacza i -tą składową PCA



$$\mathbf{y} = \mathbf{W}^T \mathbf{x}$$

$$y_i = \mathbf{w}_i^T \mathbf{x} + \mathbf{u}_i^T \mathbf{y}$$

$\mathbf{y} = (y_1, \dots, y_{i-1})^T$ wyjścia pierwszych $i - 1$ neuronów

$\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_{i-1}]^T$ macierz wag pierwszych $i - 1$ neuronów

$\mathbf{u}_i = [u_{1i}, u_{2i}, \dots, u_{(i-1)i}]^T$ wagi połączeń wewnątrz warstwy

Aktualizacja wag APEX

Aktualizacja wag $\mathbf{w}_i$ za pomocą reguły Oji

$$\Delta \mathbf{w}_i = \eta(n) y_i(n) [\mathbf{x}(n) - y_i(n) \mathbf{w}_i]$$

Połączenia między-warstwowe u_{ij} uczone regułą anty-hebbowską odpowiedzialną za ortogonalizację y_i względem poprzednich składowych

$$\Delta \mathbf{u}_i = -\eta(n) y_i(n) [\mathbf{y}(n) + y_i(n) \mathbf{u}_i]$$

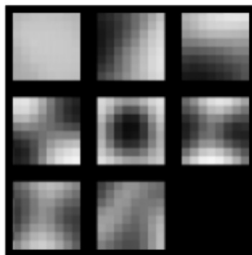
Zastosowanie PCA

- Redukcja wymiarowości
- Wizualizacja danych $nD \rightarrow 2D, 3D$
- Usuwanie szumu
- Usuwanie korelacji z danych
- Kompresja sygnału, np. kompresja obrazów
- Dane po transformacji PCA często używane do treningu algorytmów uczenia maszynowego

Kompresja danych za pomocą PCA

Wektory treningowe to ramki obrazu ($8 \times 8 = 64$ współrzędne)

Przykład: obraz 512×512 z 8 bitami na piksel



ordering
of the neurons:

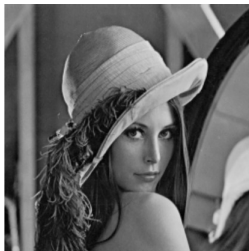
neuron 0	neuron 1	neuron 2
neuron 3	neuron 4	neuron 5
neuron 6	neuron 7	(unused)

Neurony w kolejności zawierającej najwięcej informacji potrzebnej do rekonstrukcji

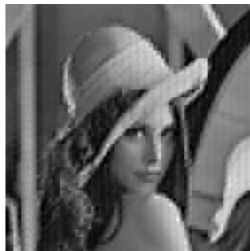
- neuron 0: średni poziom szarości
- neuron 1 i 2: gradient obrazu
- neurony 3-5: pochodna drugiego rzędu obrazu

Przykład: Kompresja obrazu za pomocą APEX

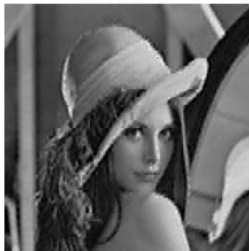
- a) oryginał
- b) 4 PC
- c) 8 PC
- d) 12 PC



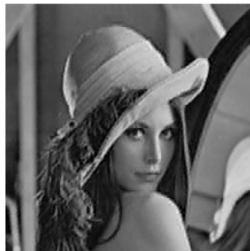
(a)



(b)



(c)



(d)

Sieci Heuaulta-Juttena

- liniowe sieci samoorganizujące oparte o regułę uczenia Hebba uogólnioną do funkcji nieliniowych
- pierwotnie stosowane do ślepej separacji sygnałów, używane również do analizy składników niezależnych (ICA), składników głównych (PCA), dekonwolucji
- struktura sieci rekurencyjna lub jednokierunkowa
- przetwarzanie sygnałów w czasie rzeczywistym
- funkcje nieliniowe bardzo istotne w procesie adaptacji ale nie wpływają na strukturę sieci

Separacja sygnałów

Dla n niezależnych sygnałów źródłowych $s_i(t)$ dostępne są tylko sygnały $x_i(t)$ będące liniową superpozycją sygnałów oryginalnych $s_i(t)$

$$x_i(t) = \sum_{j=1}^n a_{ij} s_j(t)$$

Macierz mieszająca

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

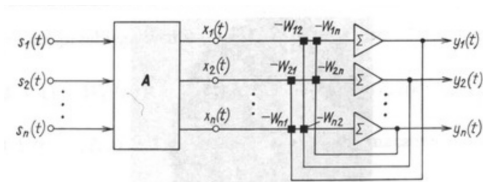
W notacji macierzowej

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$$

Ślepa separacja źródeł (BSS): zadanie odzyskania oryginalnych komponentów sygnału $\mathbf{s}(t)$ przy nieznanym $\mathbf{A}$ jak i $\mathbf{s}(t)$.

Przykład: *cocktail party problem*

Sieci rekurencyjna Heuaulta-Juttana



Działanie realizowane przez sieć

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$$

$$\mathbf{y}(t) = \mathbf{x}(t) - \mathbf{W}\mathbf{y}(t)$$

gdzie macierz wag (w oryginalnej wersji sieci)

$$\mathbf{W} = \begin{bmatrix} 0 & w_{12} & \cdots & w_{1n} \\ w_{21} & 0 & \cdots & w_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ w_{n1} & w_{n2} & \cdots & 0 \end{bmatrix}$$

Szukamy rozwiązania

$$\mathbf{y}(t) = (1 + \mathbf{W})^{-1} \mathbf{x}(t)$$

które odtwarza sygnał pierwotny $\mathbf{s}(t)$.

Warunki:

- odtwarzanie z dokładnością do skali

$$\mathbf{y}(t) = \mathbf{D}\mathbf{s}(t)$$

gdzie $\mathbf{D}$ jest macierzą diagonalną

- odtwarzanie z dokładnością do kolejności sygnałów

$$\mathbf{y}(t) = \mathbf{P}\mathbf{s}(t)$$

gdzie $\mathbf{P}$ jest macierzą permutacji

Dla małej liczby kanałów ($n = 2$) istnieje rozwiązanie analityczne.

Dla większej liczby kanałów stosuje się metodę adaptacyjną.

Reguła Heraulta-Juttana

Przy założeniu statystycznej niezależności sygnałów można wyprowadzić regułę uczenia tożsamą z regułą Hebba

$$\frac{dW_{ij}}{dt} = \eta y_i(t) y_j(t)$$

Zbieżność procesu uczącego wymaga aby w stanie ustalonym $E[\frac{dW_{ij}}{dt}] = 0$ stąd sygnały $y_i(t)$ i $y_j(t)$ nie mogą być skorelowane $E[(y_i - E[y_i])(y_j - E[y_j])] = 0$

Reguła Heraulta-Juttana

$$\frac{dW_{ij}}{dt} = \eta(t) f(y_i(t)) g(y_j(t))$$

$f()$ i $g()$ to funkcje nieparzyste, które przy założeniu statystycznej niezależności sygnałów zapewniają $E[f(y_i(t)) g(y_j(t))] = 0$, co jest warunkiem zbieżności algorytmu uczenia.

Przykłady: $f(x) = x^3, x^5, g(x) = \tanh(x), \arctan(x), x, \operatorname{sgn}(x)$

Zmodyfikowany algorytm Cichockiego

Reguła Heuaulta-Juttena skuteczna tylko dla sygnałów $s_i(t)$ o niewielkim zróżnicowaniu amplitudy (dla względnej różnicy amplitud $< 10^2$). Sygnały o małej amplitudzie nie są separowane.

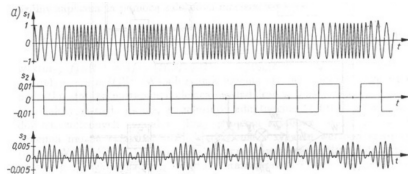
Modyfikacja Cichockiego

$$\frac{d\mathbf{W}}{dt} = \eta(t) [\mathbf{f}(\mathbf{y}(t))[\mathbf{g}(\mathbf{y}(t))]^T - \mathbf{1}]$$

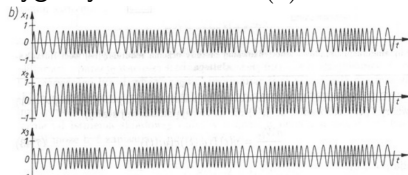
gdzie $\mathbf{y}(t) = (\mathbf{1} + \mathbf{W})^{-1}\mathbf{x}(t)$

- wprowadza sprzężenia zwrotne własne neuronu $w_{ii} \neq 0$ powodujące samonormalizację sygnałów wyjściowych $y_i(t)$
- w stanie ustalonym $E[\frac{d\mathbf{W}}{dt}] = 0$ stąd $E[(f(y_i)g(y_j))] = 0$
- separacja możliwa nawet dla sygnałów o różnicy amplitud rzędu 10^{10}

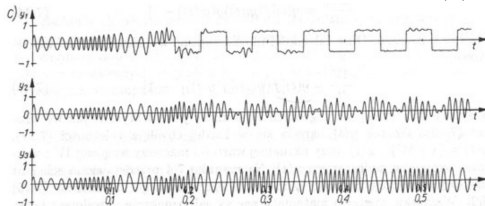
Sygnały oryginalne $s_i(t)$



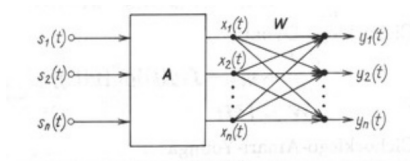
Sygnały zmieszane $x_i(t)$



Proces wydzielania sygnałów $y_i(t)$
odtworzących sygnały źródłowe $s_i(t)$



Sieć jednokierunkowa



$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

macierz $\mathbf{W} \in \mathbb{R}^{n \times n}$ jest macierzą pełną.

Sieć jednokierunkowa jest równoważna sieci ze sprzężeniami zwrotnymi gdy $\mathbf{W} = (\hat{\mathbf{W}} + \mathbf{1})^{-1}$ w której $\hat{\mathbf{W}}$ to macierz wag sieci rekurencyjnej

Zmodyfikowana reguła adaptacyjna Cichockiego

$$\frac{d\mathbf{W}}{dt} = \eta(t) \mathbf{W} [\mathbf{1} - \mathbf{f}(\mathbf{y}(t)) [\mathbf{g}(\mathbf{y}(t))]^T] \mathbf{W}$$

Sieci dynamiczne

Pamięć asocjacyjna

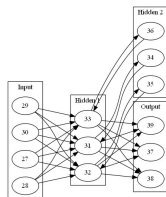
Plan

- **Sieci dynamiczne:** sieci ze sprzężeniami zwrotnymi, stan sieci w chwili t zależy od stanu w chwili $t - 1$
- Modele pamięci asocjacyjnej (autoasocjacyjnej i heteroasocjacyjnej)
- Model Hopfielda, sieć Hamminga, BAM

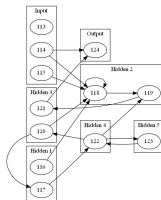
Sieci ze sprzężeniami zwrotnymi

Sieci rekurencyjne

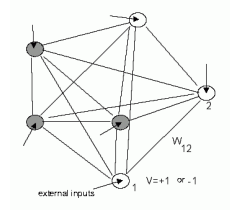
- połączenia tworzą cykle w grafie połączeń
- skomplikowana dynamika - trudna do analizy
- zbliżone do sieci biologicznych, w układach biologicznych neurony mają silne sprzężenia zwrotne



Neural Network



Neural Network



Dynamika sieci: zmiana stanu jednego neuronu przenosi się za pomocą sprzężeń zwrotnych na całą sieć, tworząc nowe stany pośrednie

Sieci ze sprzężeniami zwrotnymi

Najprostsze modele sieci z rekurencją:

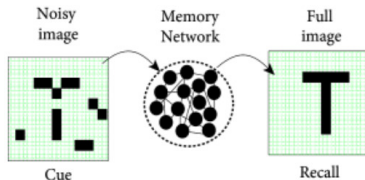
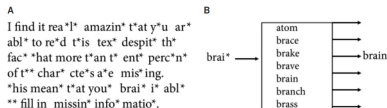
- sieci Hopfielda,
- sieć Hamminga.
- BAM
- Restricted Boltzman Machines (RBM)

Modele złożone:

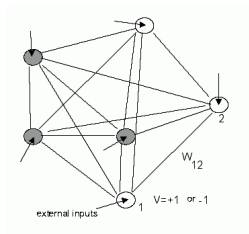
- sieć Elmana i inne o uproszczonej strukturze rekurencji
- RTRN - Real Time Recurrent Network, przetwarzająca sygnały w czasie rzeczywistym
- Głębokie sieci rekurencyjne (RNN),
LSTM, Long Short Term Memory

Model pamięci autoasocjacyjnej

Pamięć skojarzeniowa (asocjacyjna) - odtwarza zapamiętane wzorce zbliżone do sygnału wejściowego. Pamięć adresowana kontekstowo (CAM)



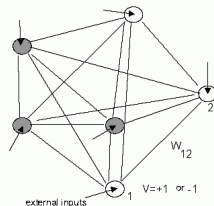
Model pamięci autoasocjacyjnej Hopfielda:



- realizuje pamięć skojarzeniową z binarnymi neuronami progowymi (dyskretne stany neuronów $y_i = \pm 1$)
- wszystkie neurony są ze sobą wzajemnie połączone, brak warstw, każdy neuron jest wejściem i wyjściem
- model stosowany do zrozumienia działania ludzkiej pamięci

Model Hopfielda (1982, 1984)

$$y_i(t+1) = \text{sgn} \left(\sum_j w_{ij} y_j(t) \right)$$



- dyskretne stany neuronów $y \in \{-1, +1\}$, wagi $w_{ij} \in \mathbb{R}$
- zmiany stanów następują w dyskretnych jednostkach czasu
- macierz wag połączeń jest symetryczna, $w_{ij} = w_{ji}$
- brak połączeń zwrotnych do tego samego neuronu $w_{ii} = 0$
- dwa tryby pracy:
 - uczenie - zapamiętywanie wzorców
 - odtwarzanie wzorców przy ustalonych wagach

Symetria wag upraszcza analizę sieci i pozwala wprowadzić funkcję energii, jednak jest nierealistyczna z biologicznego punktu widzenia. Blżej biologii są realizacje sieci o ciągłych funkcjach aktywacji, w których zmiany aktywacji następują w sposób ciągły.

Dynamika odtwarzania wzorców

- dla ustalonych wag $\mathbf{W}$ w chwili $t = 0$ doprowadzamy sygnał wejściowy $\mathbf{y}(0) = \mathbf{x}$
- w kolejnym kroku t aktualizowane są stany neuronów

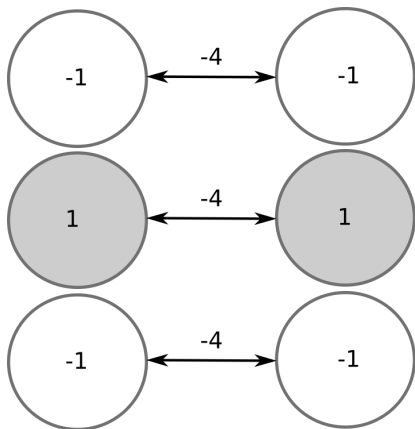
$$y_i(t+1) = \operatorname{sgn} \left(\sum_j w_{ij} y_j(t) \right) \quad i = 1, \dots, n$$

- proces asynchroniczny - w chwili t aktualizowane jest wyjście pojedynczego (losowo wybranego) neuronu, jego wyjście zależy od stanu neuronów z poprzedniego kroku
- **stan stacjonarny** - stan, w którym proces odtwarzania zatrzymuje się (atraktor), minimum lokalne funkcji energii, brak zmian wyjść przy pełnym cyklu

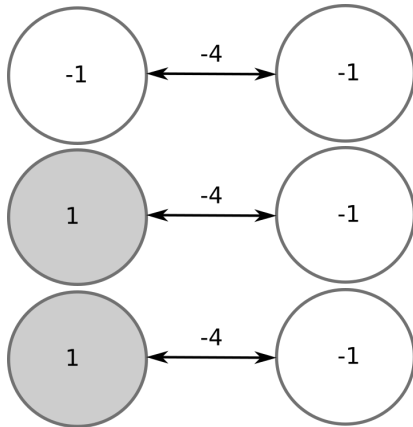
$$\mathbf{y}(t+1) = \mathbf{y}(t)$$

- asynchroniczność minimalizuje występowanie cykli i sieć zbiega do atraktora punktowego
- autoasocjacja - wyjście odtwarza pewien wzorec wejściowy $\mathbf{x}$

Zmiany asynchroniczne i synchroniczne



Zmiana synchroniczna,
oba neurony zmieniają stan,
nieskończone oscylacje



Zmiana asynchroniczna,
aktualizacja pojedynczego
losowego neuronu,
sieć zbiega do stabilnego wzorca

Źródło:

<https://julien-vitay.net/lecturenotes-neurocomputing/4-neurocomputing/2-Hopfield.html>

Minimalizacja energii

Dla sieci o symetrycznych wagach taka dynamika prowadzi do minimalizacji funkcji energii

$$E = - \sum_{i,j:i < j} w_{ij} y_i y_j - \sum_i b_i y_i$$

gdzie b_i wyraz wolny ($b_i = w_{0i}$ można włączyć do wektora wag)

Zmiana energii w czasie odtwarzania jest ujemna (energia nie rośnie)

$$\Delta E = -\Delta y_i \sum_j w_{ij} y_j = -\Delta y_i l_i$$

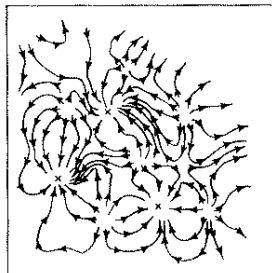
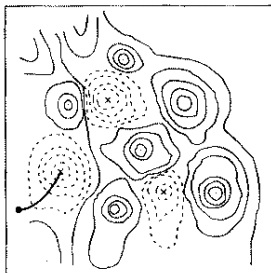
- jeżeli $l_i \geq 0$ to y_i osiąga stan $+1$ i $\Delta y_i \geq 0$
- jeśli $l_i < 0$ to y_i osiąga stan -1 i $\Delta y_i \leq 0$

Każda zmiana stanu neuronu może tylko obniżyć energię sieci

Atraktory

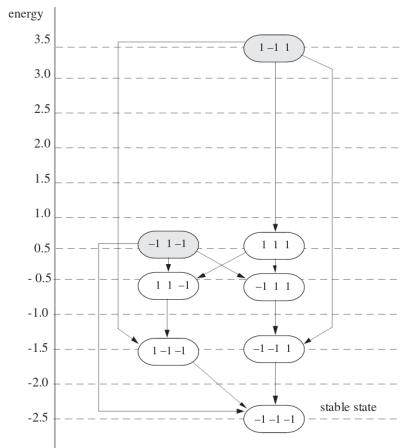
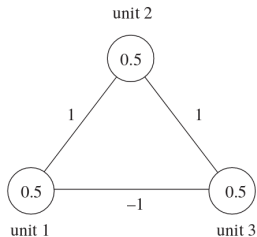
Dynamika: ruch po hiperpowierzchni energii, zależnej od potencjałów neuronów, aż do osiągnięcia lokalnego minimum na takiej powierzchni.

Portret fazowy układu dynamicznego sieci rekurencyjnej o połączeniach symetrycznych: linie ekwipotencjalne i kierunki zmian potencjału w trakcie uczenia



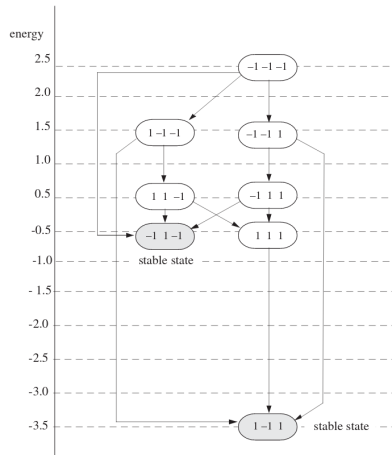
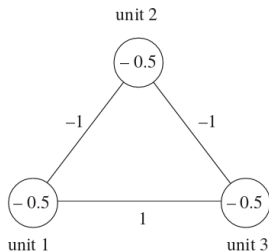
Jeśli y_i dyskretne $\{-1, +1\}$ to ruch po wierzchołkach hiperkostki

Przykład: 3 neurony



$$E = -y_1y_2 - y_2y_3 + y_1y_3 - 0.5y_1 - 0.5y_2 - 0.5y_3$$

Przykład: 3 neurony



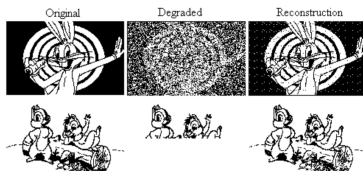
$$E = y_1 y_2 + y_2 y_3 - y_1 y_3 + 0.5 y_1 + 0.5 y_2 - 0.5 y_3$$

Zastosowanie praktyczne

Pamięć asocjacyjna

- **odtworzenie wzorców**, przy prezentacji wektora wejściowego x_i odpowiedzią sieci będzie jeden z zapamiętanych wzorców $x_1, \dots, x_p$
- **korekcja**, odtwarzany jest wzorek najbardziej „podobny” do sygnału wejściowego.
- **uzupełnianie wzorca**: prezentowany jest fragment wzorca (ciąg pierwszych bitów) a sieć odtwarza pozostałe (stany początkowych bitów się nie zmieniają)

Odtwarzanie i rekonstrukcja obrazów



Źródło:

<https://www.cs.cmu.edu/~bhiksha/courses/deeplearning/Fall.2015/slides/lec14.hopfield.pdf>

Zastosowanie praktyczne

Rozwiązywanie problemów optymalizacyjnych - odpowiednia konstrukcja sieci i funkcji energii pozwala rozwiązywać problemy optymalizacyjne, także NP-trudne

- problem komiwojażera,
- gospodarowanie zasobami,
- optymalizacja połączeń w centralach telefonicznych,
- rozmieszczanie układów scalonych,
- tworzenie rozkładów jazdy,
- rozwiązywanie problemów programowania liniowego i nieliniowego

W fizyce: modelowanie szkła spinowego

W neuronaukach: modele hipokampa, pamięć epizodyczna

Pamięć asocjacyjna

- pamięć **autoasocjacyjna** - skojarzenie tego samego obiektu (model Hopfielda)
- pamięć **heteroasocjacyjna** - skojarzone są dwa różne obiekty (dla **a** sieć odpowiada **b**), np. BAM, sieć Hamminga

Pamięć asocjacyjna vs. RAM

- **adresowanie kontekstowe** - odtworzony wzorzec uzyskiwany jest na podstawie odległości (podobieństwa) od zapamiętanego wzorca
- pamięć rozproszona - brak wyraźnie określonego miejsca przechowywania wzorca
- lokalne uszkodzenie nie powinno całkowicie blokować możliwości odczytu

Odległość Hamminga

Pamięć asocjacyjna będzie odtwarzać wektory zapamiętane najbliższe względem wektora wejściowego - dla stanów binarnych zgodnie z odległością Hamminga

Odległość Hamminga - liczba różniących się bitów

- dla wielkości binarnych $x_i, y_i = \{0, 1\}$

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n [x_i (1 - y_i) + (1 - x_i) y_i]$$

- dla wartości bipolarnych $x_i, y_i = \{+1, -1\}$

$$(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \left(n - \sum_{j=1}^n x_i y_j \right)$$

Uczenie pamięci asocjacyjnej

Dla pojedynczego wzorca $\mathbf{x}$ sieć powinna odtwarzać sygnał wejściowy

$$x_i = \operatorname{sgn} \left(\sum_{j=1}^n w_{ij} x_j \right)$$

wystarczy zażądać by:

$$w_{ij} \sim x_i x_j; \quad \text{np. } w_{ij} = \frac{1}{n} x_i x_j \quad \text{reguła Hebb'a}$$

$$\frac{1}{n} \left(\sum_{j=1}^n x_i x_j x_j \right) = \frac{1}{n} \left(\sum_{j=1}^n x_i x_j^2 \right) = \frac{1}{n} \left(\sum_{j=1}^n x_i \right) = x_i$$

Uczenie pamięci asocjacyjnej

Dla wielu wzorców $\mathbf{x}^{(i)}, i = 1, \dots, p$ korzystamy z reguły Hebba uśredniając:

$$w_{ij} = \frac{1}{n} \sum_{k=1}^p x_i^{(k)} x_j^{(k)}$$

W praktyce stosuje się często metodę iteracyjną.

Adaptacja przy prezentacji k -tego wektora $\mathbf{x}^{(k)}$

$$\begin{aligned} \Delta w_{ij} &= x_i^{(k)} x_j^{(k)} & \text{dla } i, j = 1, \dots, n \\ \Delta b_i &= x_i^{(k)} & \text{dla } i = 1, \dots, n \end{aligned}$$

Przykłady działania



odtworzenie niekompletnych lub
zazumionych danych

Stabilność pamięci

Stabilność działania wymaga

$$\operatorname{sgn} \left(\sum_{j=0}^N w_{ij} x_j^{(l)} \right) = \operatorname{sgn} \left(\frac{1}{N} \sum_{j=0}^N \sum_{k=1}^p x_i^{(k)} x_j^{(k)} x_j^{(l)} \right) = x_i^{(l)}$$

co można zapisać

$$\operatorname{sgn} \left(x_i^{(l)} + \frac{1}{N} \sum_{j=0}^N \sum_{k \neq l} x_i^{(k)} x_j^{(k)} x_j^{(l)} \right) = \operatorname{sgn} \left(x_i^{(l)} + C \right)$$

gdzie C nazywamy **przesłuchem**

Stabilność osiągamy gdy składnik przesłuchu C jest na tyle mały aby nie zmienić znaku $x_i^{(l)}$, tzn. $|C| < |x_i^{(l)}|$

Pojemność modelu

- Pojemność określa maksymalną liczbę odtwarzanych wzorców przy określonym poziomie błędu
- 2^N możliwych stanów sieci binarnej złożonej z N neuronów
- gdy ilość wzorców przekracza pojemność to sieć uogólnia, minima lokalne nie odwzorowują wówczas dokładnie wzorców wejściowych
- zbyt wiele wzorców $\Rightarrow$ chaos, zapominanie, zwiększając liczbę wzorców rośnie prawdopodobieństwo przekłamań
- **pojemność pamięci:** dla uczenia regułą Hebba liczba poprawnie pamiętanych losowych wzorców przy prawdopodobieństwie błędu 0.37% wynosi $0.138 \cdot n$ wzorców (czyli $0.138 \cdot n^2$ bitów)
- mała efektywność, bo ilość potrzebnych wag $O(n^2)$ o wartościach rzeczywistych
- metoda rzutowania pozwala uzyskać pojemność maksymalną $n - 1$ (dla wzorców niezależnych liniowo)

Stany fałszywe

W sieci mogą pojawiać się stany fałszywe lub przekłamania pamięci

- funkcja energii jest symetryczna względem polaryzacji (negatywy stanów o tej samej energii)
- występuje mieszanie różnych składowych zapamiętanych wzorców i tworzenie stabilnego zmieszanego stanu
- przy przepełnieniu pamięci powstają pośrednie minima lokalne nie odpowiadające żadnemu wzorcowi zapamiętanemu

Metoda rzutowania

Zakładamy, że każdy wzorzec $\mathbf{x}$ podany na wejście generuje natychmiastowo na wyjściu sieci stan ustalony $\mathbf{x}$

$$\mathbf{W} \cdot \mathbf{X} = \mathbf{X} \quad \Rightarrow \quad \mathbf{W} = \mathbf{X} \cdot \mathbf{X}^+$$

gdzie $\mathbf{X} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p)}]$

Dla $\mathbf{x}^{(i)}$ liniowo niezależnych możemy zastąpić pseudoinwersję macierzy $n \times p$ inwersją macierzy $p \times p$

$$\mathbf{W} = \mathbf{X} \cdot \mathbf{X}^+ = \mathbf{X} (\mathbf{X}^T \cdot \mathbf{X})^{-1} \mathbf{X}^T$$

Możliwe też wyznaczenie $\mathbf{W}$ w sposób iteracyjny podczas jednokrotnej prezentacji wzorców uczących $i = 1, \dots, p$ bez konieczności wyznaczania macierzy odwrotnej

$$\mathbf{W}^{(i)} = \mathbf{W}^{(i-1)} + \frac{\left(\mathbf{W}^{(i-1)} \mathbf{x}^{(i)} - \mathbf{x}^{(i)} \right) \left(\mathbf{W}^{(i-1)} \mathbf{x}^{(i)} - \mathbf{x}^{(i)} \right)^T}{\left(\mathbf{x}^{(i)} \right)^T \mathbf{x}^{(i)} - \left(\mathbf{x}^{(i)} \right)^T \mathbf{W}^{(i-1)} \mathbf{x}^{(i)}}$$

gdzie $\mathbf{W}^{(0)} = \mathbf{0}$

Metoda rzutowania Δ

Gradientowa odmiana metody rzutowania

$$\mathbf{W} \leftarrow \mathbf{W} + \frac{\eta}{n} \left[\mathbf{x}^{(i)} - \mathbf{W}\mathbf{x}^{(i)} \right] (\mathbf{x}^{(i)})^T$$

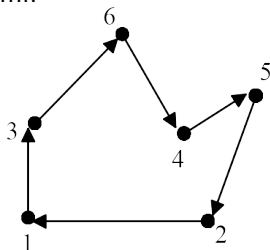
- η stała uczenia zazwyczaj z przedziału $[0.7 - 0.9]$
- wymaga wielokrotnej prezentacji wzorców aż do uzyskania zbieżności, np. gdy zmiana wartości wag nie przekracza pewnego progu tolerancji ϵ
- metoda rzutowania i rzutowania Δ zapewniają większą pojemność w stosunku do uczenia reguła Hebb'a, maksymalna liczba zapamiętanych wzorców $n - 1$

Zastosowanie w problemach optymalizacyjnych

- jeżeli optymalizowana funkcja kosztu (z warunkami) może zostać zapisana w postaci funkcji energii Hopfielda to istnieje sieć Hopfielda, której stany stabilne odpowiadają rozwiązaniu problemu optymalizacji
- dynamika odtwarzania minimalizuje energię powodując zmniejszenie wartości funkcji kosztu przy zachowaniu warunków optymalizacji
- warunki optymalizacji są „zaszyte” w wagach sieci
- problemy NP-trudne - np. problem komiwojażera

Optymalizacja - problem komiwojażera

Problem komiwojażera: znaleźć najkrótszą drogą pomiędzy N miastami.



		Kolejność					
		1	2	3	4	5	6
M i a s t o	1	1	0	0	0	0	0
	2	0	0	0	0	0	1
	3	0	1	0	0	0	0
	4	0	0	0	1	0	0
	5	0	0	0	0	1	0
	6	0	0	1	0	0	0

Problem NP-trudny, $\frac{N!}{2^N}$ wszystkich możliwości

- sieć zawierająca N^2 neuronów unipolarnych ciągłych
- stan neuronu $y_{i\alpha} = 1$ oznacza, że miasto i zostało odwiedzone w kolejności α
- numer miasta $i, j = 1, 2, \dots, N$,
kolejność odwiedzin $\alpha, \beta = 1, \dots, N$

Przykład: problem komiwojażera

Ogólna postać funkcji energii

$$E = -\frac{1}{2} \sum_{i \neq k} \sum_{\alpha \neq \beta} w_{i\alpha, k\beta} y_{i\alpha} y_{k\beta}$$

Jak dobrać W ?

Dobór wag

Kara za powtórne odwiedzenie miasta

$$E_1 = \frac{A}{2} \sum_i \sum_{\alpha \neq \beta} y_{i\alpha} y_{i\beta}$$

Kara za rozdwojenie komiwojażera (odwiedzenie dwóch miejsc w tym samym czasie)

$$E_2 = \frac{B}{2} \sum_{i \neq k} \sum_{\alpha} y_{i\alpha} y_{k\alpha}$$

Kara za wzbudzenie większej liczby neuronów niż N

$$E_3 = \frac{C}{2} \left(\sum_{i,\alpha} y_{i\alpha} - N \right)^2$$

Czynnik promujący najkrótszą trasę

$$E_4 = \frac{D}{2} \sum_{i \neq k} \sum_{\alpha} d_{ik} y_{i\alpha} (y_{k\alpha-1} + y_{k\alpha+1})$$

Dobór wag

Całkowita energia

$$E = E_1 + E_2 + E_3 + E_4$$

stąd wagi sieci

$$w_{i\alpha, k\beta} = -A(1 - \delta_{\alpha\beta})\delta_{ik} - B(1 - \delta_{ik})\delta_{\alpha\beta} - C \\ - D\delta_{ik}(1 - \delta_{ik})(\delta_{\alpha-1, \beta} + \delta_{\alpha+1, \beta})$$

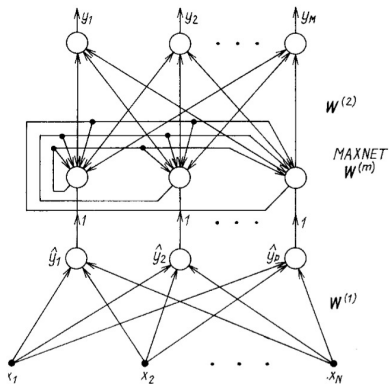
oraz wartości progowe CN

Delta Kroneckera $\delta_{ij} = 1$ dla $i = j$ oraz $\delta_{ij} = 0$ dla $i \neq j$

Wartości A, B, C, D dobierane heurystycznie, np.

$A = B = C = 500$ i $C = 200$

Sieć Hamminga

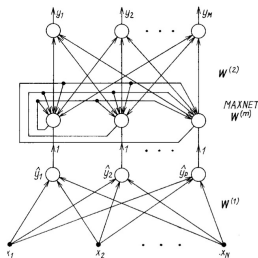


Sieć Hamminga

- 3 warstwowa sieć - rozszerzenie sieci Hopfielda
- realizuje pamięć heteroasocjacyjną:
wejście $\mathbf{x} \in \{-1, +1\}^N$, wyjścia $\mathbf{y} \in \{-1, +1\}^M$
- wyjściem jest zapamiętany wzorzec o najmniejszej odległości Hamminga względem wektora wejściowego
- warstwa MAXNET (warstwa pamięci) ze sprzężeniami zwrotnymi, połączenia każdy z każdym,
 - rozmiar równy liczbie wzorców do zapamiętania p , pojedynczy neuron jest prototypem zapamiętanego wzorca
 - w drodze iteracyjnego procesu wybierany jest neuron zwycięzca (WTA) reagujący na wzorzec wejściowy

Działanie sieci Hamminga

1. podanie stanu $\mathbf{x}$ na wejście, zainicjowanie warstwy MAXNET
2. proces iteracyjny w warstwie MAXNET trwa dopóki wszystkie jednostki oprócz jednej osiągną stan 0.
Neuron niezerowy reprezentuje klasę wektora
3. wytworzenie odpowiedzi sieci $\mathbf{y}$ na podstawie sygnału zwycięzcy



Dobór wag

1. wagi warstwy pierwszej odpowiadają sygnałom wejściowym

$$w_{ij}^{(1)} = \frac{x_j^{(i)}}{2}, \quad w_{i0}^{(1)} = \frac{N}{2}, \quad i = 1, \dots, p$$

2. wagi warstwy wyjściowej kodują sygnały wyjściowe

$$w_{ij}^{(3)} = y_j^{(i)}, \quad i = 1, \dots, p$$

3. warstwa MAXNET

- połączenia zwrotne do tych samych neuronów mają charakter pobudzający

$$w_{ii}^{(m)} = 1$$

- połączenia między różnymi neuronami inhibycyjne (hamujące)

$$-\frac{1}{p-1} < w_{ij}^{(m)} < 0$$

gdzie p - liczba wzorców. Wagi powinny się różnić aby zapewnić bezwzględną zbieżność.

Działanie sieci Hamminga

Pierwsza warstwa wyznacza odległość Hamminga od podanego wzorca $\mathbf{x}$ do wszystkich zapamiętanych wzorców wejściowych $\mathbf{x}^{(i)}$

$$y_i^{(1)}(\mathbf{x}) = \sum_{j=1}^N w_{ij}^{(1)} x_j + w_{i0} = \frac{1}{2} \left(N + \sum_{j=1}^N x_j^{(i)} x_j \right)$$

Rekurencyjny proces wyłaniania zwycięzcy warstwy MAXNET

$$y_i(k) = f \left(\sum_{j=1}^p w_{ij}^{(m)} y_j(k-1) \right)$$

gdzie funkcja wyjściowa (ReLU)

$$f(y) = \begin{cases} y & \text{dla } y \geq 0 \\ 0 & \text{dla } y < 0 \end{cases}$$

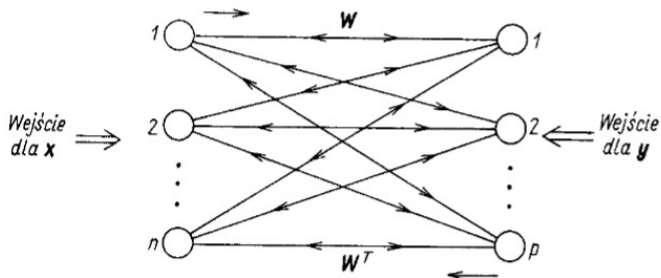
Sygnał zwycięskiego neuronu odtwarza na wyjściu wektor $\mathbf{y}$ najbliższy $\mathbf{x}$ względem odległości Hamminga

Hamming vs Hopfield

- pojemność sieci Hamminga = liczba neuronów MAXNET
- ilość wag $N \times p + p^2 + p \times M$
- dla $N = M$ możliwa realizacja pamięci autoasocjacyjnej
ilość wag $2N \times p + p^2$
- dla sieci Hopfielda ilość wag N^2
- dla $p \ll N$ sieć Hamminga potrzebuje o wiele mniej wag do osiągnięcia takiej samej pojemności
- w sieci Hamminga nie występują stany fałszywe

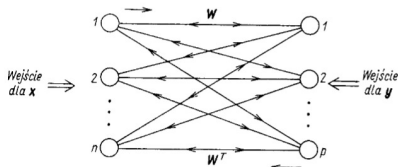
Sieć BAM

Bidirectional Associative Memory (BAM) - uogólnienie sieci Hopfielda na dwuwarstwową sieć rekurencyjną (Kasko)



BAM

- przepływ sygnału w dwóch kierunkach
- praca synchroniczna lub asynchroniczna
- aktywacje $f(x)$ typu skokowego (binarne lub bipolarne)
- $\mathbf{W}$ rzeczywista, niesymetryczna
- koduje pary wektorów $\mathbf{x}_i, \mathbf{y}_i$ - pamięć heteroasocjacyjna



Działanie BAM - zapamiętywanie

Trening: ustalenie wag wartościami macierzy korelacji

$$\mathbf{W} = \sum_{i=1}^p \mathbf{x}_i^T \mathbf{y}_i$$

Dla wejść bipolarnych $x_i, y_j \in \{-1, +1\}$ otrzymujemy uczenie hebbowskie - sygnały o tym samym znaku mają wagę wzmacniającą, o przeciwnym - hamującą

Przykład:

$$\begin{array}{ll} A_1 = (101010) & B_1 = (1100) \\ A_2 = (111000) & B_2 = (1010) \end{array} \Rightarrow \begin{array}{ll} \mathbf{x}_1 = (1, -1, 1, -1, 1, -1) & \mathbf{y}_1 = (1, 1, -1, -1) \\ \mathbf{x}_2 = (1, 1, 1, -1, -1, -1) & \mathbf{y}_2 = (1, -1, 1, -1) \end{array}$$

$$\begin{aligned} \mathbf{W} &= \mathbf{x}_1^T \mathbf{y}_1 + \mathbf{x}_2^T \mathbf{y}_2 = \\ &= \begin{pmatrix} 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & 1 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 & -2 \\ 0 & -2 & 2 & 0 \\ 2 & 0 & 0 & -2 \\ -2 & 0 & 0 & 2 \\ 0 & 2 & -2 & 0 \\ -2 & 0 & 0 & 2 \end{pmatrix} \end{aligned}$$

Działanie BAM - odtwarzanie

Odtwarzanie sygnału

$$\begin{aligned} f(\mathbf{x}_0 \mathbf{W}) = \mathbf{y}_1 &\rightarrow f(\mathbf{y}_1 \mathbf{W}^T) = \mathbf{x}_1 \rightarrow \dots \\ \dots &\rightarrow f(\mathbf{x}_k \mathbf{W}) = \mathbf{y}_k \rightarrow f(\mathbf{y}_k \mathbf{W}^T) = \mathbf{x}_k \end{aligned}$$

generuje dwa stabilne wzorce $\mathbf{x}_k$ i $\mathbf{y}_k$

Funkcja aktywacji progowa

$$f(x) = \begin{cases} 1, & x > 0 \\ 0, & x < 0 \end{cases}$$

Funkcja energii

$$E_k = -\mathbf{x}_k \mathbf{W} \mathbf{y}_k^T$$

Przykład:

$$\mathbf{A}_1 \mathbf{W} = (4, 2, -2, -4) \rightarrow (1, 1, 0, 0) = \mathbf{B}_1 \quad E_1 = -6$$

$$\mathbf{A}_2 \mathbf{W} = (4, -2, 2, -4) \rightarrow (1, 0, 1, 0) = \mathbf{B}_2 \quad E_2 = -6$$

gdy zaburzymy 1 bit w $\mathbf{A}_2$

$$(0, 1, 1, 0, 0, 0) \mathbf{W} = (2, -2, 2, -2) \rightarrow (1, 0, 1, 0) = \mathbf{B}_2 \quad \text{początkowa energia } E = -4$$

Przykład

asynchroniczny BAM

wzorce zapamiętane:

(S,E) (M,V) (G,N)

wejście $10 \times 14 = 140$

wyjście $9 \times 12 = 108$

wzorec początkowy (S,E)

z 40% szumem (99

odwrócone bity)

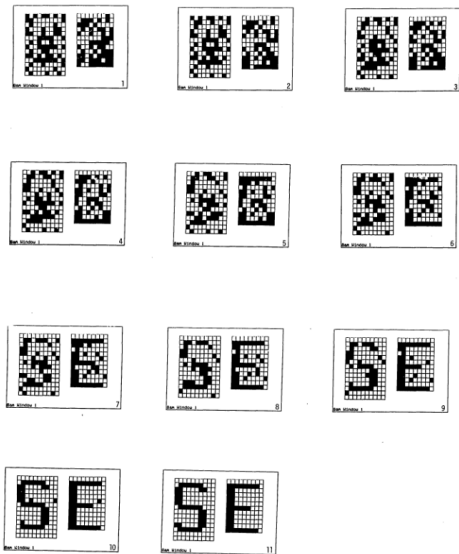


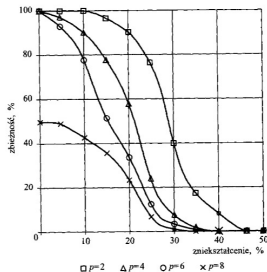
Fig. 1. Asynchronous BAM recall. Approximately six neurons update per snapshot. The associated spatial patterns (S,E), (M,V), and (G,N) are stored. Field F_A contains 140 neurons; F_B , 108. Perfect recall of (S,E) is achieved when recall is initiated with a 40% noise-corrupted version of (S,E).

Kosko (1987) Adaptive bidirectional associative memories"

Pojemność BAM

Pojemność sieci

$$p < \min(n, m)$$



Przykładowe wyniki dla $n = 25$, $m = 9$ i losowych wzorców ze średnią odległością 11

Głębokie sieci neuronowe

Głębokie sieci neuronowe - Plan

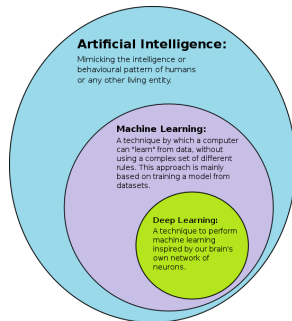
1. Głębokie MLP (DNN)
2. Sieci splotowe (CNN)
3. Sieci rekurencyjne (RNN) i uczenie sekwencji
4. Autoenkodery (AE), Deep belief network (DBN)
5. Generative Adversarial Networks (GAN)

Literatura

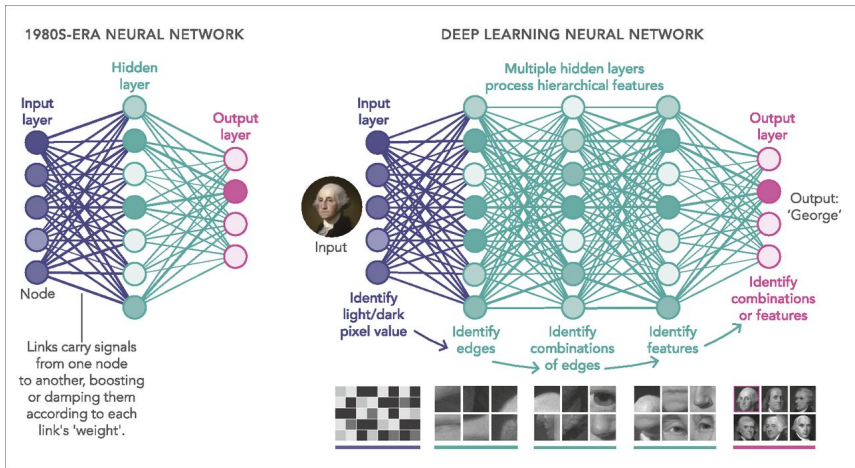
-  Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning*. MIT Press, 2016,
<http://www.deeplearningbook.org>
-  Michael Nielsen, *Neural Networks and Deep Learning*,
<http://neuralnetworksanddeeplearning.com/>
-  Aston Zhang, Zachary C. Lipton, Mu Li, Alexander J. Smola, *Dive into Deep Learning*,
<https://d2l.ai/>
-  Julien Vitay, *Neurocomputing*,
<https://julien-vitay.net/lecturenotes-neurocomputing/>
-  Denny Britz, *Deep Learning Glossary*,
<http://www.wildml.com/deep-learning-glossary/>

Głębokie uczenie to metody uczenia maszynowego

- zbudowane z **wielu warstw realizujących nieliniowe transformacje**
- tworzą **model hierarchiczny** warstwy odwzorowują kolejne poziomy abstrakcji, najniższe warstwy tworzą reprezentują najprostszych cech z (surowego) sygnału wejściowego, wyższe warstwy - generują bardziej ogólne koncepty bazując na relacjach z poprzednich warstw.
- automatycznie uczą się reprezentacji (wykrywają cechy)
- **Transfer learning** - wykorzystanie wiedzy zdobytej przy rozwiązywaniu innego (podobnego) problemu, np. „przeszczep” wag wybranych warstw sieci



Hierarchia reprezentacji

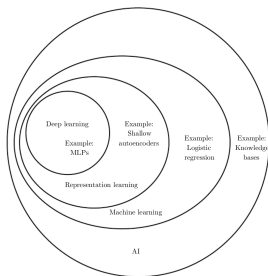


Rys:  What are the limits of deep learning? by Mitchell Waldrop

Representation learning

Metody uczące się reprezentacji

metody uczenia maszynowego posiadające mechanizmy automatycznego tworzenia takiej reprezentacji danych (ekstrakcji cech), która istotnie poprawia zdolność algorytmów w rozwiązywaniu poszczególnych zadań.



Głębokie vs. płytkie uczenie

Krótką (nieformalna) definicja: głębokie uczenie, gdy model ma więcej niż 2 nieliniowości w najkrótszej ścieżce prowadzącej od wejścia do wyjścia

Płytkie modele

- do 2 warstw (nieliniowych transformacji), np. regresja logistyczna, SVM, MLP z jedną warstwą ukrytą
- wymagają odpowiednio przygotowanych cech

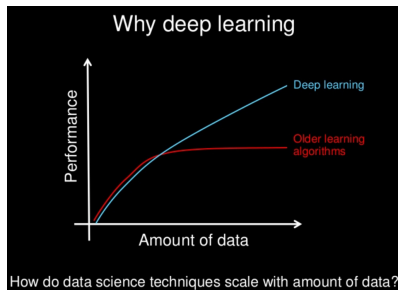
Głębokie modele

- więcej niż 2 warstwy przetwarzania, np. wielowarstwowe MLP, sieci rekurencyjne RNN, sieci splotowe CNN
- model hierarchiczny, wiele transformacji od wejścia do wyjścia
- każda „warstwa” tworzy reprezentację danych na innym poziomie abstrakcji
- możliwość pracy na „surowych” danych → modele end-to-end

Po co nurkować w głąb?

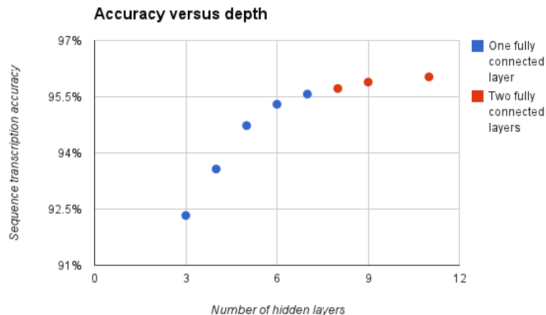
- **Uniwersalny aproksymator:** sieć MLP z jedną warstwą ukrytą jest w stanie aproksymować dowolną funkcję z dowolną dokładnością
- Deep learning nie taki prosty, problemy nie tylko z treningiem ale z interpretacją działania modeli
- Po co więc tworzyć głębsze sieci?
- „Płytkie” modele wymagają o wiele (wykładniczo) więcej neuronów
- „Płytkie” a szerokie modele łatwiej się przetrenowują, znalezienie rozwiązania złożonego jest trudniejsze niż w modelach głębszych
- Złożone problemy wymagają złożonych modeli

Skalowalność wyników



- duża dostępna ilość danych (z etykietami) - Big Data, sieci głębokie wymagają dużej liczby danych
- duża dostępna moc obliczeniowa (GPU, TPU), głębokie uczenie wiąże się z wykonywaniem wielu operacji mnożenia macierzy
- nowe metody pozwalające optymalizować głębokie modele

Głębokość a poprawność klasyfikacji



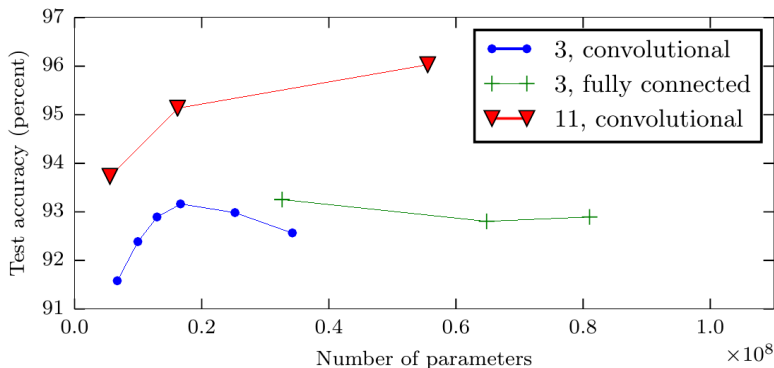
Street View House Numbers dataset (SVHN)
200k numerów domów



Głębokość niezbędna do zbudowania odpowiedniej reprezentacji i osiągnięcia lepszej poprawności

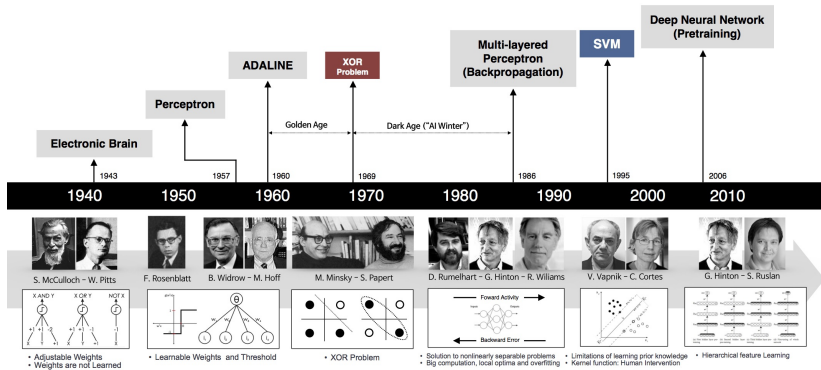
Goodfellow, et al. (2014) Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks

Głębokość ma znaczenie



Zwiększanie liczby parametrów płytkich modeli nie wystarczy do osiągnięcia poprawności modeli głębokich
Głębokość może mieć większe znaczenie od liczby parametrów modelu

Goodfellow, et al. (2014) Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks

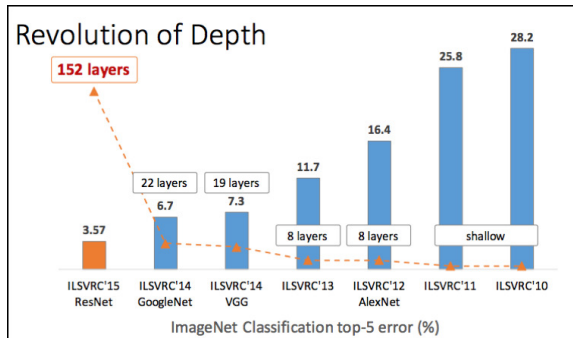


Geoffrey E. Hinton, et.al. „A fast learning algorithm for deep belief nets”, 2006
 R. Salakhutdinov, G. Hinton „Deep boltzmann machines”, 2009

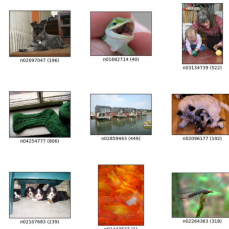
https://beamandrew.github.io/deeplearning/2017/02/23/deep_learning_101_part1.html

State of art results

👉 ImageNet Large Scale Visual Recognition Challenge (ILSVRC)



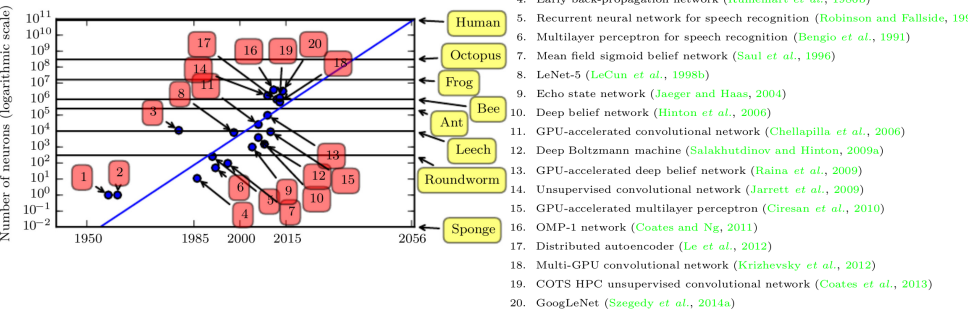
Dane ImageNet 2012:
klasyfikacja obrazów,
trening 1.26 M zdjęć,
1000 kategorii, poziom
człowieka 5%



- klasyfikacja obrazów 👉 Are we there yet?
- rozpoznawanie mowy 👉 WER are we?
- 👉 Browse State-of-the-Art

Rys: 📖 *The revolution of depth*, Armando Vieira

Ilość neuronów w modelach neuronowych



Liczba neuronów podwaja się co 2.5 roku

Rys: Gooffellow, 2016 [1]

Wszędzie AI

- rozpoznawanie obiektów (twarzy, emocji), etykietowanie i lokalizacja obiektów, segmentacja obrazów
- automatyczne tłumaczenie, analiza tekstu, mowy,
- modele języka, chatboty (ChatGPT, BERT)
- przewidywanie wyników wyborów, trzęsień ziemi, sekwencjonowanie białek (AlphaFold)
- sterowanie robotami, samochodami, gry komputerowe (AlphaGo, AlphaStar)
- kompresja sygnałów, rekonstrukcja obrazów, odszumianie sygnałów, kolorowanie czarno-białych filmów, zwiększanie rozdzielczości ...
- generowanie i synteza sygnałów: muzyki, obrazów (Modjourney, DALL-E2), tekstu, pisma ręcznego, mowy, kodu komputerowego,
- AI tworzące AI (Google AutoML), AI rozmawiające z AI (Google Duplex)
- 🖱️ 30 amazing applications of deep learning
- 🖱️ Browse State-of-the-Art

DNN

Głębokie MLP

MLP przypomnienie

Typy neuronów

- liniowe, sigmoidalne, tanh
- funkcje sigmoidalne nasycają się, co utrudnia optymalizację
- ReLU, Maxout

Optymalizacja metodami spadku gradientu lub innymi:

- wsteczna propagacja
- SGD, trening z momentem, RPROP
- metody 2 rzędu: Newtona, Levenberg-Marquarda, gradientów sprzężonych
wymagają obliczenia Hessianu lub Jakobianu
- Adam, AdaGrad, moment Nesterova

Metody regularyzacji:

- L_1 , L_2 , wczesne zatrzymanie
- dropout, batch normalization, gradient clipping, data augmentation,

Funkcje kosztu - przypomnienie

Błąd średniokwadratowy (MSE)

$$E = \frac{1}{N} \sum (f(\mathbf{x}) - y)^2$$

- problemy aproksymacji (wyjścia ciągłe)
- liniowe neurony wyjściowe

Entropia krzyżowa, Cross-Entropy (CE)

$$E = - \sum y \log f(\mathbf{x})$$

- klasyfikacja (wyjścia binarne)
- softmax lub sigmoidalne neurony wyjściowe

Wsteczna propagacja błędów

1. Zainicjuj wagi $w_{ij}^{(k)}$ małymi losowymi wartościami
2. Dopóki nie spełniony warunek stopu wykonuj
 - propagacja sygnału od wejścia do wyjścia

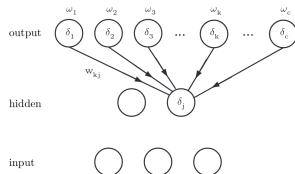
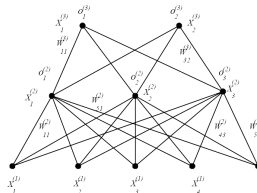
$$o_i^{(K)} = \sigma \left(\sum_j w_{ji}^{(K)} o_j^{(K-1)} \right)$$

- propagacja sygnału błędów od warstwy wyjściowej do wejściowej

$$\delta_i^{(K)} = \sum_j \delta_j^{(K+1)} w_{ij}^{(K)}$$

- aktualizuj wagi

$$\Delta w_{ij}^{(K)} = \eta \delta_j^{(K)} o_i^{(K-1)}$$



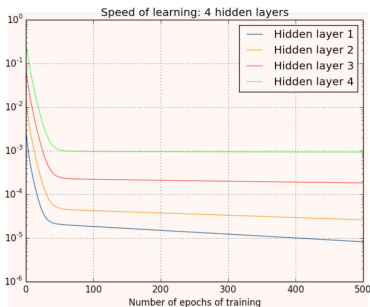
Problemy związane z uczeniem (głębokich) sieci

- **zanikający gradient** - błąd propagowany od warstwy wyjściowej zanika przy niższych warstwach
- funkcje ograniczone (sigmoidalna, tangens hiperboliczny) „nasycają się” i posiadają niezerowy gradient tylko w wąskim przedziale aktywacji bliskim 0
- **przeuczenie** - większa liczba parametrów sprzyja przetrenowaniu, wymagane są techniki regularyzacyjne
- większa ilość parametrów + duże dane → długi czas uczenia i duże wymagania dotyczące pamięci

Zanikający gradient

Przykład:

MNIST data, MLP 784x30x30x30x30x10, regularyzacja $\|\mathbf{w}\|^2$

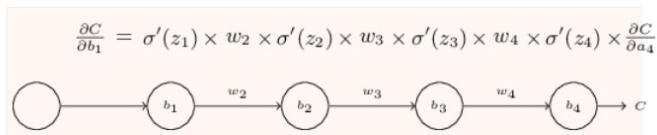


Szybkość uczenia $\|\delta\|$ drastycznie maleje w głębszych warstwach

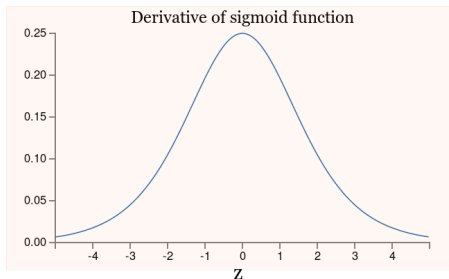
Nielsen, 2016 [2]

Znikający gradient

Przykład: sieć $1 \times 1 \times 1 \times 1$



gdy $|w_j| < 1$ wówczas $|w_j \sigma'(z_j)| < \frac{1}{4}$



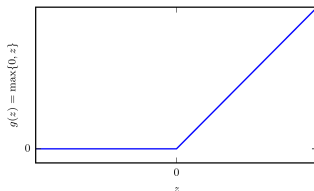
Nielsen, 2016 [2]

Problem niestabilnego gradientu

- gradient w niższych warstwach zależy od wartości wag, aktywacji i gradientów w warstwach wyższych
- **eksplodujący gradient**
gdy $w_i \gg 1$ oraz $z_i \approx 0$ wówczas $|w_i \sigma'(z_i)| > 1$, może to owocować bardzo dużymi gradientami w początkowych warstwach
- znikający i eksplodujący gradient to przypadki szczególne **problemu niestabilnego gradientu** -> wartości gradientów w poszczególnych warstwach mogą znacznie się różnić, warstwy uczą się z różnym tempem

Rectified linear unit (ReLU)

$$g(z) = \max(0, z), \quad g'(z) = \begin{cases} 1 & \text{dla } z > 0 \\ 0 & \text{dla } z < 0 \end{cases}$$



X. Glorot, A. Bordes and Y. Bengio (2011). Deep sparse rectifier neural networks

Korzyści ReLU

- gradient nie znika, gdy jednostka jest aktywna, ograniczenie problemu znikającego gradientu
- fragmentami liniowa, skuteczniejsza optymalizacja metodami gradientowymi
- efektywna obliczeniowo: mnożenie, dodawanie, warunek
- głębokie sieci ReLU nie wymagają pre-treningu (Glorot, 2011)
- rzadka reprezentacja (*sparse representation*), tylko część jednostek jest aktywna w czasie przetwarzania sygnału

Sparse representation

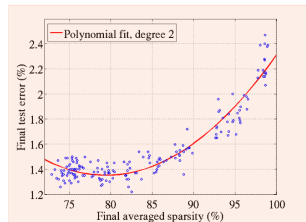
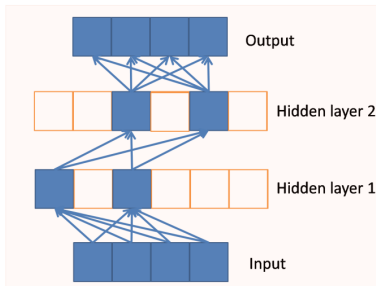


Figure 3: **Influence of final sparsity on accuracy.** 200 randomly initialized deep rectifier networks were trained on MNIST with various L_1 penalties (from 0 to 0.01) to obtain different sparsity levels. Results show that enforcing sparsity of the activation does not hurt final performance until around 85% of true zeros.

Rzadka reprezentacja dodaje nadmiarowość do modelu, zmiana wywołana przez sygnał wejściowy wprowadza zmiany tylko w podzbiorze neuronów

ReLU networks results

Sieci z jednostkami ReLU uzyskują dobre wyniki bez konieczności stosowania pre-treningu, czyli metody ustalenia początkowych wartości wag w procesie wstępnego treningu (zazwyczaj nienadzorowanego)

Neuron	MNIST	CIFAR10	NISTP	NORB
<i>With</i> unsupervised pre-training				
Rectifier	1.20%	49.96%	32.86%	16.46%
Tanh	1.16%	50.79%	35.89%	17.66%
Softplus	1.17%	49.52%	33.27%	19.19%
<i>Without</i> unsupervised pre-training				
Rectifier	1.43%	50.86%	32.64%	16.40%
Tanh	1.57%	52.62%	36.46%	19.29%
Softplus	1.77%	53.20%	35.48%	17.68%

Problemy ReLU

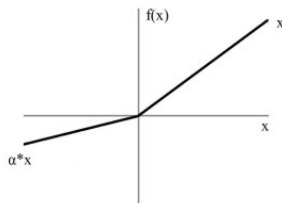
- nie jest różniczkowalne dla $z = 0$
- nie jest symetryczna i wycentrowana w 0
- nieograniczona, może powodować nieograniczony wzrost wartości aktywacji
- **problem umierających ReLU** - gdy neurony osiągają stan permanentnego braku aktywacji $g(z) = 0$ i ich wagi nie są modyfikowane w trakcie uczenia
- druga pochodna równa 0, nie można stosować metod gradientowych 2 rzędu

Uogólnienia ReLU

- **Absolute ReLU** (Jarret, 2009), $g(z) = |z|$ skuteczna w szczególnych zastosowaniach dla obrazów z symetrią
- **Leaky ReLU** (Maas, 2013), wprowadza małą ($\alpha = 0.01$) wartość aktywacji dla $z < 0$

$$g(z) = \begin{cases} z & \text{dla } z > 0 \\ \alpha z & \text{dla } z < 0 \end{cases}$$

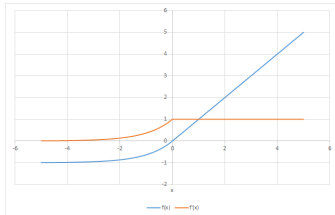
- **Parametric ReLU** odmiana Leaky ReLU, gdzie α podlega optymalizacji w trakcie uczenia



Modyfikacje ReLU

- **Noisy ReLU** dodaje niewielki szum z rozkładu normalnego
- **Exponential linear unit (ELU)** (Clevert, 2015), średnia aktywacja bliższa 0, szybsza zbieżność w stosunku do ReLU

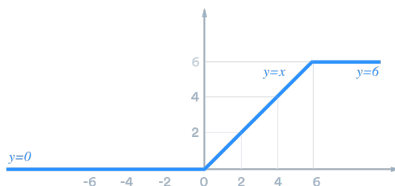
$$g(z) = \begin{cases} z & \text{dla } z > 0 \\ \alpha(e^z - 1) & \text{dla } z < 0 \end{cases} \quad g'(z) = \begin{cases} 1 & z > 0 \\ \alpha e^z, & z < 0 \end{cases}$$



Rys:  ELU as a Neural Networks Activation Function by Sefik

ReLU6

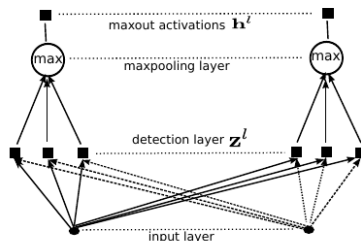
$$g(z) = \begin{cases} 6 & \text{dla } z > 6 \\ z & \text{dla } 0 < z < 6 \\ 0 & \text{dla } z < 0 \end{cases}$$



Górne ograniczenie przyspiesza tworzenie rzadkiej reprezentacji,
dobre wyniki na CIFAR-10 (Krizhevsky)

Maxout

$$g_i(\mathbf{z}) = \max_{j \in G_i} z_j$$



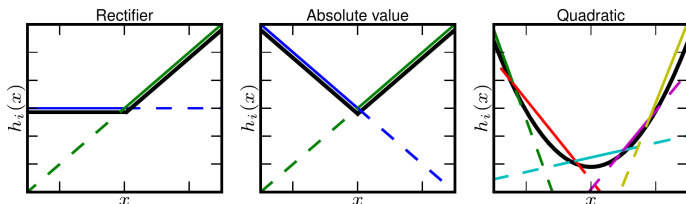
Zwraca wyjście najaktywniejszego neuronu z grupy k neuronów w warstwie, gdzie G_i to zbiór indeksów i -tej grupy

$$g'(z_j) = \begin{cases} 1 & \text{dla } j = \arg \max_i z_i \\ 0 & \text{dla } j \neq \arg \max_i z_i \end{cases}$$

Rys: Paweł Świtojowski, *Investigation of maxout networks for speech recognition*.
Goodfellow, et al, (2013). "Maxout Networks". *JMLR WCP*. 28 (3): 1319–1327.

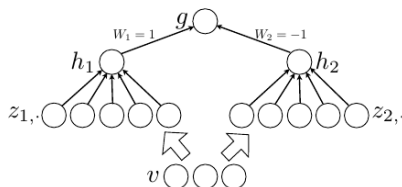
Maxout

- realizuje funkcję wypukłą fragmentami liniowa złożoną z k fragmentów, kształt funkcji jest adaptowany w treningu



- ReLU i PReLU to szczególne przypadki Maxout
- fragmentami liniowa
- utrata rzadkiej reprezentacji aktywacji (większość neuronów ma niezerowe wyjścia). Jednak gradient przybiera rzadką reprezentację, sygnał błędu propagowany tylko przez jednostkę zwycięską.

Sieć Maxout



- złożenie funkcji wypukłych pozwala modelować dowolną funkcję ciągłą
- sieć z neuronami maxout jest uniwersalnym aproksymatorem
- aktywacja zależy od grypy wag, więc wprowadza do modelu redundancję, która pozwala niwelować fenomen „katastrofalnego zapomnienia”, gdy sieć zapomina wyuczone wcześniej wzorce

Rys: Goodfellow, et al, (2013). "Maxout Networks". JMLR WCP. 28 (3): 1319–1327.

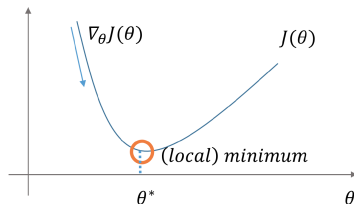
Maxout netowrk on MNIST

METHOD	TEST ERROR
RECTIFIER MLP + DROPOUT (SRI-VASTAVA, 2013)	1.05%
DBM (SALAKHUTDINOV & HINTON, 2009)	0.95%
Maxout MLP + dropout	0.94%
MP-DBM (GOODFELLOW ET AL., 2013)	0.91%
DEEP CONVEX NETWORK (YU & DENG, 2011)	0.83%
MANIFOLD TANGENT CLASSIFIER (RIFAI ET AL., 2011)	0.81%
DBM + DROPOUT (HINTON ET AL., 2012)	0.79%

Goodfellow, Maxout Networks, et al., 2013

Optymalizacja spadkiem gradientu

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla E$$
$$E = \frac{1}{N} \sum_{\mathbf{x}} E(f(\mathbf{x}), y)$$



- **Batch gradient descent** - dla całego zbioru, niepraktyczne dla dużych danych
- **SGD on-line** - stochastic gradient descent dla $N = 1$, duża wariancja
- **SGD mini-batch** estymacja błędu na podstawie paczki zawierającej n losowych przypadków

Jak dobrać stałą uczenia η ?

Rys: S. Ruder, „Optimising for Deep Learning”

SGD z pędem (momentum)

$$w(t+1) = w(t) - \eta \nabla E(t) + \gamma \Delta w(t)$$

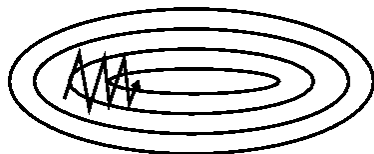
- $v(t) = \Delta w(t)$ prędkość uczenia akumuluje historyczne wartości gradientów (średnia krocząca)

$$v(t) = \gamma v(t-1) - \eta \nabla E(t)$$

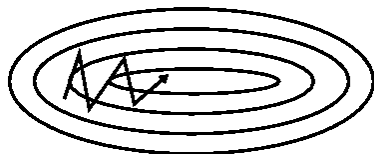
$$w(t+1) = w(t) + v(t)$$

- pęd zwiększa krok, gdy gradient nie zmienia kierunku
- redukuje oscylacje (spowalnia uczenie), gdy gradient zmienia kierunek
- gdy $\nabla E = 0$ wagi są nadal modyfikowane (bezwładność)
- typowo współczynnik $\gamma = 0.9$

Przyspieszenie zbieżności SGD za pomocą momentu



SGD bez momentu



SGD z momentem

Nesterov accelerated gradient (NAG)

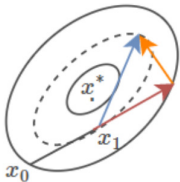
$$\begin{aligned}v(t) &= \gamma v(t-1) - \eta \nabla E(w(t) + \gamma v(t-1)) \\w(t+1) &= w(t) + v(t)\end{aligned}$$

- gradient ∇E jest liczony dla przybliżonej przyszłej pozycji
 $\hat{w}(t) = w(t) + \gamma v(t-1)$
- trening szybciej jest w stanie zareagować na zmiany gradientu

Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. Doklady ANSSSR (translated as Soviet.Math.Docl.), vol. 269, pp. 543–547.

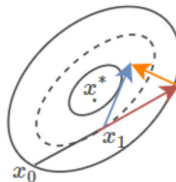
Polyak vs Nesterov

Polyak's Momentum



$$x_{t+1} = x_t - \alpha \nabla f(x_t) + \mu(x_t - x_{t-1})$$

Nesterov Momentum



$$x_{t+1} = x_t + \mu(x_t - x_{t-1}) - \gamma \nabla f(x_t + \mu(x_t - x_{t-1}))$$

Dla małych wartości η metody stają się równoważne

Rys: Ioannis Mitliagkas, *Theoretical principles for deep learning*, 2018

AdaGrad (Duchi, 2011)

$$w(t+1) = w(t) - \frac{\eta}{\sqrt{v(t)}} \nabla E(t)$$

gdzie $v(t)$ akumuluje historyczne wartości gradientów

$$v(t) = v(t-1) + [\nabla E(t)]^2 = \sum_{t' < t} [\nabla E(t')]^2$$

Efektywna stała uczenia $\frac{\eta}{\sqrt{v(t)}}$ ustalana dla każdego optymalizowanego parametru niezależnie

- strome spadki $\rightarrow$ duży gradient $\rightarrow$ mniejszy krok uczenia
szybka zbieżność dla funkcji wypukłych
- małe gradienty $\rightarrow$ większy krok uczenia
przeciwdziała efektowi znikającego gradientu
- domyślna wartość $\eta = 0.01$, zazwyczaj nie wymaga zmiany
- wada: mianownik może szybko przybrać duże wartości, co powoduje przedwczesny zanik uczenia (problem ten niwelują algorytmy Adadelta, RMSProp, Adam)

RMSProp (Hinton, 2012)

$$w(t+1) = w(t) - \frac{\eta}{\sqrt{v(t)}} \nabla E(t)$$

gdzie $v(t)$ jest wykładniczo ważoną średnią kroczącą kwadratów gradientów (EMA, *exponential moving average*)

$$v(t) = \beta v(t-1) + (1 - \beta) [\nabla E(t)]^2$$

- $RMS(\nabla E(t)) = \sqrt{v(t)}$ estymacja RMS (*root mean square*)
- modyfikacja AdaGrad zapewniająca lepszą zbieżność dla nie-wypukłych funkcji
- największy wpływ na wielkość $v(t)$ mają gradienty z ostatnich iteracji
- β kontroluje skalę długości ruchomej średniej, typowo $\beta = 0.9$ (średnia z 10 kroków)
- dobre wyniki w połączeniu z momentem Nesterova

Adadelta (Zeiler, 2012)

$$w(t+1) = w(t) - \frac{RMS(\Delta w(t-1))}{RMS(\nabla E(t))} \nabla E(t)$$

gdzie $RMS(\nabla E(t)) = \sqrt{v(t)}$

$$v(t) = \beta v(t-1) + (1 - \beta) [\nabla E(t)]^2$$

$RMS(\Delta w(t-1))$ średnia krocząca kwadratów Δw z poprzedniego kroku

- brak globalnej stałej uczenia η (ale należy dobrać współczynnik β)
- jest heurystycznym przybliżeniem metod kwadratowych (bez wyznaczania Hessianu)
- krok Δw wykonywany jest w tych samych jednostkach skali (metody 1 rzędu $\nabla E \approx \frac{1}{w}$ nie zachowują jednostki)

Adaptive Moment Estimation (Adam)

$$w(t) = w(t-1) - \frac{\eta}{\sqrt{\hat{v}(t)}} \hat{m}(t)$$

Estymaty pierwszego (średnia) i drugiego (wariancja) momentu gradientów

$$m(t) = \beta_1 m(t-1) + (1 - \beta_1) \nabla E(t)$$

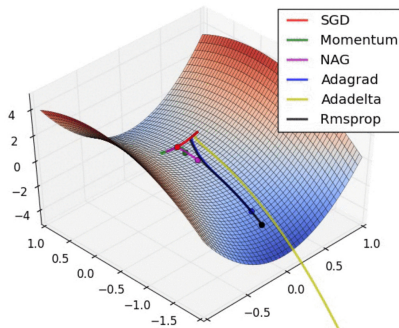
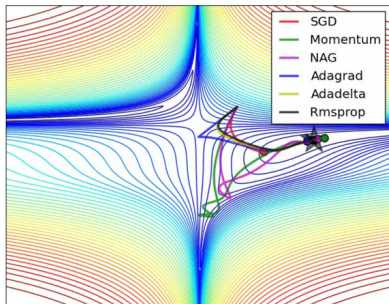
$$v(t) = \beta_2 v(t-1) + (1 - \beta_2) [\nabla E(t)]^2$$

Poprawka na odchylenie w stronę wartości 0 (z powodu zerowych początkowych wartości estymaty)

$$\hat{m}(t) = \frac{m(t)}{1 - \beta_1}, \quad \hat{v}(t) = \frac{v(t)}{1 - \beta_2}$$

współczynniki zanikania $\beta_1 = 0.9$, $\beta_2 = 0.999$
stała uczenia $\eta = 0.01$

Porównanie



👉 An overview of gradient descent optimization algorithms

Podsumowanie

- DNN dobre do analizy danych bez wyraźnej struktury np. przestrzennej lub czasowej (sekwencje)
- w większości zastosowań aktywacja ReLU sprawdza się bardzo dobrze oraz algorytm adam z domyślnymi parametrami η, β_1, β_2
- SGD z zanikającą wykładniczo wartością stałej uczenia może dorównać algorytmom takim jak adam

CNN

Convolutional Neural Networks

Sieci splotowe

Convolutional Neural Networks (CNN, ConvNet) wariant MLP inspirowany biologicznie, gdzie mnożenie macierzy wag i sygnału wejściowego zastąpione jest operacją splotu

Zdolne do generalizacji sygnału posiadającego relacje przestrzenne

- 1D sygnały czasowe
- 2D obrazy
- 3D fMRI, video, obrazy RGB
- sygnały wielokanałowe

Potrafią być odporne na przestrzenne transformacje sygnału (skalowanie, obrót, przesunięcie)

Splot 1D

- **Splot** (convolution) funkcji $x(t)$ oraz $w(t)$

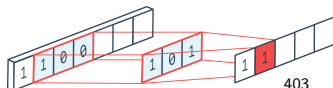
$$(x \star w)(t) = \int x(\tau)w(t - \tau)d\tau$$

- **Korelacja wzajemna** (*cross correlation*)

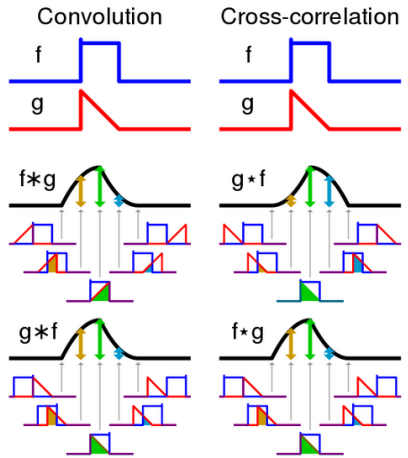
$$(x \star w)(t) = \int x(\tau + t)w(\tau)d\tau$$

- wynikiem jest funkcja, np. uśredniona wartość x względem wszystkich wartości w jeśli w spełnia wymagania gęstości prawdopodobieństwa
- W sieciach spłotowych CNN:
 x - sygnał wejściowy
 w - kernel (filtr), wagi połączeń neuronu, niezerowe tylko w ograniczonym obszarze (pole recepcyjne)
- wartości wyjściowe tworzą **mapę cech** (*feature map*)

$$s(i) = \sum_m x(m + i)w(m)$$

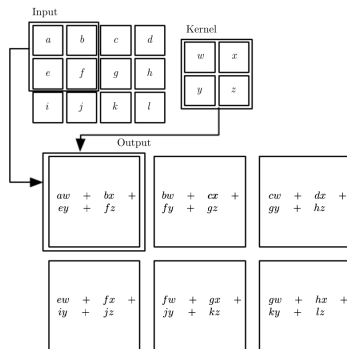


Splor i korelacja wzajemna



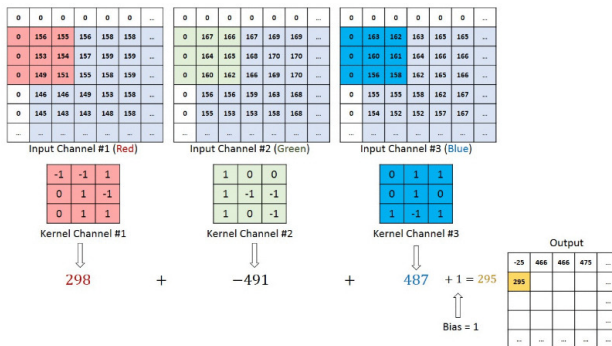
<https://en.wikipedia.org/wiki/Cross-correlation>

Splot 2D - wejście 1 kanał



$$s(i, j) = \sum_l \sum_m w(l, m) \cdot x(i + l, j + m)$$

Spot 2D - wejście 3 kanały (RGB)



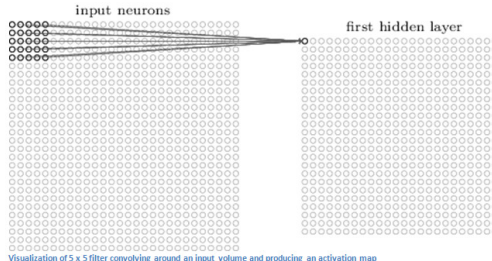
$$s(i, j, k) = \sum_m \sum_n \sum_k w(m, n, k) \cdot x(i + m, j + n, k)$$

Filtry graficzne 2D - przykłady

Operation	Filter	Convolved Image
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	
Gaussian blur (approximation)	$\frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	

<https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/>

Pola recepcyjne i mapy cech

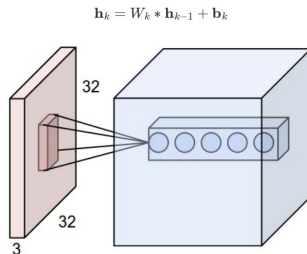
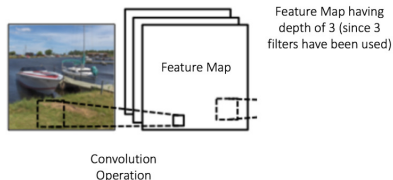


- **Pole recepcyjne** - obszar „widziany” przez neuron (rozmiar filtra)
- wyjście neuronu: splot sygnału i liniowego filtra, ewentualny wyraz wolny (bias) i nieliniowa funkcja wyjściowa (ReLU, tanh), generują **mapę cech** (*feature map*)

$$s(i, j) = \sigma \left(b + \sum_m \sum_n \sum_k w(m, n, k) \cdot x(i + m, j + n, k) \right)$$

Pola recepcyjne i mapy cech

- N neuronów (filtrów) tworzy N map (objętość, tensor n wymiarowy)

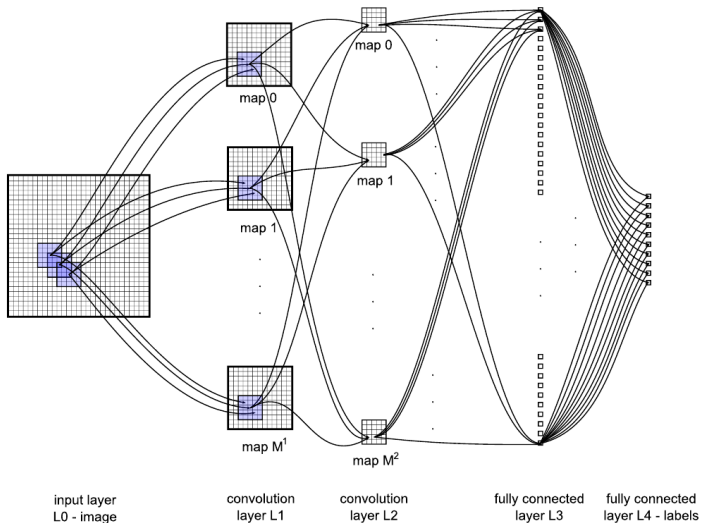


- Przy przetwarzaniu obrazów sygnał wejściowy x to tensor 3D: szerokość, wysokość, kanał (RGB). Dla filtra l

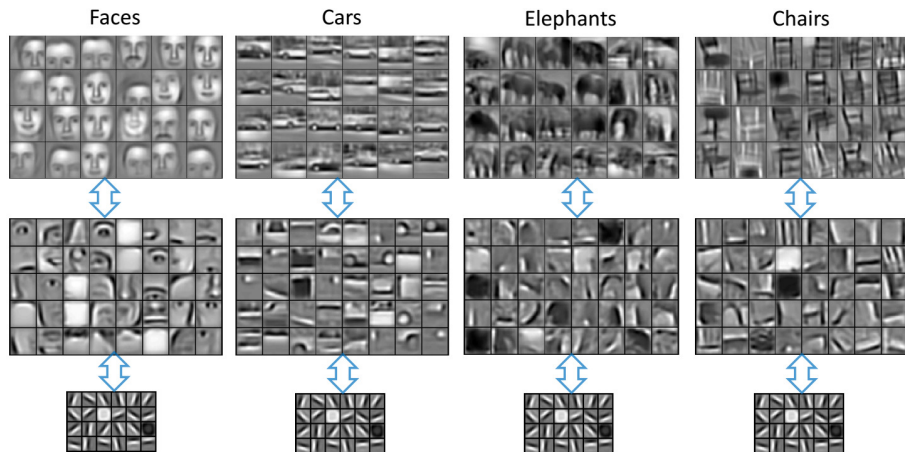
$$s(i, j, l) = \sum_{mnk} x(i + m, j + n, k) w(m, n, k, l)$$

- tensor wyjściowy warstwy zawierającej N filtrów (M_x, M_y, N) , gdzie M_x, M_y - rozmiary mapy wyjściowej
- w praktyce x rozszerzony do 4D o wymiar związany z rozmiarem mini-batcha

Typowa architektura CNN



Przykłady filtrów



Gooffellow, 2016 [1]

Rozmiar mapy cech

- Warstwa zawierająca N neuronów (filtrów) tworzy N map (objętość, tensor n wymiarowy)
- Rozmiar mapy wyjściowej (M_x^n, M_y^n) w warstwie n :

$$M_x^n = \frac{M_x^{n-1} - K_x^n + 2P_x^n}{S_x^n} + 1 \quad M_y^n = \frac{M_y^{n-1} - K_y^n + 2P_y^n}{S_y^n} + 1$$

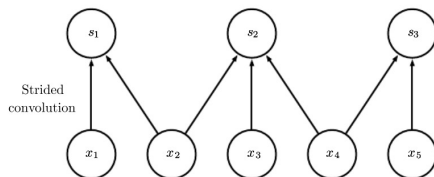
K_x, K_y rozmiar filtra (szerokość i wysokość)
 S_x, S_y przesunięcie (*stride*) w każdym z wymiarów,
splot 2D

$$s(i, j, k) = \sum_{lmn} x(l, j \times S_x + m, k \times S_y + n) w(i, l, m, n)$$

P_x, P_y wielkość rozszerzenia brzegu (*zero padding*),

Stride - krok przesunięcia

Przykład w 1D: S=2, K=3, P=1

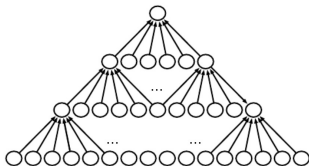


$$M^n = \frac{M^{n-1} - 3 + 2}{2} + 1$$

- redukcja wymiaru - zmniejszenie wymogów obliczeniowych i pamięciowych
- zmniejszenie rozdzielczości sygnału
- kosztem mniej dokładnej reprezentacji

Zero padding

valid zero padding, brak rozszerzenia na brzegach ($P = 0$)



$$M^n = \frac{M^{n-1} - K^n}{S^n} + 1$$

- rozmiar map maleje w kolejnych warstwach
- ogranicza to możliwość budowania głębokich sieci i wymuszaAn Interactive Node-Link Visualization of Convolutional Neural Networks by Adam W. Harley.a stosowanie małych filtrów
- sygnał wejściowy na brzegach ma mniejszy wpływ na sygnał wyjściowy

Zero padding

same zero padding, brzegi wypełnione dodatkowymi wartościami (zerami) aby zapewniać odtworzenie wymiaru sygnału wejściowego
 $P = \lfloor \frac{K}{2} \rfloor, S = 1$



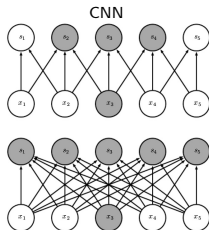
- pozwala budować bardzo głębokie sieci o dowolnych wielkościach filtrów $M_i^n = \lceil \frac{M_i^{n-1}}{S_i} \rceil$
- sygnał wejściowy na brzegach ma mniejszy wpływ na sygnał wyjściowy (na brzegach 0)

Właściwości CNN

- **rzadka reprezentacja** - rozmiar filtra jest dużo mniejszy od rozmiaru sygnału wejściowego, pojedyncze wejście oddziałuje tylko na grupę neuronów
- **współdzielenie parametrów** - warstwa splotowa może być widziana jako w pełni połączona warstwa ze współdzielonymi wagami (dużo mniejsza liczba parametrów w stosunku do MLP)
- **równoważność względem przesunięcia** sygnału - ta sama cecha znajdująca się w różnych miejscach obrazu będzie aktywowała ten sam filtr
- możliwość użycia sygnału wejściowego o zmiennym rozmiarze - większy obraz wejściowy wygeneruje większe mapy, w przypadku MLP zwiększenie rozmiaru wektora wejściowego wymaga rozbudowy architektury

Rzadka reprezentacja

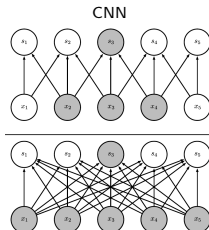
pojedyncze wejście aktywuje
tylko grupę neuronów w
małym obszarze



MLP

MLP: mnożenie macierzy $O(m \times n)$ parametrów
CNN: warstwa splotowa $O(k \times n)$, gdzie $k \ll m$

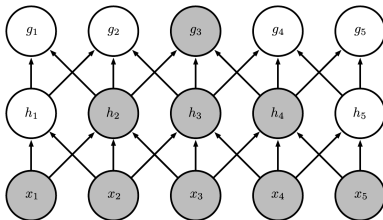
wyjście neuronu zależne od
małego obszaru sygnału
wejściowego



MLP

Efektywne pole recepcyjne

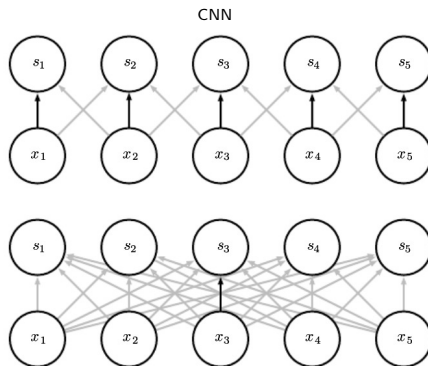
- **Efektywne pole recepcyjne** - obszar sygnału wejściowego pokryty przez neurony w wyższych warstwach, rośnie z głębokością



- stride, pooling i dilation (rozrzedzony spłot) dodatkowo zwiększają efektywne pole recepcyjne
- pomimo rzadkich połączeń sieć jest w stanie w ten sposób modelować złożone (odległe) zależności

Współdzielenie wag

- ta sama waga jest używana przy przetwarzaniu każdego punktu wejściowego
- pojedynczy filtr pozwala wykryć tę samą cechę w różnych położeniach obrazu wejściowego (*equivariance to translation*)



Źródło grafiki: Goofellow, 2016 [1]

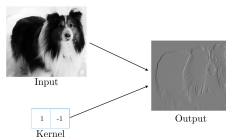
MLP

Sieci Neuronowe

419

Rzadkie połączenia i współdzielenie wag

Przykład: wykrywanie pionowych krawędzi obrazu $n \times n$



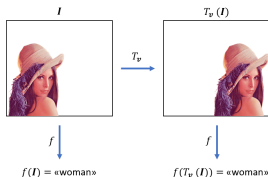
- MLP: mnożenie pełnej macierzy $n^2 \times n^2 = n^4$ wag, $n^2 \times (n^2 + 1)$ operacji
- CNN: splot filtrem 1×2 (2 wagi) wymaga $n^2 \times 3$ operacji (2 mnożenia i dodawanie)
- warstwa splotowa istotnie wydajniejsza w mapowaniu powtarzających się relacji w małych, lokalnych rejonach sygnału wejściowego

Równoważność przy przesunięciu

Przesunięcie sygnału wejściowego powoduje identyczne przesunięcie sygnału na mapie cech.

$$T_v(\mathbf{x}) = \mathbf{x} + \mathbf{v}$$

Aktywowane są te same filtry wykrywające obiekt



Równoważność translacji jest pożądaną własnością w rozpoznawaniu obrazów, gdzie lokalne cechy (np. krawędzie) mogą wystąpić w każdym miejscu na obrazie

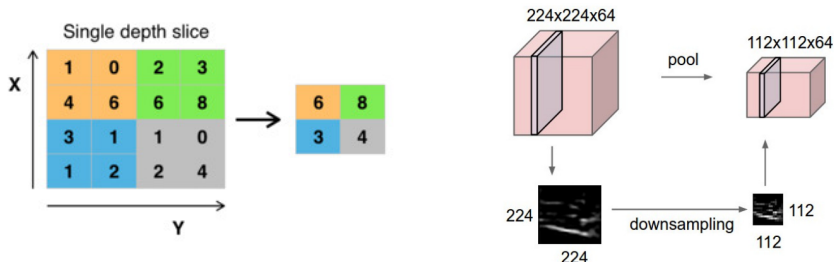
W przypadku sieci MLP przesunięcie sygnału wejściowego powoduje zmianę aktywacji wszystkich

Równoważność przy przesunięciu

- współdzielenie wag dla całego wejścia nie zawsze jest pożądane, można zastosować różne filtry w różnych obszarach obrazu, np. rozpoznawanie twarzy ze zdjęć paszportowych, które posiadają charakterystyczne cechy występujące wyłącznie w określonych rejonach sygnału wejściowego
- Sieci CNN splot nie są niezmiennicze względem zmiany skali, obrotu oraz przekształceń afinicznych obrazu. Taką własność osiąga się poprzez rozszerzenie zbioru treningowego o przypadki zdeformowane (szum, skalowanie, obrót, zmiana kontrastu, etc..)

Pooling

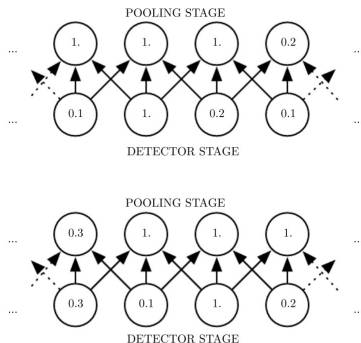
- typowy blok realizujący warstwę splotową CNN:
 - liniowa aktywacja (splot)
 - detekcja (nieliniowość np. ReLU)
 - redukcja (pooling) - „uogólnienie” wartości sąsiadujących wyjść
- **Max pooling** - maksimum z pewnego podobszaru



- redukcja wymiarowości - przesunięcie S filtra typowo równe jego wielkości K (maksimum z rozłącznych obszarów)
- inne podejścia: *avg. pooling* (średnia z sąsiedztwa), *norm pooling* (norma z sąsiadujących wyjść), ważona średnia od centrum, niektóre architektury rezygnują z tej warstwy

Pooling względem obrazu wejściowego

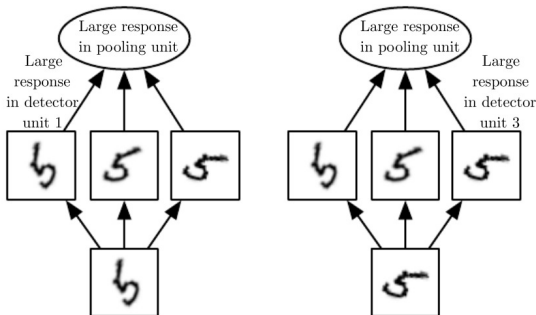
Pooling zapewnia niezmienniczość względem drobnego przesunięcia obrazu wejściowego



Źródło grafiki: Gooffellow, 2016 [1]

Pooling względem map cech

Redukcja względem wyjść różnych splotów umożliwia wprowadzenie do modelu niezmienniczości względem pewnych transformacji sygnału wejściowego (np. obrotu)



Źródło grafiki: Gooffellow, 2016 [1]

MLP - oznaczenia

M - liczba warstw ($M = 3$)

$w_{ij}^{(l)}$ - waga łącząca element i należący do warstwy $l - 1$ oraz element j z warstwy l

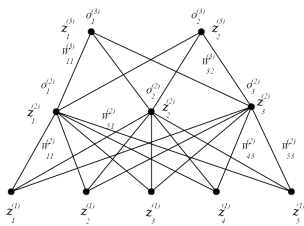
$z_j^{(l)}$ - aktywacja neuronu j w warstwie l

$$z_j^{(l)} = \sum_i w_{ij}^{(l)} o_i^{(l-1)}$$

$o_j^{(l)}$ - sygnał wychodzący z elementu j należącego do warstwy l

$$o_j^{(l)} = \sigma(z_j^{(l)}) = \sigma \left(\sum_i w_{ij}^{(l)} o_i^{(l-1)} \right)$$

$f_i(\mathbf{x}) = o_i^{(M)}$ funkcja realizowana przez MLP (i -te wyjście)



Algorytm wstecznej propagacji dla sieci MLP

Sygnał propagowany od wejścia do wyjścia

$$o_i^h = f(z_i^h) \quad z_i = \sum_j w_{ji}^h o_j^{h-1} + b_i^h$$

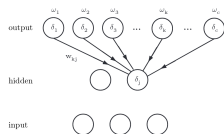
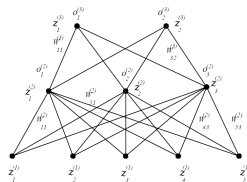
Sygnał błędu neuronów δ_i^k propagowany od warstwy wyjściowej do wejścia

$$\delta_i^{out} = \frac{\partial E}{\partial z_i^h} = f'(z_i^{out})(y_i - f(z_i^{out}))$$

$$\delta_i^h = \frac{\partial E}{\partial z_i^h} = f'(z_i^h) \sum_j \delta_j^{h+1} w_{ij}^{h+1}$$

Aktualizacja wag w kierunku spadku gradientu

$$\Delta w_{ij}^h = -\eta \frac{\partial E}{\partial w_{ij}^h} = \eta \delta_j^h o_i^{h-1}$$



Wsteczna propagacja błędu w sieciach CNN

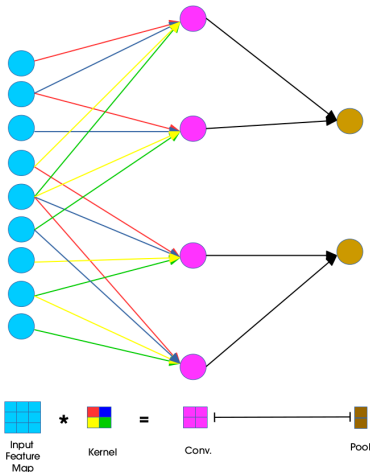
Sygnał propagowany od wejścia do wyjścia (dla jednego kanału $H \times W$ i filtra $k_1 \times k_2$)

$$o_{ij}^h = f(z_{ij}^h)$$

$$z_{ij} = \sum_{m=0}^{k_1-1} \sum_{n=0}^{k_2-1} w_{mn}^h o_{i+m,j+n}^{h-1} + b^h$$

onz. * operacja splotu

$$o_{ij}^h = f\left((\mathbf{w}^h * \mathbf{o}^{h-1})_{ij} + b^h\right)$$

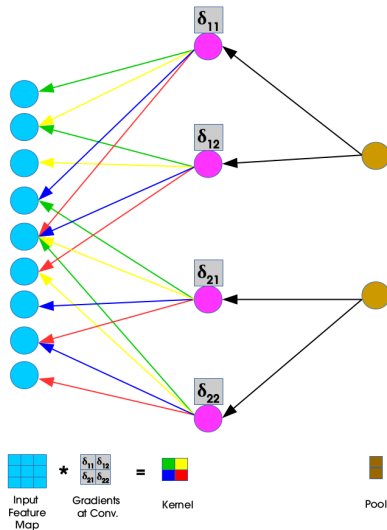
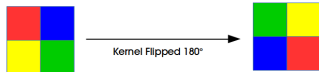


Źródło
<https://www.jefkine.com/general/2016/09/05/backpropagation-in-convolutional-neural-networks/>

Wsteczna propagacja błędu w sieciach CNN

Sygnał błędu neuronów δ_i^k od warstwy wyjściowej do wejścia

$$\begin{aligned}\delta_{ij}^h &= \frac{\partial E}{\partial z_{ij}^h} = \\ &= f'(z_{ij}^h) \sum_{m=0}^{k_1-1} \sum_{n=0}^{k_2-1} \delta_{i-m,j-n}^{h+1} w_{mn}^{h+1} \\ &= f'(z_{ij}^h) \left(\delta^{h+1} * \text{rot}_{180^\circ}(\mathbf{w}^{h+1}) \right)_{ij}\end{aligned}$$

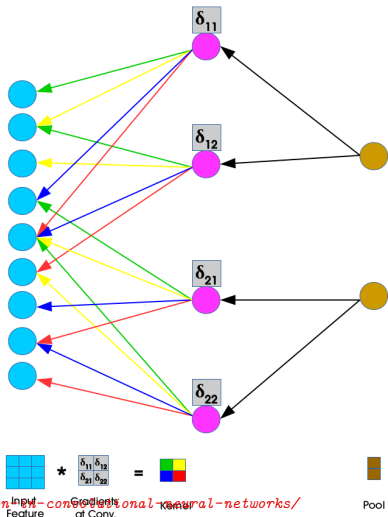


Źródło
<https://www.jefkine.com/general/2016/09/05/backpropagation-in-convolutional-neural-networks/>

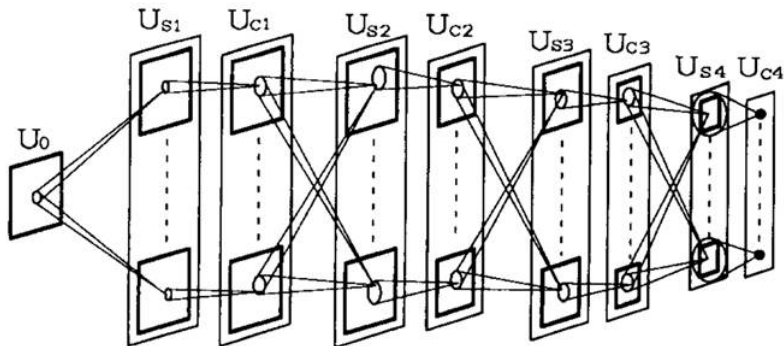
Wsteczna propagacja błędu w sieciach CNN

Splot mapy wejściowej o wymiarze $H \times W$ i filtra o wymiarze $k_1 \times k_2$ tworzy mapę wyjściową o wymiarze $Hk_1 + 1 \times Wk_2 + 1$. Aktualizacja wag w kierunku spadku gradientu:

$$\begin{aligned}\Delta w_{ij}^h &= -\eta \frac{\partial E}{\partial w_{ij}^h} = \\ &= \eta \sum_{m=0}^{H-k_1} \sum_{n=0}^{W-k_2} \delta_{mn}^h o_{m+i, n+j}^{h-1} = \\ &= \eta (\delta^h * \mathbf{o}^{h-1})_{ij}\end{aligned}$$



Neocognitron (Fukushima, 1980)



- pierwsza sieć CNN do rozpoznawania ręcznie pisanych cyfr
- architektura inspirowana biologicznie ludzkim układem przetwarzania obrazu w mózgu
- uczenie nienadzorowane (bez wstecznej propagacji błędów)

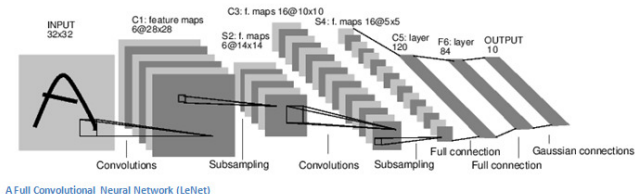
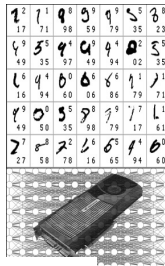
LeNet5 (LeCun 1998)

klasyfikacja cyfr pisanych ręcznie
(odczytywanie czeków)

👉 MNIST

trening: 60k cyfr, 250 osób,

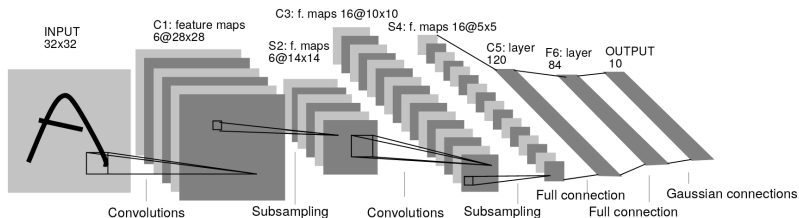
test: 10k cyfr



👉 LeNet-5, convolutional neural networks

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. *Gradient-based learning applied to document recognition. Proceedings of the IEEE, november 1998*

LeNet-5



Layer	kernels	size	output	connections	parameters
Input			1x32x32		
C1	6	5x5	6x28x28	122304	156
S2		2x2	6x14x14	5880	12
C3	16	5x5	16x10x10	151600	1516
S4		2x2	16x5x5	2000	32
C5	120	5x5	120x1x1	48120	48120
F6	84		84x1x1	10164	10164
Output	10		10x1x1		

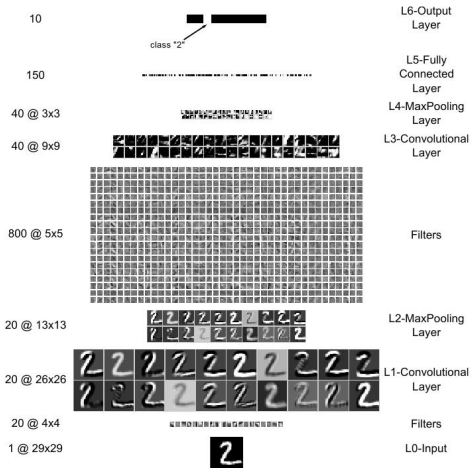
LeNet-5 połączenia S2-C3

Mapy C3 połączone wyłącznie z wybranymi mapami z S2

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	X				X	X	X			X	X	X	X		X	X
1	X	X				X	X	X			X	X	X	X		X
2	X	X	X				X	X	X			X		X	X	X
3		X	X	X			X	X	X	X			X		X	X
4			X	X	X			X	X	X	X		X	X		X
5				X	X	X			X	X	X	X		X	X	X

- redukcja liczby połączeń
- przełamanie symetrii sieci - różne sygnały wejściowe pozwalają uzyskać różnorodne (być może komplementarne) zestawy detektorów cech

Przykład wizualizacji aktywacji sieci

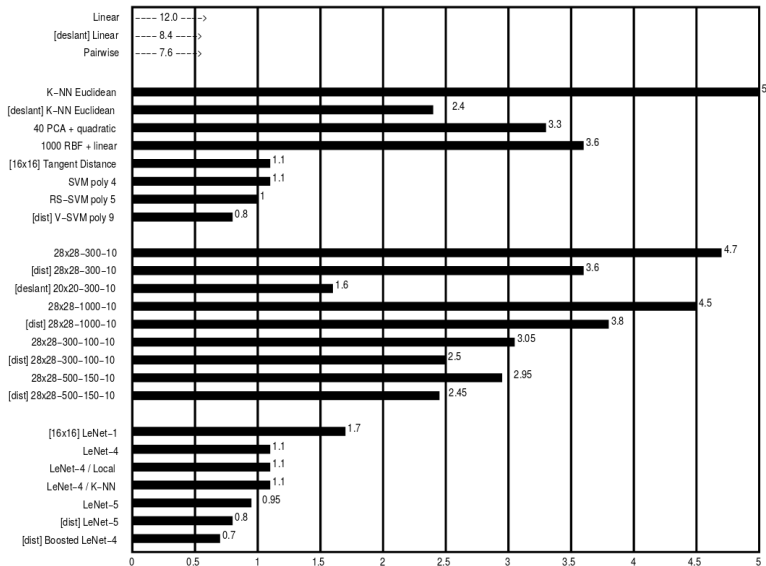


👉 Wizualizacja sieci plotowej

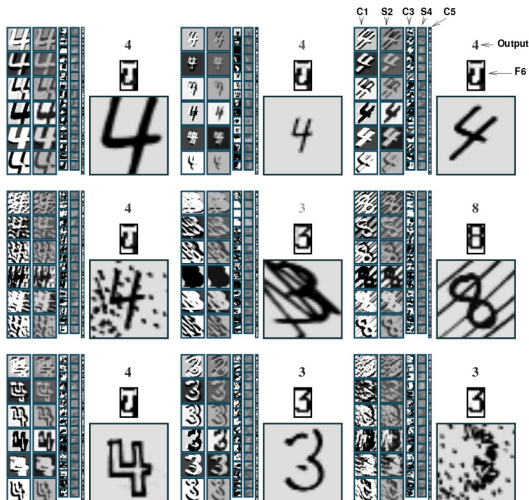
LeNet-5 błędy klasyfikacji



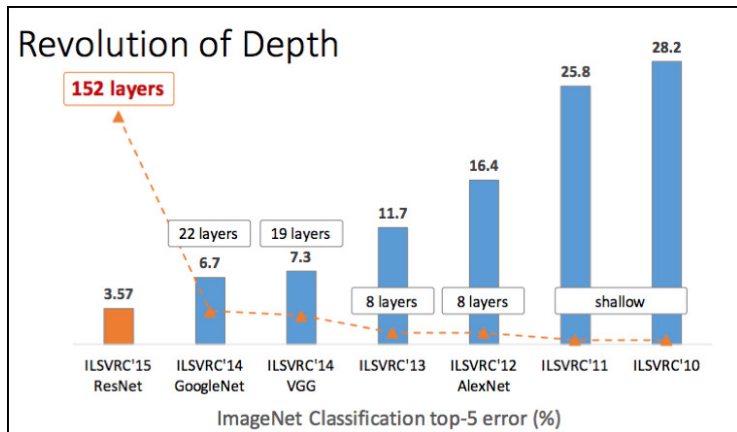
LeNet-5 porównanie z innymi metodami



LeNet-5 odporność na szum, zniekształcenia i nietypowe przypadki



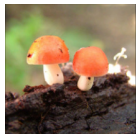
ImageNet



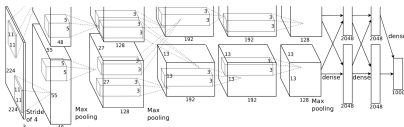
- 👉 CNNs Architectures used on ImageNet

AlexNet (A. Krizhevsky, 2012)

- Zwycięzca 🖱️ ImageNet w 2012, błąd klasyfikacji 15.3% (poprawa z 26%)
- 1.2M obrazów, 1000 klas



mushroom

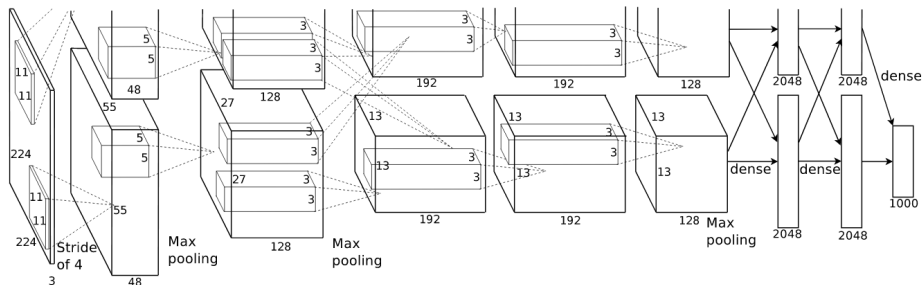


- architektura wzorowana LeNet5 ale głębsza (8 warstw) i więcej filtrów na warstwę, filtry 11x11, 5x5, 3x3
- sekwencje warstw spłotowych z jednostkami ReLU
- regularyzacja: dropout, L2, rozszerzanie danych, normalizacja wyjść wybranych warstw
- SGD z momentem, 20 epok, 6 dni treningu (2x NVIDIA GTX 580 3GB GPUs), 60M parametrów

A. Krizhevsky, I. Sutskever, G.E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, 2012

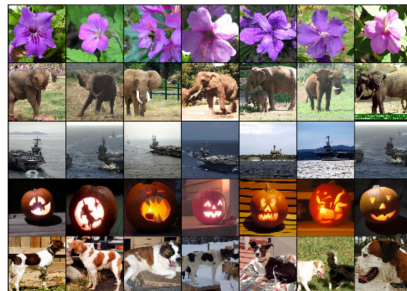
AlexNet

Layer	kernels	size	stride	output
Input				224x224x3
C1 + LRN	96	11x11	4x4	
MaxPooling		3x3	2x2	55x55x96
C2 + LRN	256	5x5		
MaxPooling		3x3	2x2	27x27x256
C3	384	3x3		13x13x384
C4	384	3x3		13x13x384
C5	256	3x3		13x13x256
MaxPooling		3x3	2x2	13x13x256
F6 + dropout($p = 0.5$)	4096			
F7 + dropout($p = 0.5$)	4096			
Output softmax	1000			

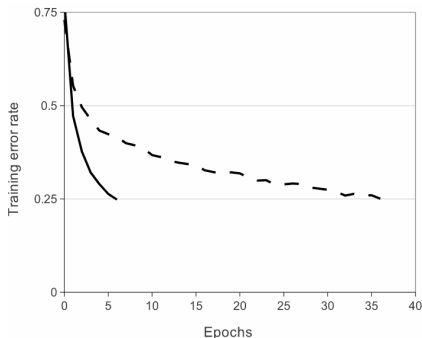




5 najsilniejszych odpowiedzi



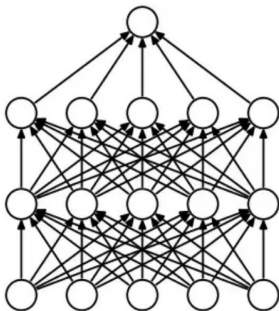
obraz wejściowy i 6 najbliższych wzorców względem odległości Euklidesowej wektora pobudzeń ostatniej warstwy ukrytej



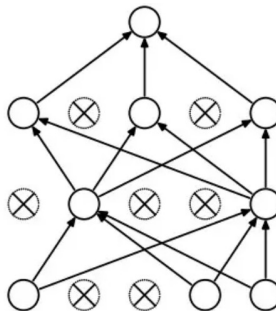
- Porównanie szybkości zbieżności 4 warstwowej sieci spłotowej z jednostkami ReLU i tanh na danych ImageNet
- Zastosowanie ReLU do kilkukrotnego przyspieszenia zbieżności
- Inicjalizacja: wagi z rozkładu Gaussa $N(0, 0.01)$, obciążenia (bias) $b = 1$, stąd większość ReLU aktywnych na początku treningu

Dropout

Dropout (Srivastava et al., 2014) dla każdego wzorca treningowego z prawdopodobieństwem p „wyrzuca” jednostki (zeruje ich aktywacje), odrzucone jednostki nie biorą udziału w treningu



(a) Standard Neural Net



(b) After applying dropout.

Rys:  ResNet, AlexNet, VGGNet, Inception: Understanding various architectures of Convolutional Networks by Koustubh Sinhal

Dropout

- Silnie zapobiega przeuczeniu
- Używany wyłącznie w czasie treningu. Podczas predykcji (*inference*) dropout jest wyłączany, czyli $p = 1$
- Każdy krok uczenia odbywa się ze zmienioną losowo architekturą sieci, uczenie przebiega innymi ścieżkami grafu połączeń (ale w CNN wagi są współdzielone)
- Dla n jednostek mamy 2^n możliwych architektur, trening w mini-batchu uśrednia gradient względem różnych, wylosowanych sieci
- Przeciwdziała powstawaniu złożonych relacji pomiędzy wieloma neuronami, stąd pojedynczy neuron jest zmuszony wykrywać bardziej wartościowe cechy, niezależne od wpływu innych neuronów
- Wolniejsza zbieżność, np. AlexNet 2 x wolniej)

Local response normalization

Normalizacja wartości $a_{x,y}^i$ pikseli w pozycji x, y i kanału (mapy cech) i wejściowych do warstwy spłotowej

$$b_{x,y}^i = a_{x,y}^i / \left(k + \alpha \sum_{j=\max(0, i-n/2)}^{\min(N-1, i+n/2)} (a_{x,y}^j)^2 \right)^\beta$$

- $k = 2, n = 5, \alpha = 10^{-4}, \beta = 0.75$ dobrane heurystycznie ze zbioru walidacyjnego
- realizuje normalizację jasności dla n sąsiadujących kolejnych map cech
- w AlexNet poprawa o 1.4% (top 1) oraz 1.2% (top 5)

Dodatkowa regularyzacja

Sztuczne rozszerzenie zbioru danych (*Data augmentation*)

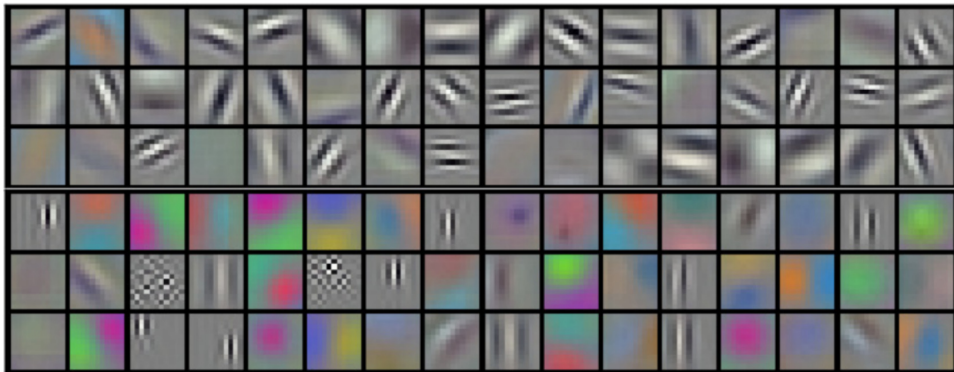
- przesunięcie i odbicia: losowanie obrazków 224x224 z 256x256, zwiększenie zbioru 2048 razy niezbędne do uniknięcia przeuczenia przy tak dużej sieci
- losowa modyfikacja intensywności kanałów RGB
- poprawa ponad 1% (top 1, top 5)

MaxPooling z nakrywaniem

- kernel 3x3, stride 2x2
- poprawa 0.4% (top 1) i 0.3% (top 5) względem próbkowania bez nakrywania (kernel 2x2)
- obserwacja: AlexNet rzadziej ulegał przeuczeniu

Zastosowanie 2 rdzeni GPU pozwoliło na szkolenie sieci o większej liczbie filtrów co zaowocowało poprawą 1.7% (top-1) oraz 1.2% (top-5) w porównaniu z mniejszą siecią na 1 rdzeniu

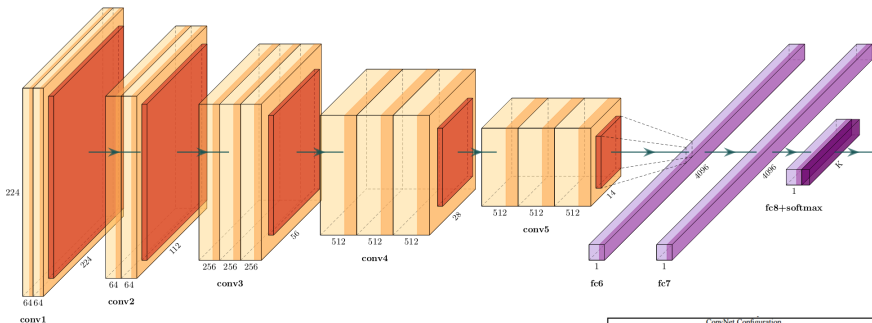
Filtry w pierwszej warstwie



96 filtrów pierwszej warstwy spłotowej, osobne kolumny warstw spłotowych uczą się wykrywania innego typu relacji, obserwowane przy każdym treningu

VGGNet (Simonyan and Zisserman, 2014)

Visual Geometry Group, University of Oxford



- bloki warstw spłotowych + max pooling
- wielkość map zmniejszana o połowę po każdym bloku
- ilość filtrów zwiększana dwukrotnie po każdym bloku
- małe filtry (3×3), gęsty spłot (stride=1)

ConvNet Configuration					
A	B	C	D	E	
11 weight layers	11 weight layers	13 weight layers	16 weight layers	19 weight layers	
input (224 × 224 RGB image)					
conv 3-64	conv 3-64	conv 3-64	conv 3-64	conv 3-64	conv 3-64
	L2N				
maxpool					
conv 3-128	conv 3-128	conv 3-128	conv 3-128	conv 3-128	conv 3-128
maxpool					
conv 3-256	conv 3-256	conv 3-256	conv 3-256	conv 3-256	conv 3-256
conv 3-256	conv 3-256	conv 3-256	conv 3-256	conv 3-256	conv 3-256
maxpool					
conv 3-512	conv 3-512	conv 3-512	conv 3-512	conv 3-512	conv 3-512
conv 3-512	conv 3-512	conv 3-512	conv 3-512	conv 3-512	conv 3-512
maxpool					
conv 3-512	conv 3-512	conv 3-512	conv 3-512	conv 3-512	conv 3-512
conv 3-512	conv 3-512	conv 3-512	conv 3-512	conv 3-512	conv 3-512
maxpool					
FC-4096					
FC-1000					
softmax					

VGGNet

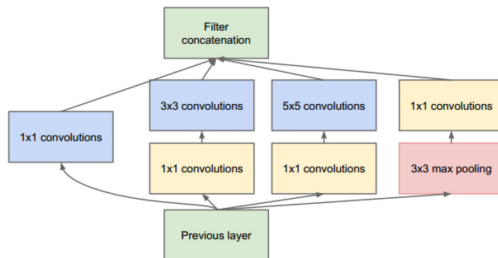
- zamiast dużych filtrów (w AlexNet 11x11, 7x7, 5x5) VGG stosuje we wszystkich warstwach filtry 3x3 zwiększając ich efektywne pola recepcyjne sekwencjami warstw splotowych
- małe filtry 3x3 zdolne do wykrycia bardziej subtelnych relacji w obrazach
- zwiększenie głębokości pozwala trenować bardziej złożone cechy
- stride=1, brak utraty informacji, gęsta konwolucja
- kilka architektur o różnej głębokości od 11 do 19 warstw (VGG16, VGG19)
- 138M parametrów
- trening 2-3 tygodnie (4x GPU), duże wymagania pamięciowe i obliczeniowe

K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv technical report, 2014

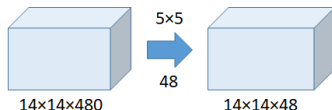
GoogLeNet/Inception

Sieć złożona z wielu bloków inceptyjnych (*inception module*) skonstruowanych wg. zasad:

- rozmiar filtru odwrotnie proporcjonalny do liczby filtrów, np. 128 filtrów 3x3 i 32 filtry 5x5
- równoległe przetwarzanie w kilku warstwach splotowych o zróżnicowanej wielkości filtrów 1x1, 3x3, 5x5 - detektory cech w różnej skali
- duże filtry poprzedzone splotem 1x1 w celu redukcji liczby parametrów. Splot 1x1 redukuje liczbę kanałów do 1.

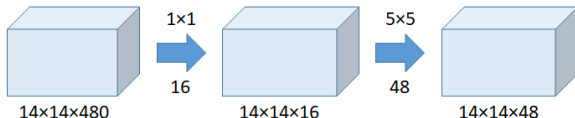


Redukcja złożoności spłotem 1x1



ilość parametrów $5 \times 5 \times 48 \times 480 = 576K$

ilość operacji $(14 \times 14 \times 48) \times (5 \times 5 \times 480) = 112.9M$



ilość parametrów $1 \times 1 \times 16 \times 480 + 5 \times 5 \times 48 \times 16 = 26.88K$

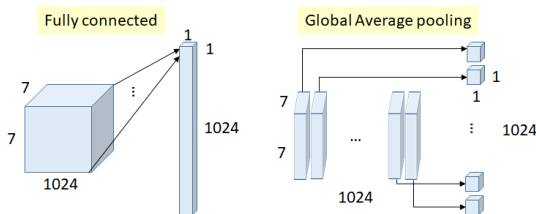
ilość operacji

$(14 \times 14 \times 16) \times (1 \times 1 \times 480) + (14 \times 14 \times 48) \times (5 \times 5 \times 16) =$
 $1.5M + 3.8M = 5.3M$

Rys:  Review: GoogLeNet (Inception v1)— Winner of ILSVRC 2014 (Image Classification) by Sik-Ho Tsang

Global average pooling

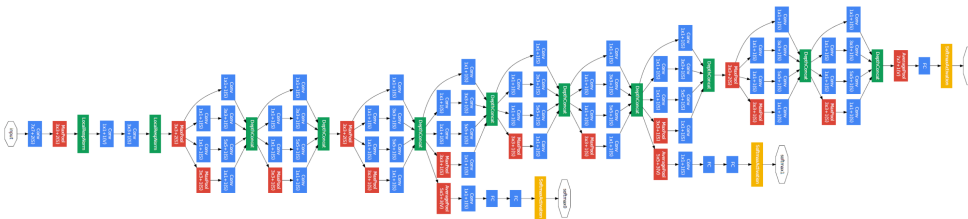
Globalne uśrednienie względem kanałów (global average pooling) zamiast pełnej połączonej warstwy za ostatnią warstwą splotową



$$7 \times 7 \times 1024 \times 1024 = 51.3\text{M}$$

- Warstwy w pełni połączone zawierają najwięcej parametrów (w AlexNet 90% parametrów) w sieci
- global average pooling: 0 wag, uśrednienie wartości dla poszczególnych kanałów
- w GoogLeNet uśrednianie poprawiło wynik top-1 o 0.6%

GoogLeNet/Inception



- 22 warstwy
- warstwy przewężające (bottleneck) dodatkowo redukują złożoność sieci
- kilka wyjść softmax na różnych poziomach sieci, stabilizują trening i przyspieszają zbieżność, zapobiegają zanikaniu gradientu

funkcja kosztu $J(\theta) = \mathbb{E}_D[-\mathbf{t} \log \mathbf{y}_1 - \mathbf{t} \log \mathbf{y}_2 - \mathbf{t} \log \mathbf{y}_3]$

- poprawność 93.3% top-5 na ImageNet, dużo szybszy w treningu od VGG
- kolejne iteracje: Inception v2, v3, InceptionResNet, Xception

ResNet (Residual Neural Network, Kaiming He et. al 2015)

Bloki ze skrótami (residual module)

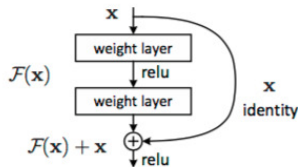
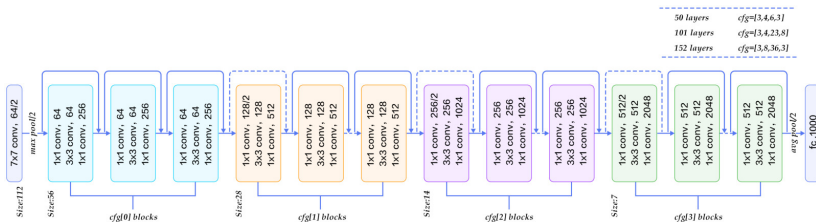


Figure 2. Residual learning: a building block.

- „skrót” łączy wejścia bloku z wyjściem poprzez odwzorowanie jednostkowe
- przejścia „skrótowe” pomagają uczyć bardzo głębokie sieci (152 warstwy)
- zapobiegają zanikaniu gradientu (sygnał może być propagowany skrótami)

ResNet



- wiele bloków zawierających sploty 1x1, 3x3, 1x1
- regularyzacja: batch normalization

Batch normalization

Batch normalization (Ioffe, 2015) normalizuje wejścia $x_i^{(k)}$ warstwy względem wartości średniej $\mu^{(k)}$ i wariancji $\sigma^{(k)}$ wzdłuż wymiaru k dla paczki danych w danej iteracji

$$y_i^{(k)} = \gamma^{(k)} \hat{x}_i^{(k)} + \beta^{(k)}$$

gdzie γ i β podlegają adaptacji w czasie treningu (uczenie normalizacji).

Czynnik normalizujący każde wejście

$$\hat{x}_i^{(k)} = \frac{x_i^{(k)} - \mu_B^{(k)}}{\sqrt{\sigma_B^{(k)^2} + \epsilon}}$$

średnia i odchylenie dla minipakietu o rozmiarze m

$$\mu_B = \frac{1}{m} \sum_{i=1}^m x_i, \quad \sigma_B^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_B)^2$$

Batch normalization

- przyspiesza zbieżność, poprawia stabilność treningu oraz polepsza generalizację
- rozwiązuje problem przesunięcia kowariancji - gdy zmienia się wyjście warstwy poprzedniej wówczas warstwa następna musi dopasować się do nowego rozkładu danych
- można stosować większe kroki uczenia bez obawy znikającego lub eksplodującego gradientu
- większa odporność na wpływ różnorodnej specjalizacji i metod treningu
- wygładza powierzchnię błędu (Santurkar, 2018)

Batch normalization w ResNet

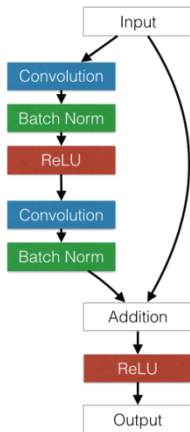


Figure 1. A ResNet basic block

HighNets: Highway networks

Sieci rezydualne, które starają się balansować wpływ głównej ścieżki (primary pathway) oraz ścieżki skrótowej (skip pathway)

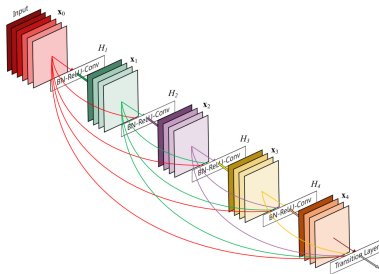
$$\mathbf{h}_n = T_{W'} f_W(h_{n-1}) + (\mathbf{1} - T_{W'}) h_{n-1}$$



- nawet do 1000 warstw
- wyniki state-of-the-art na MNIST i CIFAR 2015

DenseNets: Dense networks

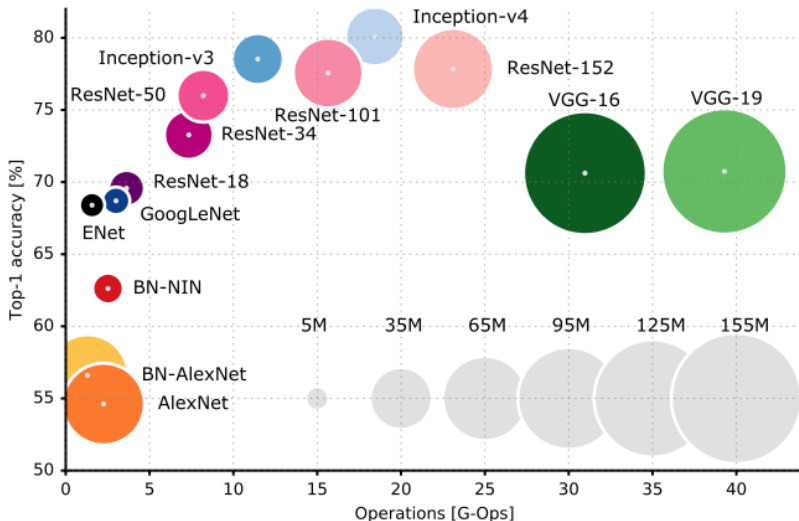
Sieci rezydualne, które posiadają połączenia skrótowe między wszystkimi blokami warstw sieci (do 5-ciu bloków)



- do 100 warstw
- wyniki state-of-the-art na wielu danych benchmarkowych 2018

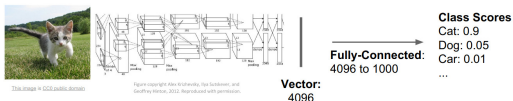
Porównanie popularnych sieci CNN

Poprawność klasyfikacji (top-1) zbioru Imagenet, liczba parametrów (zużycie RAM), szybkość działania (flops)

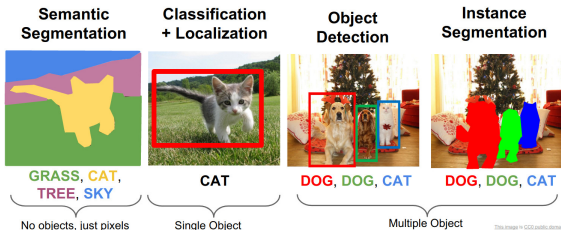


Analiza obrazów

- klasyfikacja (rozpoznawanie) obiektów



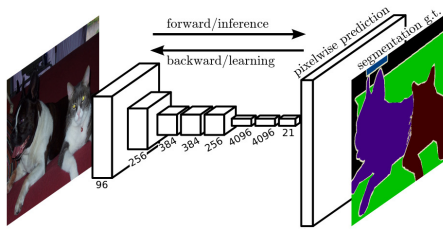
- detekcja i lokalizacja obiektów
- semantyczna segmentacja



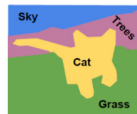
Rys: Fei-Fei Li, et.al., "Detection and Segmentation", lecture

Fully Convolutional Network (FCN)

Segmentacja semantyczna - predykcja obiektów na poziomie pikseli. Wyjście sieci jest rozmiaru $W \times H \times K$, gdzie $W \times H$ - rozmiar obrazu wejściowego, K - ilość obiektów do segmentacji (tło, kot, pies, itd...)



Input Image

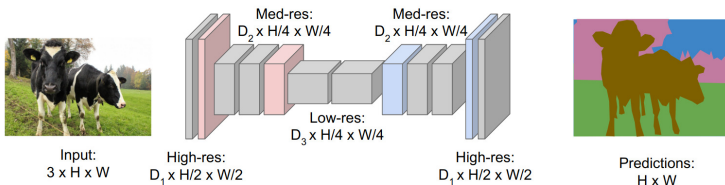


Expected Output (Source - Ref. 3)

Backbone network (kręgosłup) - wykorzystanie sieci do klasyfikacji obrazów (AlexNet, Googlenet, ResNet, VGG) jako część składowa FCN. Kręgosłup może być wstępnie wytrenowany na innym problemie analizy obrazów (np. ImageNet).

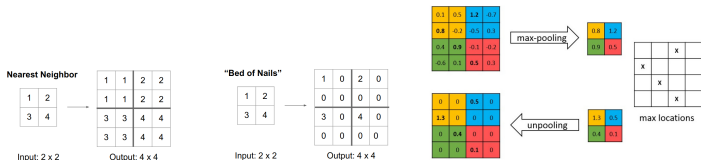
FCN: Downsampling and upsampling

- głębokie przetwarzanie obrazu o dużej rozdzielczości bardzo kosztowne
- FCN używa upsamplingu do odtworzenia oryginalnego rozmiaru obrazu

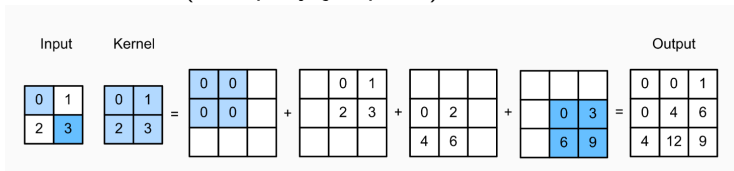


Unpooling i dekonwolucja

- unpooling



- deconvolution (transpozycja spłotu)



RNN

Sieci rekurencyjne i
modelowanie sekwencji

Modelowanie sekwencji

Sieć jednokierunkowa realizuje odwzorowanie jeden do jednego

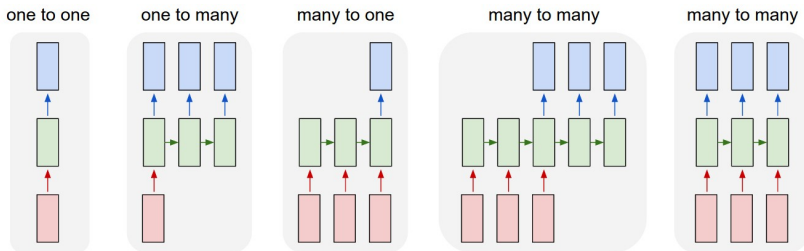
$$F(\mathbf{x}_i; \theta) = \mathbf{y}_i$$

Modelując szeregi czasowe chcemy uzyskać odpowiedź na podstawie sekwencji historycznych wartości

$$F(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_t; \theta) = \mathbf{y}$$

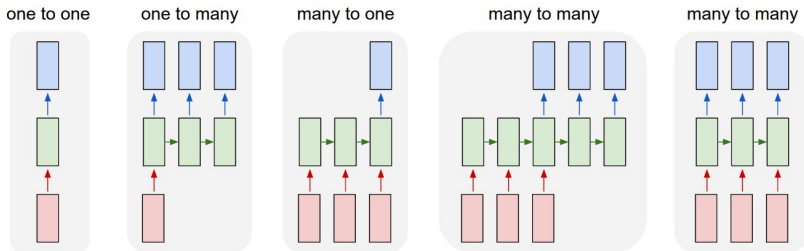
- Dane sekwencyjne lub zależne od czasu: sygnał audio, tekst, ruch obiektów, EEG, ...
- Modelowanie sekwencji: przewidywanie kolejnych wartości szeregu (np. przewidywanie pogody, notowań giełdowych), generowanie tekstu (chatboty), generowanie muzyki, sterowanie ruchem robota, ...
- Przetwarzanie sekwencji: maszynowe tłumaczenie, rozpoznawanie mowy, rozpoznawanie pisma, przetwarzanie strumieni wideo, ...
- Klasyfikacja sygnałów czasowych, etykietowanie obrazów, wykrywanie wzorców w sekwencjach

Modelowanie sekwencji



- *one-to-one*: pojedynczy sygnał wejściowy produkuje pojedynczy sygnał wyjściowy, brak informacji o sekwencji
- *one-to-many*: pojedyncze wejście generuje sekwencję na wyjściu, np. opis obrazów
- *many-to-one*: sekwencja wejściowa produkuje pojedynczy sygnał wyjściowy, np. klasyfikacja tekstów

Modelowanie sekwencji



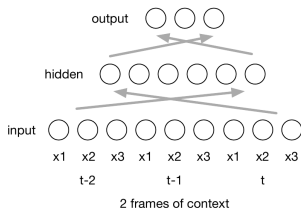
many-to-many: wejście i wyjście sekwencyjne

- bez dopasowania ramek, długość sekwencji wejściowej jest różna od długości sekwencji wyjściowej, np. tłumaczenie maszynowe, rozpoznawanie mowy
- z dopasowaniem ramek (*alignment*), np: etykietowanie ramek wideo, klasyfikacja fonemów w sygnale mowy (model akustyczny)

Modelowanie sekwencji sieciami jednokierunkowymi

Sieci MLP: dodanie **stałego kontekstu** do wejścia,
przykład: 2 ramki z lewej i prawej (poprzednie i przyszłe wektory wejściowe)

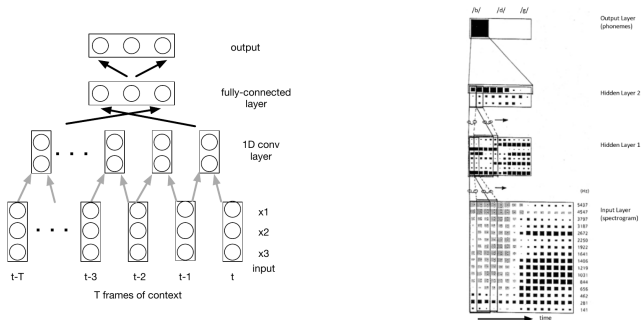
$$\hat{\mathbf{x}}^t \leftarrow [\mathbf{x}^{t-2}, \mathbf{x}^{t-1}, \mathbf{x}^t, \mathbf{x}^{t+1}, \mathbf{x}^{t+2}]$$



- zwiększenie wymiaru wejściowego istotnie zwiększa złożoność modelu, więc potrzebna jest większa liczba próbek uczących
- modelowanie tylko zależności krótkiego zasięgu (*short term memory*), ograniczonych wielkością okna kontekstu

Uczenie sekwencji z MLP i CNN

Sieci CNN: spłot względem wymiaru czasu (spłot 1D), stały kontekst odpowiada wielkości efektywnego pola recepcyjnego w wymiarze czasu.

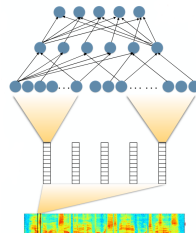


TDNN (Waibel, 1987), *time-delay neural network*, model ze spłotem 1D w domenie czasu.

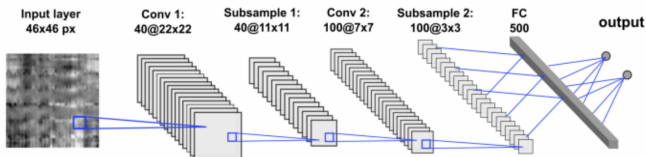
Źródło grafik: Steve Renals, *Machine Learning Practical — MLP Lecture 9*
Alexander Waibel, *Phoneme Recognition Using Time-Delay Neural Networks*, 1987.

Przykłady

Model akustyczny, rozpoznawanie fonemów z nagrań audio



Analiza sygnałów EEG, CNN + widma EEG

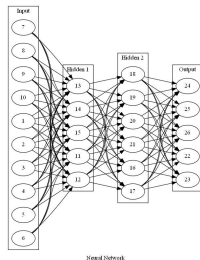


Nurse, E., Mashford, B. S., Yepes, A. J., Kiral-Kornek, I., Harrer, S., & Freestone, D. R. (2016). Decoding EEG and LFP Signals using Deep Learning: Heading TrueNorth

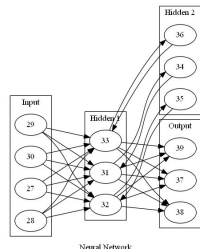
RNN - Sieci rekurencyjne

- **Sieci jednokierunkowe** używają jedynie kontekstu o skończonej długości, nie są w stanie wykryć zależności w dłuższych okresach czasu (*long-term*)
- **Sieci rekurencyjne (RNN)** posiadają połączenia do wcześniejszych warstw, wyjście w chwili t zależy od stanu z czasu $t - 1$ oraz z kroków poprzednich dzięki rekurencji

Sieć jednokierunkowa

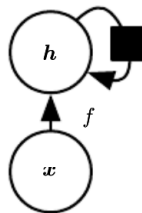


Sieć rekurencyjna Elmana



Stan ukryty jednostek RNN

$$\mathbf{h}^t = f(\mathbf{h}^{t-1}, \mathbf{x}^t; \theta)$$

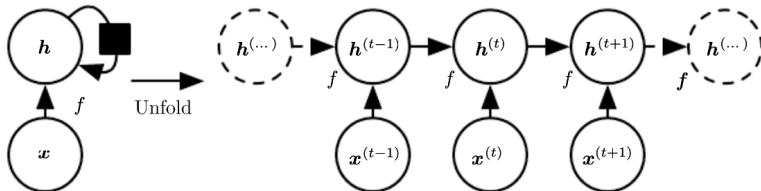


- neurony z połączeniami rekurencyjnymi działają jak pamięć, potencjalnie są w stanie modelować relacje występujące z dowolnie długim odstępem czasowym (nieskończony kontekst)

$$\mathbf{h}^t = g(\mathbf{x}^t, \mathbf{x}^{t-1}, \dots, \mathbf{x}^1)$$

- wektor $\mathbf{h}^t$ (aktywacje neuronów warstwy rekurencyjnej) tworzy skompresowaną reprezentację sekwencji
- stan początkowy $\mathbf{h}^0$ może być interpretowany jako dodatkowy sygnał wejściowy, typowo inicjowany wartością zero

Rozwinięcie rekurencji w czasie

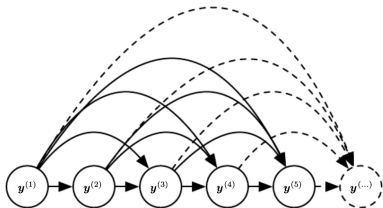


Rozwinięcie sieci RNN do postaci sieci jednokierunkowej, gdzie wagi połączeń są współdzielone.

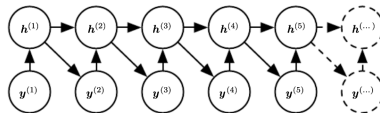
Głębokie uczenie w sieciach RNN: od pierwszego sygnału $\mathbf{x}^1$ do wyjścia $\mathbf{h}^t$ jest wiele warstw

Efektywność kodowania

- Sieć RNN z rekurencją w warstwach ukrytych spełnia wymagania **uniwersalnej maszyny Turinga**
- Ilość parametrów opisujących stan $\mathbf{h}$ ma wpływ na wielkość pamięci ale nie musi zależeć od długości sekwencji

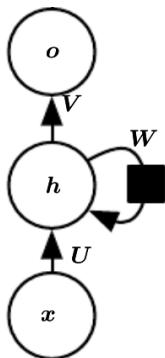


pełny model sekwencji, ilość parametrów $O(k^\tau)$



RNN, stan obecny zależy pośrednio od stanów z przeszłości, ilość parametrów stała względem długości sekwencji $O(1)$

Sieć z jedną warstwą rekurencyjną



$$o_i^{(t)} = \text{softmax}_i \left(\sum_{r=1}^k v_{ir} h_r^{(t)} + b_i \right)$$

$$h_i^{(t)} = \sigma \left(\sum_{j=1}^d u_{ij} x_j^{(t)} + \sum_{r=1}^k w_{ir} h_r^{(t-1)} + c_i \right)$$

To samo zwięźlej w zapisie macierzowym

$$\mathbf{o}^{(t)} = \text{softmax} \left(\mathbf{V} \mathbf{h}^{(t)} + \mathbf{b} \right)$$

$$\mathbf{h}^{(t)} = \sigma \left(\mathbf{U} \mathbf{x}^{(t)} + \mathbf{W} \mathbf{h}^{(t-1)} + \mathbf{c} \right)$$

$\mathbf{x}^{(t)} \in \mathbb{R}^d$ sygnał wejściowy w chwili t

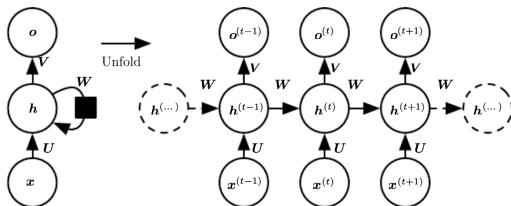
$\mathbf{U} \in \mathbb{R}^{d \times k}$, $\mathbf{W} \in \mathbb{R}^{k \times k}$

d ilość wejść sieci

k ilość neuronów w warstwie ukrytej

Wsteczna propagacja w czasie

Rozwinięta sieć odpowiada głębokiej sieci jednokierunkowej ze współdzielonymi wagami $\mathbf{W}$, $\mathbf{V}$, $\mathbf{U}$

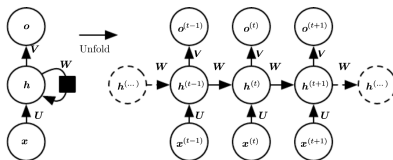


BPTT *backpropagation through time* - trening za pomocą wstecznej propagacji sieci „rozwiniętej w czasie”, odbywa się analogicznie jak w jednokierunkowych sieciach.

Aktualizacja wagi w chwili t wymaga:

- całej historii sygnałów przeszłych $\mathbf{x}^1, \dots, \mathbf{x}^t$
- historii stanów neuronów $\mathbf{h}^1, \dots, \mathbf{h}^t$ i ich pochodnych $\mathbf{h}'^1, \dots, \mathbf{h}'^t$

Wsteczna propagacja w czasie



Funkcja kosztu dla całej sekwencji (model *many-to-many*):

$$L\left(\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(\tau)}\}, \{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(\tau)}\}\right) = \frac{1}{\tau} \sum_{t=1}^{\tau} L^{(t)}$$

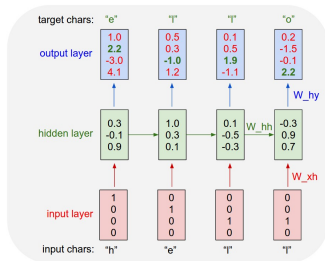
$L^{(t)}$ - koszt w kroku t (np. cross-entropy) zależny od $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(t)}$

$$\partial_w L = \sum_{t=1}^{\tau} \partial_w L\left(y^{(t)}, o^{(t)}\right)$$

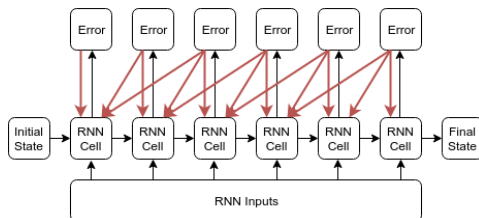
Aktualizacje wag są akumulowane dla sygnału błędu propagowanego w kolejnych krokach czasu

Przykład: CharRNN

- model języka oparty na znakach, wejście i wyjście stanowi sekwencja znaków
- znaki kodowane z pomocą wektorów binarnych o długości równej ilości wszystkich znaków w alfabecie (kodowanie *one-hot*)
- problem klasyfikacji przewidywania kolejnego znaku w sekwencji $f(x(t)) \rightarrow x(t+1)$
- wyjście softmax modeluje rozkład prawdopodobieństwa $P(x_{t+1} | x_1, x_2, \dots, x_t)$

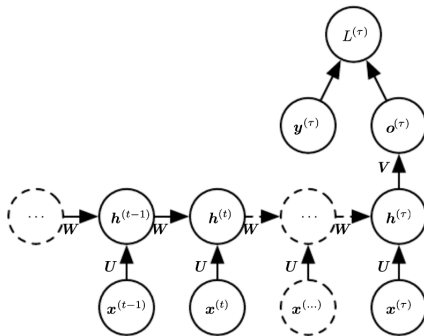


Wsteczna propagacja w czasie



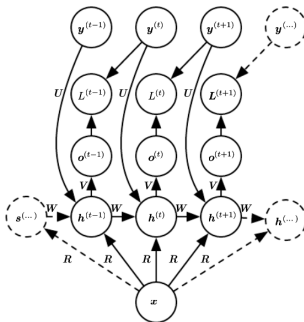
- Rozwinięta w czasie sieć tworzy głęboką sieć neuronową ze wszystkimi tego konsekwencjami (zanikający gradient dla relacji odległych w czasie, wybuchający gradient)
- Obliczenia dla długich sekwencji są kosztowne
- Nie zawsze potrzebujemy informacji z odległej przeszłości
- **Truncated BPTT** - ograniczenie długości sekwencji wstecznej propagacji do stałej wartości T (np. $T = 20$). Gradient z dalszej przeszłości $t - T - 1$ równy 0
- problem z wykrywaniem odległych relacji (*long-term*) przez obcięcie T oraz zanikający gradient

Pojedyncze wyjście na końcu sekwencji



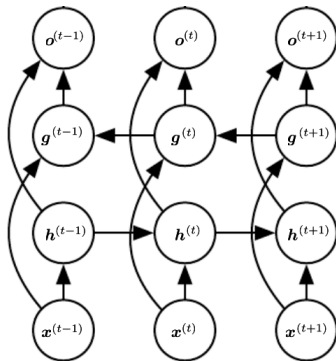
Sieć produkuje wyjście po obejrzeniu całej sekwencji,
np. klasyfikacja wpisów w portalach społecznościowych
(tzw. analiza sentymentów)

Pojedyncze wejście generujące sekwencję



- Sieć produkuje sekwencję wyjść przy prezentacji stałego wzorca (np. etykietowanie zdjęć, opis sceny)
- Sygnał wejściowy x pełni rolę kontekstu
- Modelowanie sekwencji $y^t = f(y^{t-1}; x)$

Dwukierunkowa sieć RNN



Sieci dwukierunkowe RNN łączą warstwy z rekurencją zależną od sygnału przeszłego **h** z rekurencją zależną od przyszłego sygnału **g**

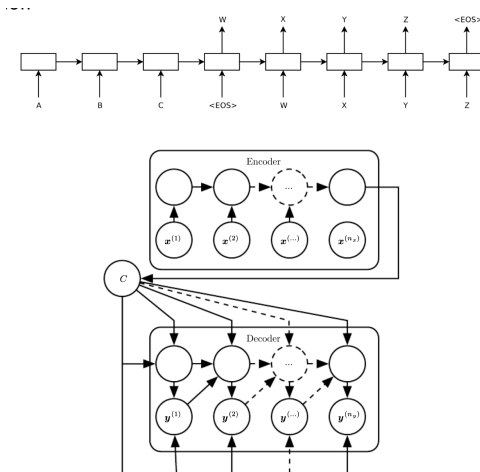
Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks. Signal Processing, 1997

Dwukierunkowa sieć RNN

- informacja z przyszłości może zawierać istotną informację, np. dźwięki kolejnych fonemów (koartykulacja), znaczenie słów w zdaniu
- wyjścia sieci w dowolnym momencie zależą od całej sekwencji wejściowej, możliwe tylko uczenie *offline*
- zastosowanie: rozpoznawanie mowy (Graves, 2008, 2009), bioinformaryka (Baldi, 1999), rozpoznawanie pisma (Liwicki, Marcus, 2007)
- dla obrazów 2D możliwe zastosowanie sieci 4 kierunkowych, co pozwala modelować odległe relacje, jednak nie są one tak wydajne w tych zastosowaniach jak CNN

Model Encoder-Decoder

Transformacja $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n_x)} \Rightarrow \mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(n_y)}$



Sutskever et al (2014). "Sequence to Sequence Learning with Neural Networks", NIPS.

K Cho et al (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation", EMNLP.

Model Encoder-Decoder

seq2seq (Sutskever, 2014) to dwie sieci RNN uczone jednocześnie:

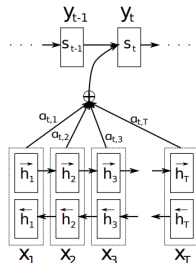
- **enkoder** (reader, input) przetwarza sekwencję wejściową generując kontekst C
- **dekoder** (writer, output) zależnie od kontekstu C generuje sekwencję wyjściową
- wielkość kontekstu C może ograniczyć zdolność sieci do reprezentacji dalekich w czasie relacji
- dodanie uwagi (*attention model*, Bahdanan, 2015) - uwzględnienie w C więcej informacji z enkodera, które skalują wyjścia dekodera

Model Encoder-Decoder z systemem uwagi

- **enkoder** dwukierunkowa sieć RNN, przetwarza całą sekwencję, generuje wektory stanów w kolejnych chwilach $\mathbf{h}_i$
- kontekst uwagi $\mathbf{c}$ jest średnią ważoną stanów, wagi a_i podlegają uczeniu i są unormowane

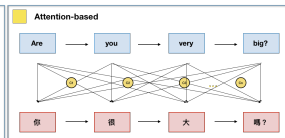
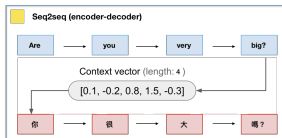
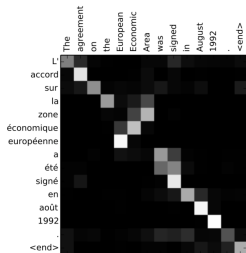
$$\mathbf{c} = \sum_{i=0}^T a_i \mathbf{h}_t \quad a_i = \frac{\exp e_i}{\sum_j \exp e_j}$$

- mechanizm uwagi waży wpływ kontekstu w wybranym momencie sekwencji wejściowej
- **dekoder** emuluje przeszukiwanie sekwencji wejściowej podczas transkrypcji, macierz uwagi A_t jest łączona ze stanem $\mathbf{h}_{t-1}$

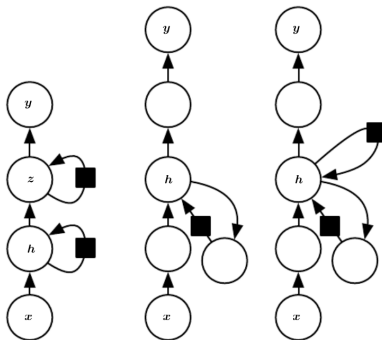


$$\mathbf{x}_t^{\text{decoder}} = \left[\mathbf{h}_{t-1}^{\text{decoder}} ; A_t^{\text{decoder}} \right]$$

seq2seq z mechanizmem uwagowym



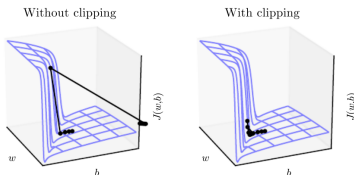
Deep RNN



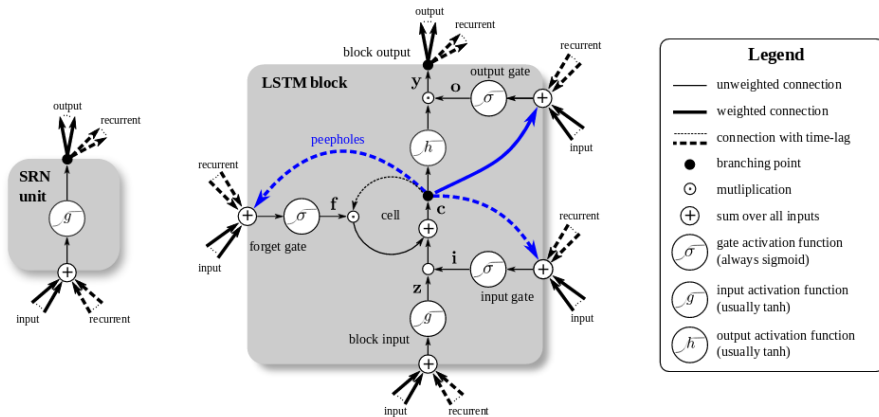
Wielowarstwowe RNN - kolejne warstwy tworzą hierarchię reprezentacji (Graves, 2013)

Problemy uczenia głębokich sieci RNN

- znikający gradient dla odległych w czasie relacji (zanika wykładniczo z czasem), rozwiązaniem LSTM
- niestabilność uczenia - eksplodujący gradient, rozwiązanie: przycinanie gradientu i regularyzacja
- **gradient clipping** - ograniczenie wartości gradientu dla wszystkich wymiarów w mini-batchu, może zmienić kierunek spadku gradientu
- **norm clipping**: jeśli $\|g\| > \nu$ to $g \leftarrow \frac{g\nu}{\|g\|}$



Long Short-Term Memory



Sepp Hochreiter, Jürgen Schmidhuber, „Long short-term memory”, 1997

LSTM

kandydat nowego stanu

$$\mathbf{z}^t = g(\mathbf{W}_z \mathbf{x}^t + \mathbf{R}_z \mathbf{y}^{t-1} + \mathbf{b}_z)$$

bramka wejściowa

$$\mathbf{i}^t = \sigma(\mathbf{W}_i \mathbf{x}^t + \mathbf{R}_i \mathbf{y}^{t-1} + \mathbf{b}_i)$$

bramka zapominania

$$\mathbf{f}^t = \sigma(\mathbf{W}_f \mathbf{x}^t + \mathbf{R}_f \mathbf{y}^{t-1} + \mathbf{b}_f)$$

stan wewnętrzny

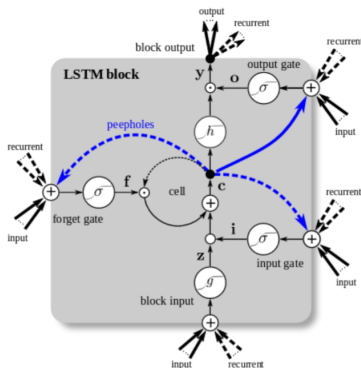
$$\mathbf{c}^t = \mathbf{i}^t \odot \mathbf{z}^t + \mathbf{f}^t \odot \mathbf{c}^{t-1}$$

bramka wyjściowa

$$\mathbf{o}^t = \sigma(\mathbf{W}_o \mathbf{x}^t + \mathbf{R}_o \mathbf{y}^{t-1} + \mathbf{b}_o)$$

sygnał wyjściowy

$$\mathbf{y}^t = \mathbf{o}^t \odot h(\mathbf{c}^t)$$



Sepp Hochreiter, Jürgen Schmidhuber, „Long short-term memory”, 1997

LSTM

- LSTM niweluje wpływ zanikania gradientu, dzięki czemu pozwala mapować odległe w czasie relacje
- wprowadza dodatkową wewnętrzną jednostkę stanu $\mathbf{c}$
- stan $\mathbf{c}^t$ zależy w sposób liniowy od stanu poprzedniego $\mathbf{c}^{t-1}$ (liniowa rekurencja), więc nawet przy bardzo długich rekurencjach gradient ma szansę nie zaniknąć (*long-term memory*)
- iloczyn Hadamarda $\mathbf{x} \odot \sigma(\mathbf{y})$ (po współrzędnych), realizacja bramki wyłączającej/włączającej sygnał $\mathbf{x}$
- bramki wejściowa, wyjściowa i bramka zapominania sterują zapamiętywanymi informacjami, zależą od sygnału wejściowego i stanu poprzedniego
- funkcja różniczkowalna - można uczyć metodami spadku gradientu i wsteczną propagacją

LSTM

- bramka wejściowa ma wpływ na to jaki sygnał zostanie zapisany w jednostce stanu $\mathbf{c}$
- bramka zapominania steruje wpływem poprzedniego stanu $\mathbf{c}^{t-1}$ na obecny stan $\mathbf{c}^t$
- bramka wejściowa i zapominania wpływają na to co umieszczane jest i co jest usuwane ze stanu
- bramka wyjściowa steruje ilością informacji przepływającą ze stanu wewnętrznego do wyjścia jednostki, pozwala np. zachować na później informacje, które w tym momencie nie są istotne

Peephole LSTM

$$\mathbf{z}^t = g(\mathbf{W}_z \mathbf{x}^t + \mathbf{R}_z \mathbf{y}^{t-1} + \mathbf{b}_z)$$

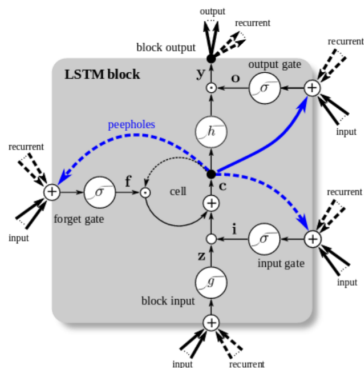
$$\mathbf{i}^t = \sigma(\mathbf{W}_i \mathbf{x}^t + \mathbf{R}_i \mathbf{y}^{t-1} + \mathbf{p}_i \odot \mathbf{c}^{t-1} + \mathbf{b}_i)$$

$$\mathbf{f}^t = \sigma(\mathbf{W}_f \mathbf{x}^t + \mathbf{R}_f \mathbf{y}^{t-1} + \mathbf{p}_f \odot \mathbf{c}^{t-1} + \mathbf{b}_f)$$

$$\mathbf{c}^t = \mathbf{i}^t \odot \mathbf{z}^t + \mathbf{f}^t \odot \mathbf{c}^{t-1}$$

$$\mathbf{o}^t = \sigma(\mathbf{W}_o \mathbf{x}^t + \mathbf{R}_o \mathbf{y}^{t-1} + \mathbf{p}_o \odot \mathbf{c}^t + \mathbf{b}_o)$$

$$\mathbf{y}^t = \mathbf{o}^t \odot h(\mathbf{c}^t)$$

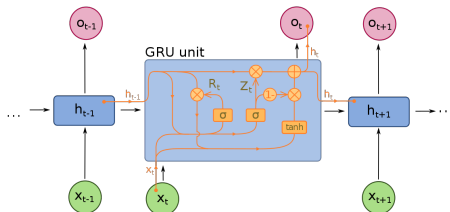


aktywacja bramek uzależniona dodatkowo od poprzedniego stanu $\mathbf{c}^{t-1}$ poprzez połączenia „podglądające” $\mathbf{p}_i$, $\mathbf{p}_o$, $\mathbf{p}_f$

Sepp Hochreiter, Jürgen Schmidhuber, „Long short-term memory”, 1997

Gated Recurrent Unit

- GRU (Kyunghyun Cho et al., 2014) brak jednostki wewnętrznej c i bramki wyjściowej
- bramki: resetu r_t i aktualizacji z_t
- mniej parametrów względem LSTM, porównywalna jakość wyników dla wielu problemów



Źródło: https://en.wikipedia.org/wiki/Recurrent_neural_network

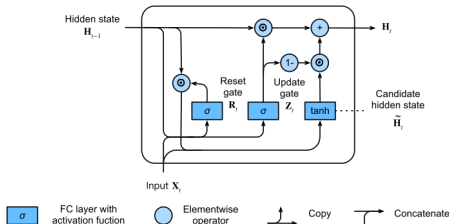
Gated Recurrent Unit

$$\mathbf{z}^t = \sigma(\mathbf{W}_z \mathbf{x}^t + \mathbf{U}_z \mathbf{h}^{t-1} + \mathbf{b}_z)$$

$$\mathbf{r}^t = \sigma(\mathbf{W}_r \mathbf{x}^t + \mathbf{U}_r \mathbf{h}^{t-1} + \mathbf{b}_r)$$

$$\tilde{\mathbf{h}}^t = \tanh(\mathbf{W}_h \mathbf{x}^t + \mathbf{U}_h (\mathbf{r}^t \circ \mathbf{h}^{t-1}) + \mathbf{b}_h)$$

$$\mathbf{h}^t = (1 - \mathbf{z}^t) \circ \mathbf{h}^{t-1} + \mathbf{z}^t \circ \tilde{\mathbf{h}}^t$$

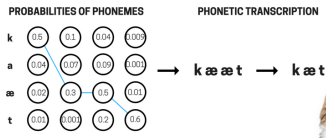


- stan proponowany $\tilde{\mathbf{h}}^t$, dla $\mathbf{r}^t = 0$ sprowadza się do warstwy MLP
- bramka aktualizacji $\mathbf{z}^t$ określa stopień wpływu poprzedniego stanu $\mathbf{h}^{t-1}$ oraz obecnego stanu proponowanego $\tilde{\mathbf{h}}^t$
- bramka resetu odpowiada za relacje krótkie w czasie a bramka aktualizacji - za długie

Connectionist Temporal Classification

- CTC to metoda modelowania sekwencji oraz nazwa funkcji kosztu, znajduje zastosowanie, gdy sekwencja wejściowa $\mathbf{x}_1, \dots, \mathbf{x}_T$ nie posiada dostępnego dopasowania ramek (*alignementu*) z sygnałem wyjściowym $\mathbf{y}_1, \dots, \mathbf{y}_K$. Długość sekwencji wejściowej T jest dłuższa od sekwencji wyjściowej ($T > K$), np. transkrypcja tekstu na podstawie nagrań audio
- wyjście softmax uzupełnione o dodatkową etykietę 'blank'
 $\mathcal{L}' = \mathcal{L} \cup \{blank\}$
- prawdopodobieństwo $P(\mathbf{y}'|\mathbf{x})$ zaobserwowania sekwencji etykiet (ścieżki) $\mathbf{y}'$ o długości T

$$P(\mathbf{y}'|\mathbf{x}) = \prod_{t=1}^T P(y'_t|\mathbf{x})$$



Connectionist Temporal Classification

CTC poszukuje sekwencji etykiet o największej wiarygodności $P(\mathbf{y}|\mathbf{x})$ marginalizując po wszystkich możliwych ścieżkach $\mathbf{y}'$ z blankiem

$$p(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{y}' \in \mathcal{M}^{-1}(\mathbf{y})} P(\mathbf{y}'|\mathbf{x})$$

gdzie $\mathcal{M}$ mapuje sekwencję etykiet o długości T do sekwencji krótszej poprzez usunięcie wszystkich blanków i powtarzających się etykiet, np.: $\mathcal{M}(a - ab -) = \mathcal{M}(-aa - -abb) = aab$

Funkcja kosztu

$$L = - \sum \log p(\mathbf{y}|\mathbf{x})$$

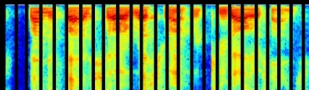
Graves, A., Fernández, S., Gomez, F., Schmidhuber, J., Connectionist temporal classification: labeling unsegmented sequence data with recurrent neural networks. 2006.

Connectionist Temporal Classification (CTC)

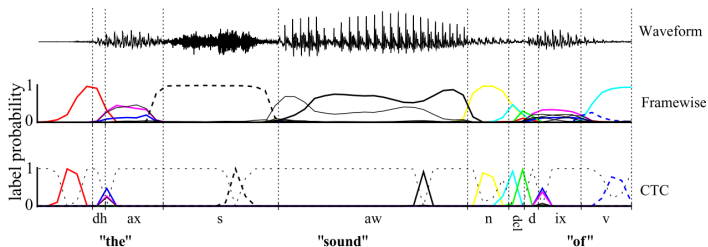
Deep Speech - CTC

No alignment needed!

$$\begin{array}{c} P(_ _ \text{TH} _ _ _ _ \text{E} _ _ _ _ \text{C} _ _ \text{AAA} _ _ \text{TT} _ _ _) \\ + \\ \vdots \\ + \\ P(_ \text{T} _ _ _ _ \text{H} _ _ _ _ \text{EE} _ _ _ _ _ _ _ _ \text{C} _ _ _ _ \text{AA} _ _ _ _ \text{T} _ _ _ _ _) \end{array} \quad \left. \vphantom{\begin{array}{c} P(_ _ \text{TH} _ _ _ _ \text{E} _ _ _ _ \text{C} _ _ \text{AAA} _ _ \text{TT} _ _ _) \\ + \\ \vdots \\ + \\ P(_ \text{T} _ _ _ _ \text{H} _ _ _ _ \text{EE} _ _ _ _ _ _ _ _ \text{C} _ _ _ _ \text{AA} _ _ _ _ \text{T} _ _ _ _ _) \end{array}} \right\} P(\text{THE} \text{---} \text{CAT} \text{---})$$

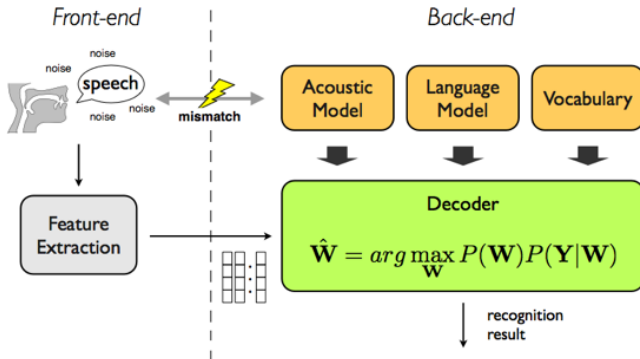


Predykcja CTC



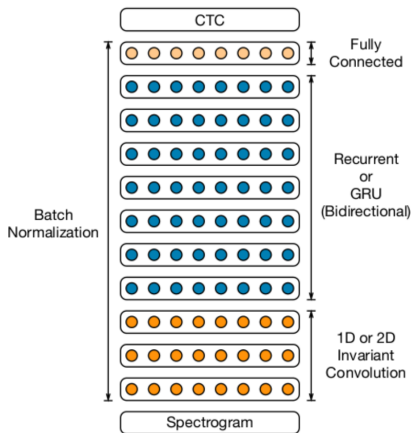
- trening z CTC generuje sekwencję pików etykiet (fonemów) przedzielonych etykietą 'blank'
- w treningu z alignmencem stosuje się funkcję kosztu CrossEntropy, przesunięcie granic fonemu względem poprawnego dopasowania powoduje błąd nawet gdy fonem jest poprawnie przewidziany
- wyjście CTC może być wykorzystane bezpośrednio (bez konieczności dalszego przetwarzania) do transkrypcji

Automatyczne rozpoznawanie mowy



Źródło grafiki: <https://www.esat.kuleuven.be/psi/spraak/theses/08-09-en/MDT.php>

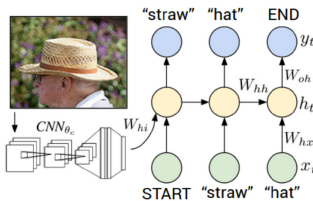
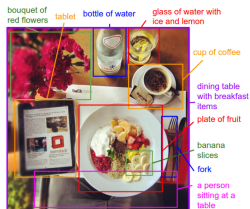
Przykład: Deep Speech 2



Read Speech			
Test set	DS1	DS2	Human
WSJ eval'92	4.94	3.60	5.03
WSJ eval'93	6.94	4.98	8.08
LibriSpeech test-clean	7.89	5.33	5.83
LibriSpeech test-other	21.74	13.25	12.69

Amodei, Dario, et al. "Deep speech 2: End-to-end speech recognition in english and mandarin."(2015).

Zastosowania: generator opisów zdjęć

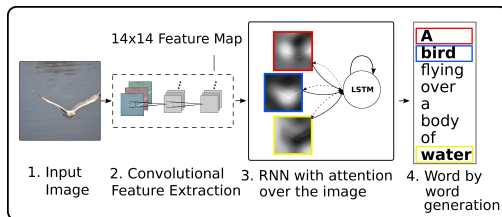


Multimodal Recurrent Neural Network NeuralTalk2

Karpathy A., Fei-Fei L. „Deep Visual-Semantic Alignments for Generating Image Descriptions”, 2015

Zastosowania: generator opisów zdjęć

Show, attend and tell (Xu et al., 2015)



↑ a living room with a couch and a television



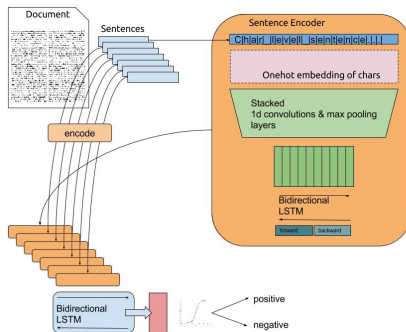
↑ a man riding a bike on a beach



a man is walking down the street with a suitcase ↗

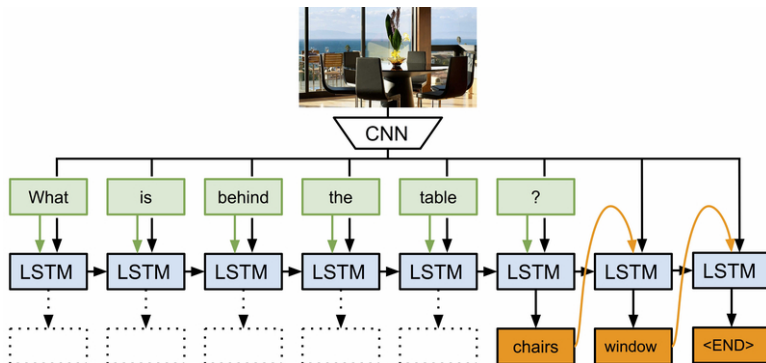
Mechanizm uwagi pozwala skupić się na wybranych fragmentach obrazu podczas generowania opisu

Zastosowania: analiza sentymentów



<https://offbit.github.io/how-to-read/>

Zastosowania: odpowiedzi na pytania



Uczenie nienadzorowane

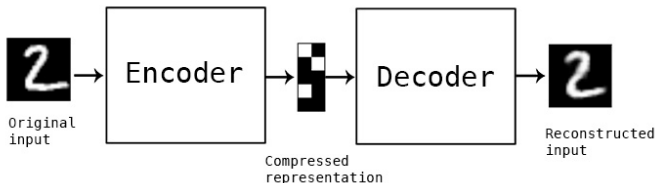
Autoenkodery

Modele generatywne

Uczenie nienadzorowane

- Uczenie **nadzorowane** (*supervised*) - dane wejściowe **X** posiadają spodziewane wyjście (etykiety) (**X, Y**)
- Uczenie **nienadzorowane** (*unsupervised*) używa wyłącznie danych wejściowych **X**. Dane bez etykiet są powszechnie występujące.
 - trening automatycznie wykrywa istotne cechy w danych,
 - modeluje rozkład prawdopodobieństwa danych
- Modele sieci:
 - autoenkodery (AE), splotowe AE (CAE), Variational AE (VAE)
 - Restricted Boltzman Machines (RBM)
 - Generative Adversarial Networks (GAN)
- Zastosowanie:
 - kodowanie sygnału, wykrywanie istotnych cech, usuwanie szumu i rekonstrukcja sygnału, głęboka klaseryzacja
 - generowanie sygnałów, transfer stylu
 - inicjowanie głębokich sieci DNN (stacked autoencoders)

Autoenkodery



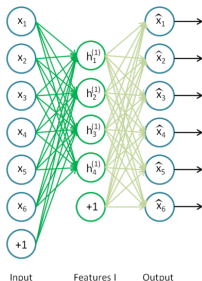
- autoenkodery uczą się odtwarzać sygnał wejściowy
- warstwy pośrednie kodują specyficzne cechy danych
- kompresja stratna - wymagamy aby model nie był jednostkowym odwzorowaniem
- AE uczą się reprezentacji danych (ekstrakcja cech), modelują rozkład prawdopodobieństwa sygnału
- koder i dekodek mogą być sieciami neuronowymi, całość uczona wsteczną propagacją błędu

Autoenkodery

Encoder:

$$\mathbf{h} = f(\mathbf{x})$$

$$= \sigma(\mathbf{W}\mathbf{x} + \mathbf{b})$$



Decoder:

$$\hat{\mathbf{x}} = g(\mathbf{h})$$

$$= \sigma(\mathbf{W}'\mathbf{h}(\mathbf{x}) + \mathbf{b}')$$

- **Autoencoder** - sieć jednokierunkowa, której celem jest rekonstrukcja sygnału wejściowego
- warstwa ukryta $\mathbf{h}$: koduje sygnał wejściowy, detektor cech, skompresowana reprezentacja danych
- często wymaga się aby $\mathbf{W}' = \mathbf{W}^T$, współdzielone wagi kodera i dekodera

Koszt rekonstrukcji

Funkcja rekonstrukcji $L(\mathbf{x}, \hat{\mathbf{x}})$

- MSE dla danych wejściowych rzeczywistych

$$L(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{2} \sum_k (\hat{x}_k - x_k)^2$$

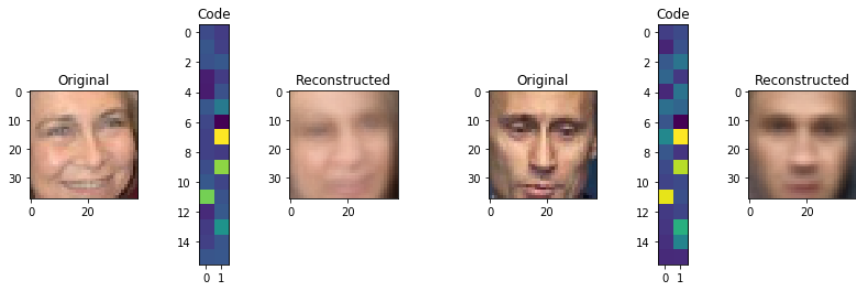
- liniowa aktywacja warstwy wyjściowej $\hat{\mathbf{x}} = \mathbf{W}'\mathbf{h} + \mathbf{b}'$
- Cross Entropy dla danych binarnych (lub z zakresu $[0, 1]$)

$$L(\mathbf{x}, \hat{\mathbf{x}}) = - \sum_k (x_k \log \hat{x}_k + (1 - x_k) \log (1 - \hat{x}_k))$$

- trening wsteczną propagacją błędu
- ograniczenia zapobiegające tworzeniu odwzorowania jednostkowego: niekompletny AE, regularyzowany AE, kurczliwy AE, ...

Niekompletny autokoder

- **Niekompletny AE** (*undercomplete*) - rozmiar warstwy **h** mniejszy od rozmiaru wejścia,
- kompresja sygnału do rozmiaru **h**
- dla liniowego dekodera wynik jest równoważny z PCA, wagi neuronów tworzą składowe główne maksymalizujące wariancję
- Przykład: input-output: 32x32 (1024 piksele) ,
hidden (code size): 32



Regularyzowany AE

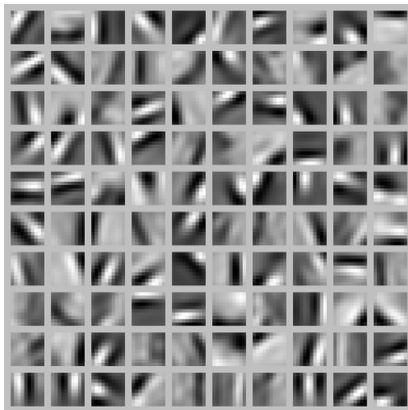
- wielkość $\mathbf{h}$ decyduje o pojemności sieci i jakości odwzorowania. Dla niekompletnych AE jest ograniczenie zwiększania rozmiaru warstwy ukrytej
- **Nadkompletny AE** (*overcomplete*) gdy rozmiar $\mathbf{h}$ większy od rozmiaru sygnału wejściowego $\mathbf{x}$,
 - brak kompresji
 - nie gwarantują uzyskania wartościowych reprezentacji bez dodania odpowiedniej regularyzacji wymuszającej odpowiednią reprezentację w $\mathbf{h}$ (np. rzadkość)
- **Rzadki autoencoder** z czynnikiem kary za brak rzadkiej reprezentacji

$$L(\mathbf{x}, \hat{\mathbf{x}}) + \lambda \sum |h_i|$$

- w celu uzyskania rzadkich aktywacji można też użyć ReLU (Glorat, 2011)

Przykład

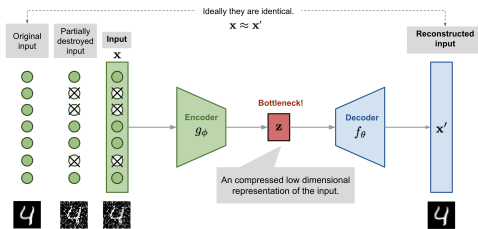
Cechy uzyskane w warstwie ukrytej przez autoenkoder rzadki (regularyzacja L_1) trenowany na naturalnych obrazach (bez etykiet!)



Oja, E. (1982). A simplified neuron model as a principal component analyzer. Journal of mathematical biology, 15(3), 267–73.

Autoenkoder z odszumianiem (DAE)

- **Denoise autoencoder (DAE)** usuwa szum (lub inne zniekształcenia) sygnału wejściowego.
- Sygnał wejściowy $\tilde{x}$ jest zniekształconą kopią x
- AE uczy się odtwarzać oryginalny sygnał bez deformacji



src: Building Autoencoders in Keras

From Autoencoder to Beta-VAE

- **Contractive autoencoder (CAE)**

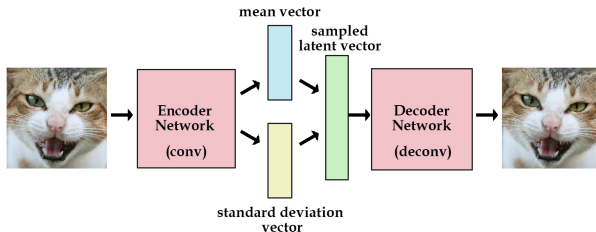
$$L(\mathbf{x}, \hat{\mathbf{x}}) + \lambda \|\nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x})\|_F^2$$

- kara za dużą zmienność reprezentacji przy zmianie sygnału (norma Frobeniusa)

$$\|\nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x})\|_F^2 = \sum_i \sum_j \left(\frac{\partial h_i(\mathbf{x})}{\partial x_j} \right)^2$$

Wariacyjny AE

- **Variational autoencoders (VAE)** jawnie modeluje rozkład gęstości danych, kodowana reprezentacja to parametry funkcji gęstości, np. σ_i, μ_i rozkładu normalnego $N(\mu_i, \sigma_i)$



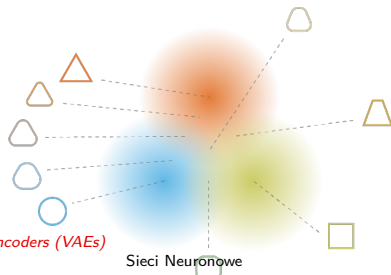
- sygnał wejściowy dekodera jest próbkowany z rozkładu określonego przez parametry warstwy kodującej
- model generatywny - umożliwia próbkowanie danych wyjściowych, tworzenie nowych sygnałów o charakterystyce zbliżonej do danych treningowych

Kullback–Leibler divergence

- czynnik regularyzujący, dodawany do funkcji kosztu (np. MSE), wymuszający odpowiedni rozkład danych w warstwie kodującej (KL divergence), np. preferujący rozkład $N(0, 1)$

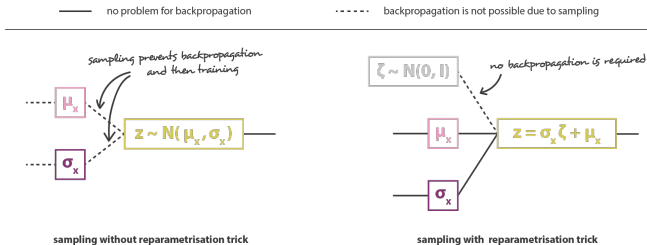
$$\sum_{i=1}^n \sigma_i^2 + \mu_i^2 - \log(\sigma_i) - 1$$

- osiąga minimum dla $\sigma_i = 1, \mu_i = 0$
- wymusza rozkład zakodowanych danych w centrum przestrzeni ukrytej, minimalizacja kosztu rekonstrukcji konkuruje z regularyzacją próbując rozdzielić grupy różnych danych



Representation trick (stochastyczne BP)

W jaki sposób zrealizować wsteczną propagację błędu?



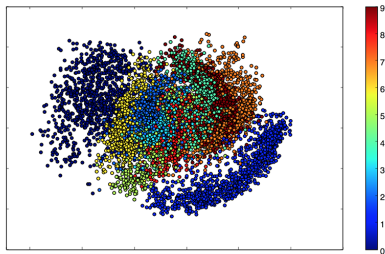
$$\mathbf{z} = \mu_{\mathbf{x}} + \sigma_{\mathbf{x}}\xi$$

Próbka $\xi \sim \mathcal{N}(0, 1)$ stanowi kolejne wejście sieci neuronowej jednak nie posiadające parametrów optymalizacyjnych (nie propagujemy błędu tym wejściem).

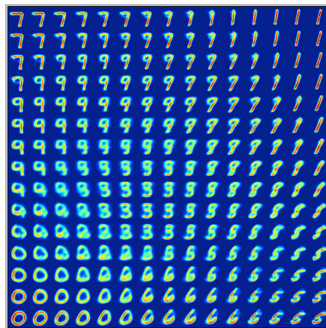
Całą sieć można więc trenować wsteczną propagacją.

VAE na MNIST

wizualizacja danych treningowych
w 2 wymiarowej warstwie ukrytej
kodera



generowanie cyfr dla siatki 15x15
wartości 2 wymiarowej warstwy
ukrytej dekodera

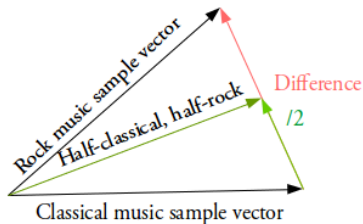


Architektura: Input-512-2- (μ, σ) -2-512-Output

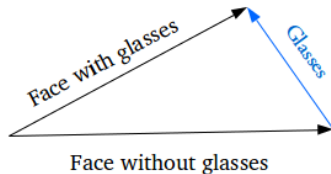
src:  Building Autoencoders in Keras

Interpolowanie i generowanie pojęć

Arytmetyka w przestrzeni zmiennych utajonych



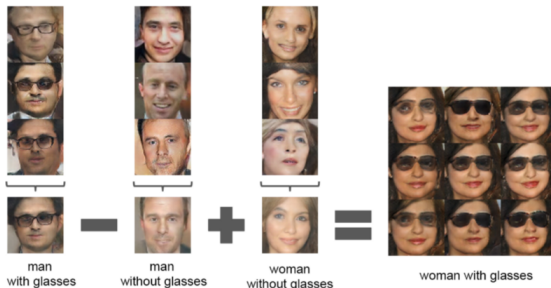
Środek między wektorami μ
kodującymi pojęcia



Różnica pomiędzy pojęciami
może być dodana do innego
pojęcia

Ukryta reprezentacja pojęć

- wyuczona reprezentacja danych koduje pojęcia
- arytmetyczne operacje na wektorach kodujących odpowiadają relacjom semantycznym pomiędzy pojęciami



Przykład: generowanie twarzy

Wejście



Rekonstrukcja



Losowe

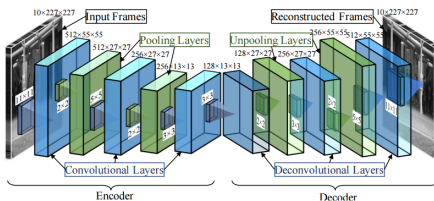


👉 Deep Feature Consistent Variational Autoencoder

📄 Variational auto-encoder trained on celebA

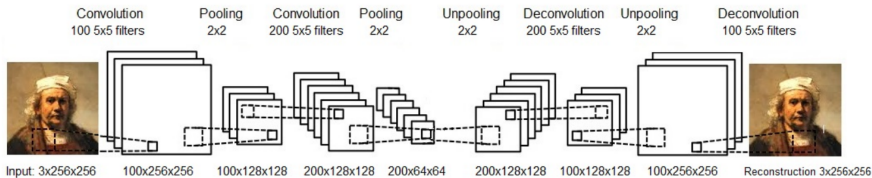
Inne typy autoenkoderów

- Convolutional AE (CAE), dekodery jest symetryczną względem enkodera siecią dekonwolucyjną

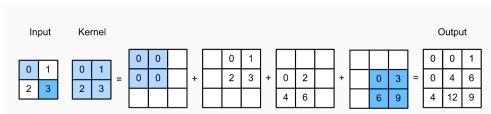


- sequence-to-sequence: AE z użyciem jednostek rekurencyjnych (np. LSTM)
- głębokie autoenkodery - więcej warstw sprzyja kompresji danych o złożonych rozkładach
- stacked autoencoders - tworzenie głębokich sieci z uczonych sekwencyjnie autoenkoderów (metoda inicjalizacji)

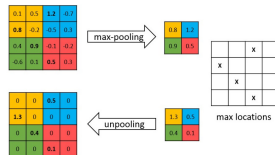
Deep Convolutional Autoencoder



deconvolution
transpozycja spłotu



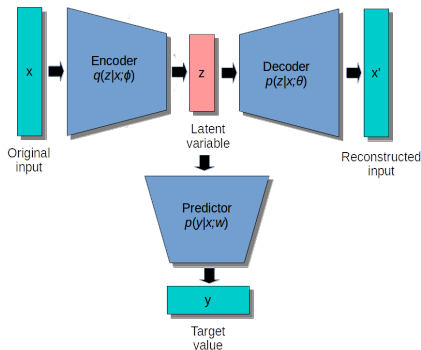
unpooling



Omid E. David, Nathan S. Netanyahu, DeepPainter: Painter Classification Using Deep Convolutional Autoencoders

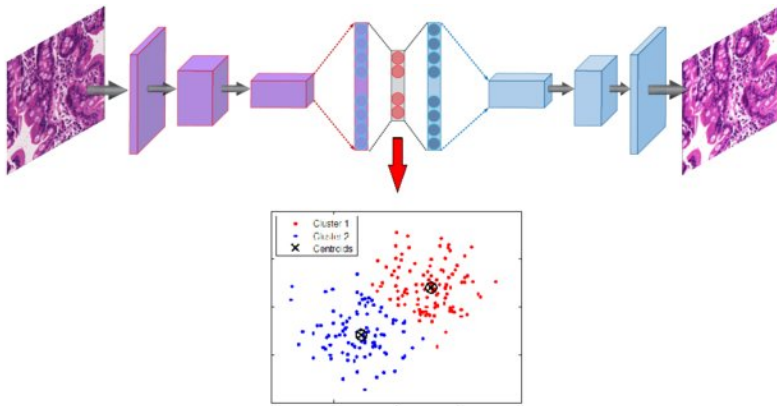
Uczenie pół-nadzorowane (samouczenie)

Uczenie pół-nadzorowane (**semi-supervised**) lub samouczenie (**self-taught**): najpierw trenujemy AE na dużym zbiorze danych w sposób nienadzorowany, następnie reprezentacja danych uzyskana z kodera wykorzystywana jest do uczenia płytkiego klasyfikatora na małym zbiorze z etykietami



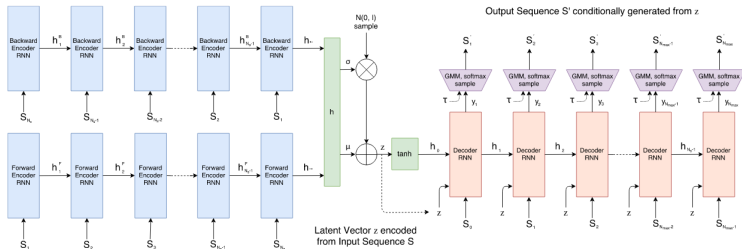
Guo et al., 2017, Deep convolutional autoencoder

Głębokie grupowanie



Przykład: Sketch-RNN

Sequence-to-Sequence Variational Autoencoder do trenowania reprezentacji rysowania odręcznego

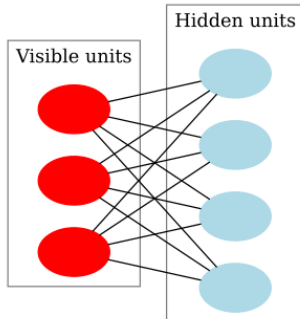


👉 Draw Together with a Neural Network

David Ha, Douglas Eck, A Neural Representation of Sketch Drawings

Restricted Boltzman Machine

- **Restricted Boltzman Machines (RBM)** (Hinton 2000), wcześniej Harmonium (P. Smolensky, 1986), to szczególny przypadek Maszyny Boltzmana
- stochastyczna sieć neuronowa modelująca rozkład prawdopodobieństw wejść
- *restrictive* - brak połączeń wewnątrz warstwy
- algorytm uczenia: *contrastive divergence* (Hinton)



Źródło grafiki: Wikipedia

Budowa RBM

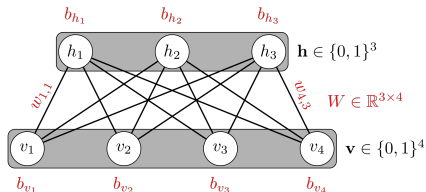
- binarne jednostki umieszczone w dwóch warstwach: widzialnej $\mathbf{v}$ i ukrytej $\mathbf{h}$
- sygnał wejściowy podawany do jednostki widzialnej $\mathbf{v} = \mathbf{x}$
- energia układu

$$E(\mathbf{x}, \mathbf{h}) = -\sum a_i x_i - \sum b_i h_i - \sum \sum x_i w_{ij} h_j$$

gdzie w_{ij} to wagi połączeń między warstwami, a_i i b_i to wyrazy wolne związane z warstwą widzialną i ukrytą

- prawdopodobieństwo rozkładu

$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{x}, \mathbf{h})}$$



src: <https://kwonkyo.wordpress.com/>

- rozkład brzegowy

$$P(\mathbf{x}) = \frac{1}{Z} \sum_{\mathbf{h}} e^{-E(\mathbf{x}, \mathbf{h})}$$

- brak połączeń wewnątrz warstw, więc stany w warstwach są niezależne. Prawdopodobieństwo aktywacji:

$$P(\mathbf{h}|\mathbf{x}) = \prod_{j=1}^n P(h_j|\mathbf{x}), \quad P(\mathbf{x}|\mathbf{h}) = \prod_{i=1}^m P(x_i|\mathbf{h})$$

$$P(h_i = 1|\mathbf{x}) = \sigma(\mathbf{w}_i \mathbf{x} + b), \quad P(x_k = 1|\mathbf{h}) = \sigma(\mathbf{h}^T \mathbf{w}_k + c)$$

Contrastive divergence CD

- cel treningu: maksymalizacja prawdopodobieństwa $P(\mathbf{x})$

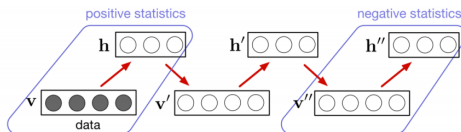
$$L = \frac{1}{N} \sum \log P(\mathbf{x})$$

- algorytm optymalizacji *Contrastive divergence* (Hinton)
 - dla każdego $\mathbf{x}$ wejściowego wygeneruj $\mathbf{x}'$ z k krotnego samplowania Gibbsa
 - aktualizacja

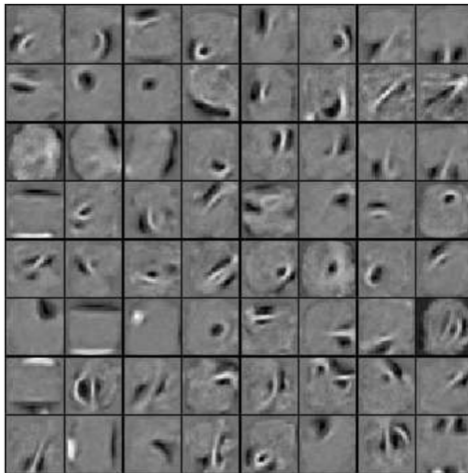
$$\Delta \mathbf{W} = \epsilon(\mathbf{x}\mathbf{h}^T - \mathbf{x}'\mathbf{h}'^T)$$

$$\Delta \mathbf{a} = \epsilon(\mathbf{x} - \mathbf{x}'), \quad \Delta \mathbf{b} = \epsilon(\mathbf{h} - \mathbf{h}')$$

- w praktyce $k = 1$ daje dobre rezultaty

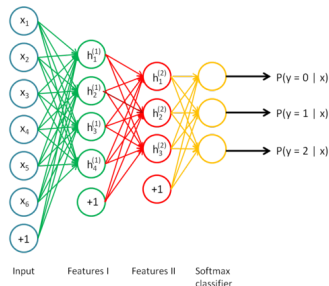


Filtry RBM na MNIST



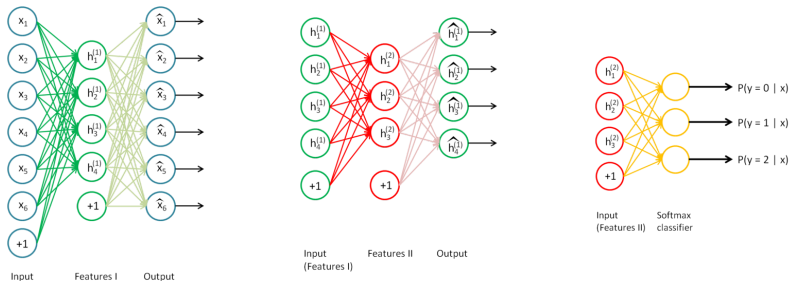
Erhan, Bengio, (2010) Why Does Unsupervised Pre-training Help Deep Learning?

Stacked autoencoders



- Sygnał na wyjściu warstwy ukrytej autoenkodera staje się wejściem kolejnego autoenkodera
- Hierarchia reprezentacji: pierwsza warstwa koduje najprostsze warstwy, kolejne bardziej ogólne relacje, itd.
- Podobna architektura z wielowarstwowym RBM to Deep Belief Networks (DBN)

Inicjalizacja sieci DNN



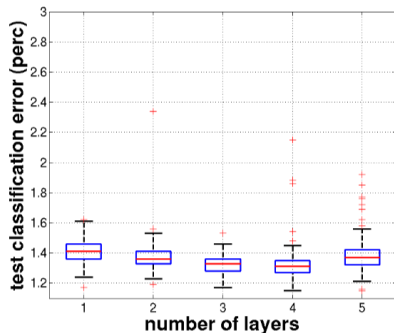
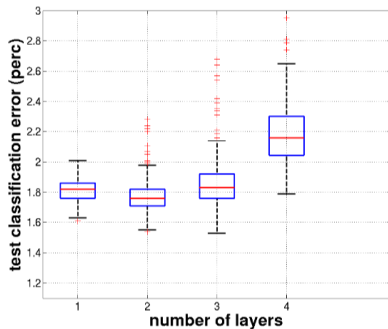
1. Greedy layer-wise pre-training:

- trening nienadzorowany autoenkodera wsteczną propagacją
- usuwamy warstwę wyjściową (dekoder) i uczymy kolejny AE, którego wejściem jest sygnał z warstwy ukrytej poprzedniego AE przy niezmiennych wartościach wag w poprzednich warstwach
- powtarzamy procedurę dla każdej kolejnej warstwy

2. fine tuning: dodajemy w pełni połączoną warstwę wyjściową (lub kilka warstw) i uczymy całą sieć w sposób nadzorowany

Pre-trening DNN

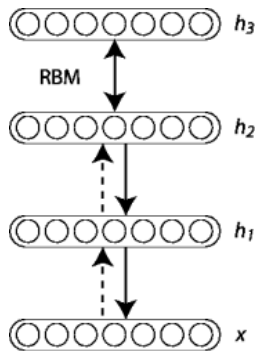
Pre-trening poprawia generalizację, działa jak regularyzacja



Erhan, Bengio, (2010) Why Does Unsupervised Pre-training Help Deep Learning?

Deep Belief Networks (DBN)

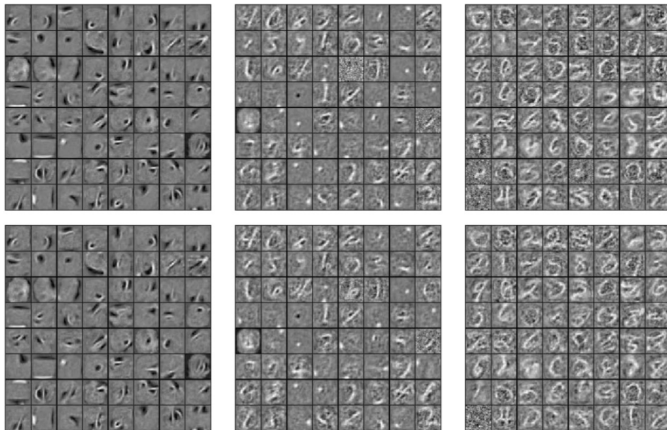
- DBN (Teh, Hinton 2006) wielowarstwowy RBM,
- uczy się głębokiej, hierarchicznej reprezentacji danych, każda kolejna warstwa analizuje sygnał widoczny w warstwie ukrytej poprzedniego RBM
- uczenie zachłanne: kolejne warstwy dodawane i uczone pojedynczo
- pierwszy efektywny model głęboki
- to nie jest sieć jednokierunkowa



src: <http://deeplearning.net/tutorial/DBN.html>

Filtry DBN na MNIST

pre-training

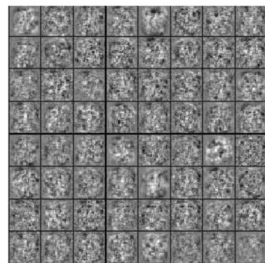
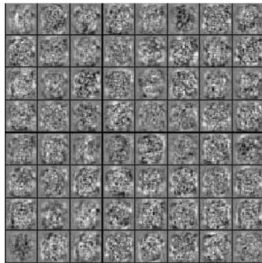
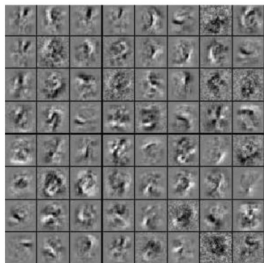


fine tune

Erhan, Bengio, (2010) Why Does Unsupervised Pre-training Help Deep Learning?

Filtry DBN na MNIST

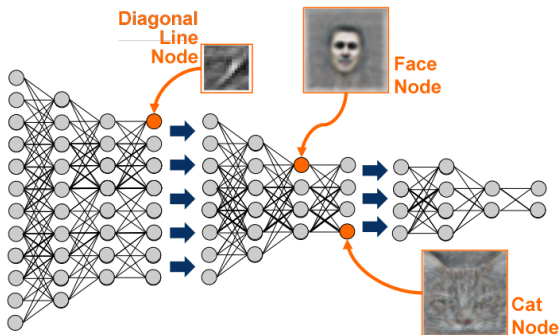
without pre-training



Erhan, Bengio, (2010) Why Does Unsupervised Pre-training Help Deep Learning?

Czym jest kot?

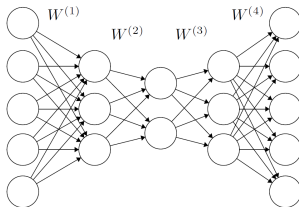
Sieć DBN trenowana na obrazach 200x200 z 10M losowych nagrań wideo z YouTube. Dane bez etykiet - nie było żadnego celu treningu. Trening rozproszony na 1000 maszynach (16K CPU) przez 3 dni.



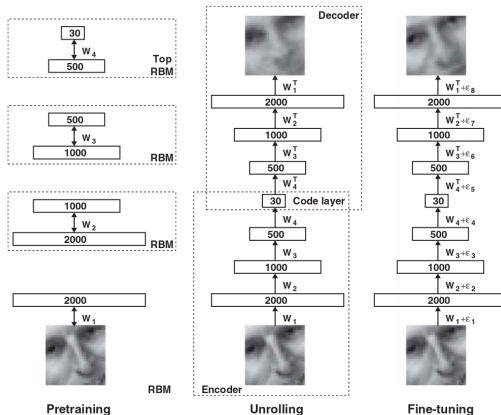
Pojedyncze neurony reagują na prezentację twarzy lub kota.

Głębokie autoenkodery

- dodatkowa warstwa ukryta zwiększa możliwości enkodera w odwzorowaniu kodu (uniwersalny aproksymator jest w stanie nauczyć się dowolnego kodowania)
- dodanie warstw pozwala zmniejszyć złożoność reprezentacji trudnych problemów
- głębokie AE pozwalają uzyskać większą kompresję (Hinton 2006)
- niekompletne głębokie AE - żadna warstwa ukryta nie powinna być mniejsza niż warstwa kodująca



Pretraining dla głębokich AE

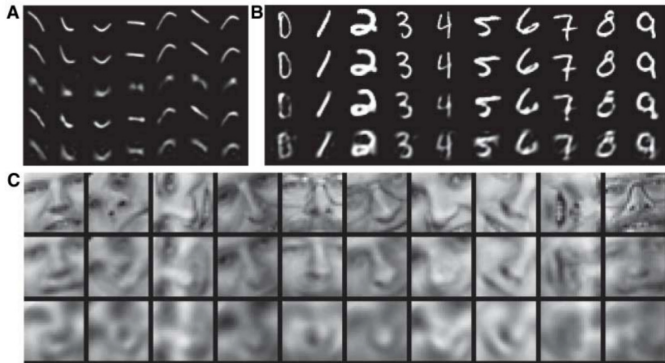


- RBM używany do inicjacji kodera i dekodera

G. Hinton, R. Salakhutdinov, *Reducing the Dimensionality of Data with Neural Networks*, Science 2006

Kompresja obrazów: Deep AE vs. PCA

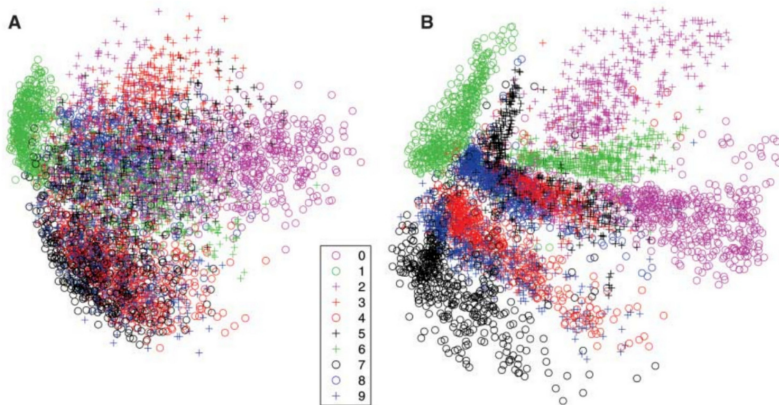
Fig. 2. (A) Top to bottom: Random samples of curves from the test data set; reconstructions produced by the six-dimensional deep autoencoder; reconstructions by "logistic PCA" (8) using six components; reconstructions by logistic PCA and standard PCA using 18 components. The average squared error per image for the last four rows is 1.44, 7.64, 2.45, 5.90. (B) Top to bottom: A random test image from each class; reconstructions by the 30-dimensional autoencoder; reconstructions by 30-dimensional logistic PCA and standard PCA. The average squared errors for the last three rows are 3.00, 8.01, and 13.87. (C) Top to bottom: Random samples from the test data set; reconstructions by the 30-dimensional autoencoder; reconstructions by 30-dimensional PCA. The average squared errors are 126 and 135.



G. Hinton, R. Salakhutdinov, Reducing the Dimensionality of Data with Neural Networks, Science 2006

Projekcja do 2D (dwa neurony kodujące)

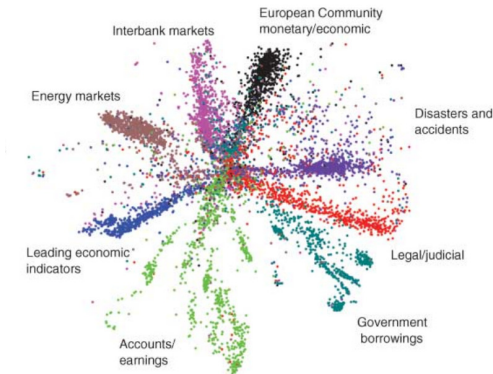
MNIST: PCA vs. Deep AE (784-1000-500-250-2)



G. Hinton, R. Salakhutdinov, Reducing the Dimensionality of Data with Neural Networks, Science 2006

Projekcja do 2D (dwa neurony kodujące)

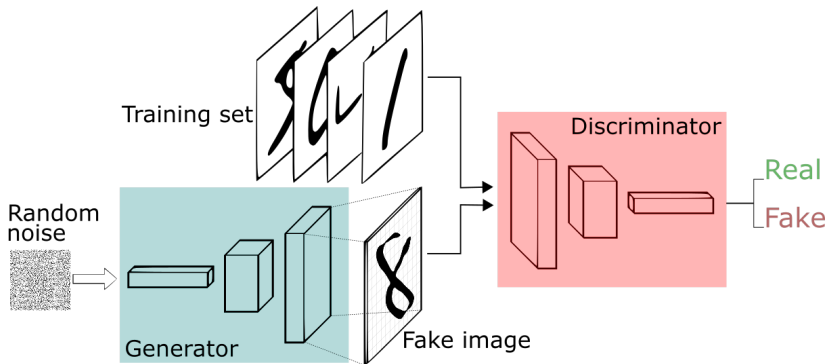
Wizualizacja dokumentów: architektura 2000-500-250-125-2



G. Hinton, R. Salakhutdinov, Reducing the Dimensionality of Data with Neural Networks, Science 2006

Generative Adversal Network (GAN)

Goodfellow 2014



src: <https://deeplearning4j.org/generative-adversarial-network>

Generative Adversal Network (GAN)

- GAN (Goodfellow 2014) - dwie sieci (generująca i oceniająca) rywalizujące ze sobą w grze o sumie zerowej
- model GAN uczy się generować dane o właściwościach zbliżonych do zbioru treningowego, np. zdjęć o realistycznych cechach
- sieć **oceniająca** (dyskryminująca, np. CNN) stara się odróżnić prawdziwy sygnał od wygenerowanego przez sieć generującą
- sieć **generująca** (np. sieć dekonwolucyjna) tworzy sygnał z pewnego rozkładu starając się „oszukać” sieć oceniającą, trening dąży do maksymalizacji błędu dyskryminacji
- obie sieci uczone wsteczną propagacją, sieć generująca tworzy coraz bardziej realistyczne obrazy, sieć oceniająca specjalizuje się w rozróżnianiu coraz subtelniejszych różnic pomiędzy obrazami prawdziwymi i wygenerowanymi

Trening GAN

Funkcja kosztu dyskryminatora - binarna entropia krzyżowa (spodziewana odpowiedź 1 dyskryminatora dla obrazu prawdziwego)

$$L_D = -\frac{1}{2}\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} \log D(\mathbf{x}) - \frac{1}{2}\mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} \log(1 - D(G(\mathbf{z})))$$

gdzie $p_{\text{data}}(\mathbf{x})$ rozkład danych prawdziwych, $p_z(\mathbf{z})$ rozkład danych generowanych

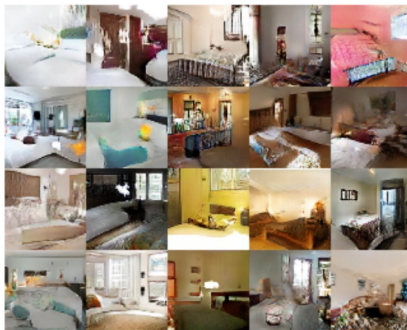
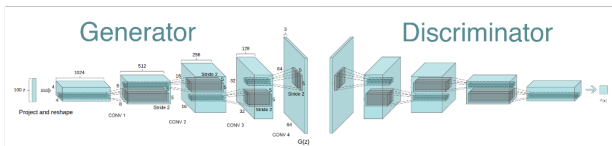
Funkcja kosztu generatora

$$L_G = -\frac{1}{2}\mathbb{E}_z \log D(G(\mathbf{z}))$$

maksymalizuje prawdopodobieństwo popełnienia błędu przez dyskryminator (spodziewana odpowiedź 1 dyskryminatora dla obrazu fałszywego)

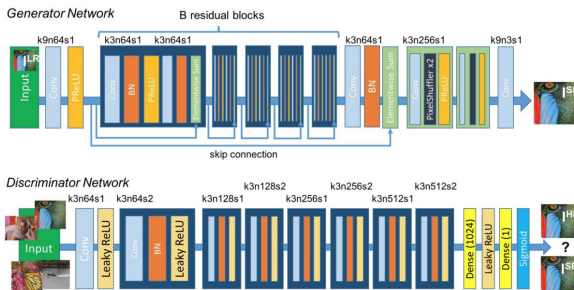
DCGAN - Deep Conv GAN

Generowanie obrazów za pomocą splotowej sieci GAN



Radford, A., et al. (2015) Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks

SRGAN - super resolution GAN

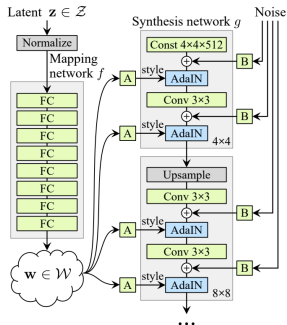


Rys:  Understanding and building Generative Adversarial Networks (GANs)

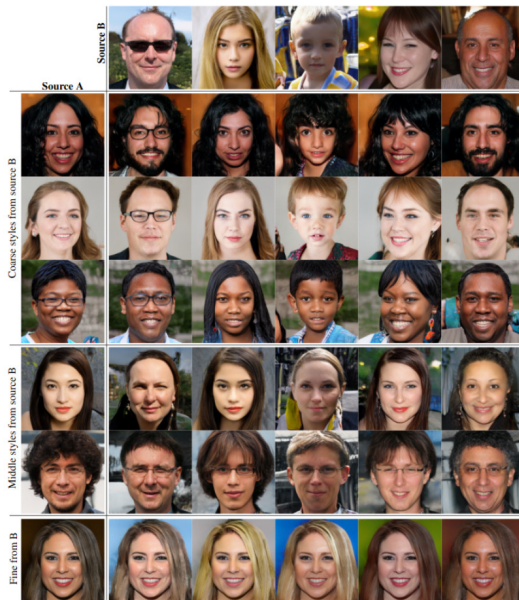
Figure 4: Architecture of Generator and Discriminator Network with corresponding kernel size (k), number of feature maps (n) and stride (s) indicated for each convolutional layer.

StyleGAN

Generator



Demo: 🖱️ Generator twarzy



T. Karras, S. Laine, T. Aila (2019) A Style-Based Generator Architecture for Generative Adversarial Networks

Inne zastosowania

- rekonstrukcja obrazów 3D ze zdjęć
- generowanie tekstu, synteza mowy, ...
- modyfikacje obrazów: zmiana twarzy, mimiki, fryzury, postarzanie osób na zdjęciach
- transfer stylu
- generowanie scen w grach wideo, zwiększanie rozdzielczości tekstur
- bezpieczeństwo: podnoszenie skuteczności algorytmów w wykrywaniu ataków cybernetycznych, wykrywanie fałszerstw, wykrywanie anomalii
- generowanie dane na potrzeby uczenia maszynowego
- generowanie obrazów z tekstu, generowanie nagrań wideo
- 📖 Curated list of awesome GAN applications and demo