#nn

Sieci neuronowe 2024

Wykład 1

Czym wyróżnia się perceptron prosty (Model Pittsa-Mc Cullocha)
?
Funkcja aktywacji zwraca dwie wartości - neuron jest pobudzony albo nie. Dodatkowo nie ma elementu nauki.

Wykład 2

Czym wyróżnia się ADALINE
?
Podobne do perceptronu prostego ale celem jest znalezienie wag. Może mieć wiele wejść, ma jedno wyjście

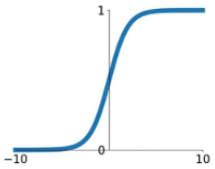
Jakiej funkcji kosztu użyjemy dla sieci rozwiązującej problem regresji?
?
Mean Squared Error

Jakiej funkcji kosztu użyjemy dla sieci rozwiązującej problem klasyfikacji?
?
Cross Entropy - $H(p,\hat{p}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]$.
W przypadku wieloklasowej klasyfikacji gdzie wyjścia są zakodowane jako 0 lub 1 stosuje się $\mathcal{L}_{CE} = -\sum_{k=1}^{M} y_k \ln(\hat{y}_k)$ (patrz wykład 3).

Jaka jest formuła sigmoidu
?

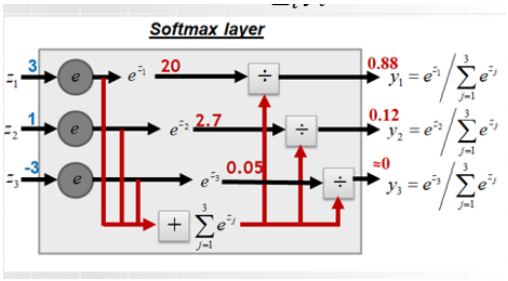
Sigmoid

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



Czym jest gradient funkcji wielu zmiennych
?
Wektor wskazujący kierunek największego wzrostu. Składa się z pochodnych cząstkowych ∂ funkcji straty po wagach ∂w_i .

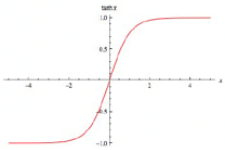
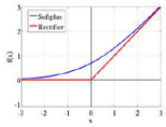
Czym jest softmax
?



$$P(Y = i|x, W, b) = softmax_i(Wx + b) = \frac{e^{W_i x + b_i}}{\sum_j e^{W_j x + b_j}}$$

Wykład 3

Jak wyglądają funkcje aktywacji i ich pochodne: **sigmoid, tangens hiperboliczny, relu, softplus**.

	$f(z) = \frac{1}{1+\exp(-z)}$	$f'(z) = f(z)(1-f(z))$
	$f(z) = \tanh(z)$	$f'(z) = (1-f^2(z))$
	$f(z) = \begin{cases} z, & z \geq 0 \\ 0, & z < 0 \end{cases}$	[*] $f'(z) = \begin{cases} 1, z \geq 0 \\ 0, z < 0 \end{cases}$
	$f(z) = \log(1+\exp(z))$	$f'(z) = \frac{1}{1+\exp(-z)}$

Czym jest N-Fold Cross-Validation i kiedy warto jej użyć
?

Idea #4: **Walidacja krzyżowa**: podział danych na foldy (części). Każdy **fold** jest używany do walidacji i do uśredniania wyników

fold 1	fold 2	fold 3	fold 4	fold 5	test
fold 1	fold 2	fold 3	fold 4	fold 5	test
fold 1	fold 2	fold 3	fold 4	fold 5	test

Dobre rozwiązanie dla małych zbiorów danych, ale nie dla dużych zbiorów dla modeli głębokich

Wykład 4

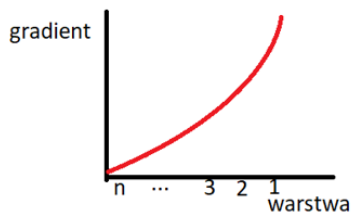
Jakie są strategie zmiany wag
?

- **Stochastic Gradient Descent**: uaktualnianie wag (parametrów) po każdym wzorcu. Cechuje się dużą wariancją funkcji kosztu :(
- **Minibatch Gradient Descent**: uaktualnianie wag po m wzorca (m rozmiar minibatcha). Redukuje wariancją (lepiej zbiega).
- **Batch Gradient Descent**: uaktualnianie wag po wszystkich wzorcach uczących (po całej epoce, of line training), co zwiększa stabilność treningu (bo liczymy gradient dopiero po przejsciu przez cały zbiór uczący)

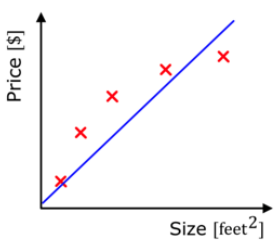
Jak za pomocą sigmoidu wyrazić tangens hiperboliczny
?
$\tanh(x) = 2sigm(2x) - 1$

Czym jest problem zanikającego gradientu
?
Bardzo duże lub małe pobudzenie neuronów, powoduje że aktywacja dla sigmoida lub tanh wynosi zawsze 0 lub 1. To prowadzi do niskich wartości gradientu, co przekłada się na wartość zmiany wag (a raczej jej brak). Poza funkcją aktywacji, odpowiedzialne za to zjawisko może być zła inicjalizacja wag -
nadanie im zbyt dużej wartości początkowej. Jeśli mamy do czynienia z zanikającym gradientem, warto zastosować RELU.

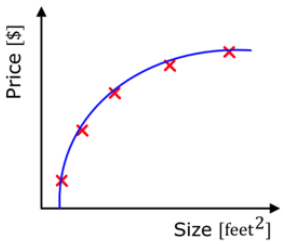
Czym jest problem wybuchającego gradientu
?
Korzystanie z RELU może powodować wysokie wartości gradientu, które w wyniku metody łańcuchowej mogą się zbytnio wyskalować. Aby temu zapobiec można zastosowoać obcinanie gradientu powyżej ustalonego progu albo leaky RELU



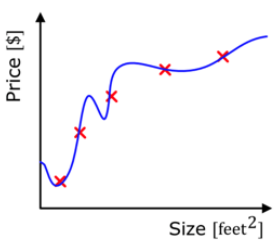
Bias a wariancja
?



Duży bias



ok



duża wariancja

niedouczenie

przeuczenie

Wykład 5

Jak sobie radzić z przetrenowaniem modelu?
?

1. Regularyzacja wag (L1 lub L2)
2. Dropout
3. Wczesne zatrzymanie uczenia
4. Sztuczne powiększane zbioru uczącego

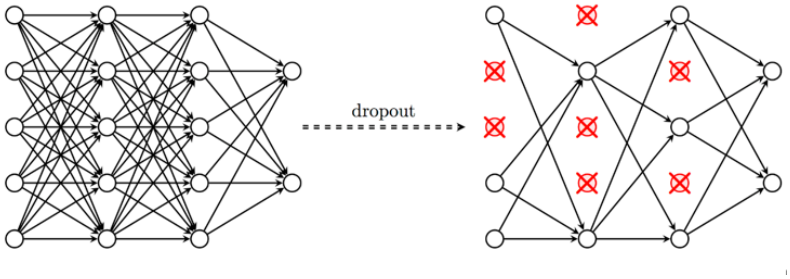
Na czym polega L2 (Ridge Regression)
?

1. **wzór**: $L(\theta) = L_0(\theta) + \lambda\frac{1}{2}\|\theta\|^2$, gdzie $\|\theta\|_2 = (w_1)^2 + (w_2)^2 + \ldots$, a λ to hiperparametr.
2. **opis**: Na zwiększeniu funkcji straty o kwadraty wag. Im bardziej skomplikowany model, tym większa kara.
3. **wpływ na aktualizację wag:** zmniejsza wartości proporcjonalnie do ich wartości, ale nie ustawia ich na 0

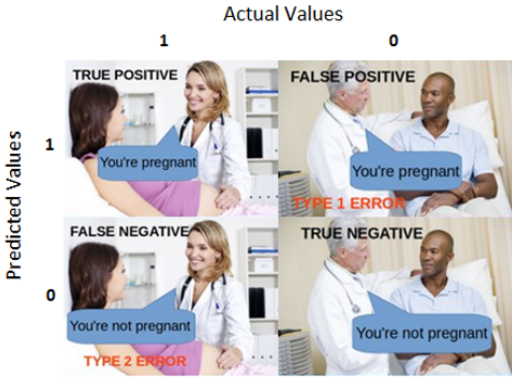
Na czym polega L1 (Lasso Regression)
?

1. **wzór:** $L(\theta) = L_0(\theta) + \lambda\|\theta\|$, gdzie $\|\theta\| = |w_1| + |w_2| + \ldots$, a λ to hiperparametr.
2. **opis:** Na zwiększeniu funkcji straty o wartości bezwzględne wag. Im bardziej skomplikowany model, tym większa kara.
3. **wpływ na aktualizację wag:** zmniejsza wagi niezależnie od ich wartości, w szczególności może je zerować - tym samym wyłączać neurony

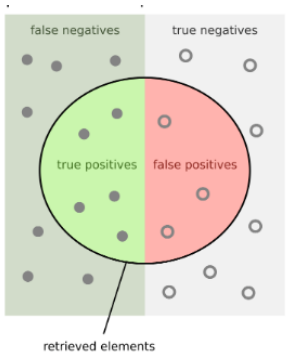
Na czym polega Dropout
?
Z określonym prawdopodobieństwem na okres batcha wyłączamy losowe neurony w warstwie ukrytej. Przy każdej paczce wyłączamy inną część sieci. Możemy postrzegać tak wytrenowany model jako rodzina klasyfikatorów (wiele sieci złączone w jedną). Dropout wykorzystywany jest tylko przy treningu

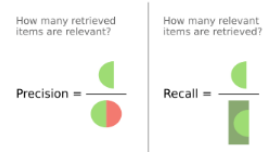


Czym jest macierz pomyłek (confusion matrix)
?



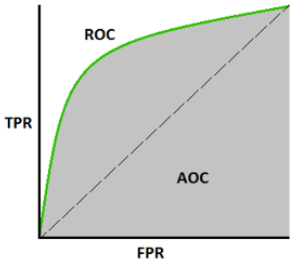
Metryki
?





Czym jest F1 score
?
Średnia harmoniczna precyzji i recall

Czym jest krzywa AUC-ROC (Area Under Curve - Receiver Operating Characteristics)
?
Wykres True-Positive-Rave (**Specifity**) od False-Positive-Rate (Sensitivity). Pożądane jest aby pole pod wykresem było 1.

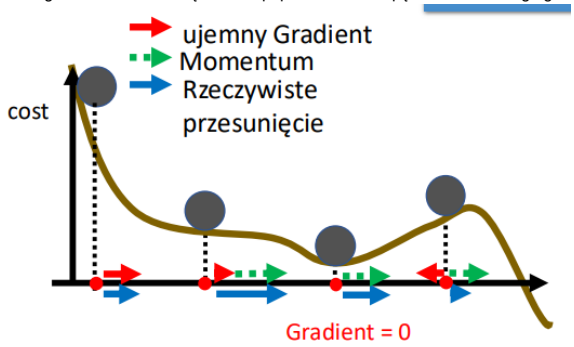


Wymień dobre praktyki w projektowaniu sieci MLP
?

- sprawdzenie czy zbiór jest zrównoważony
- kodowanie wartości nominalnych (kategorycznych)
- skalowanie i normalizacja cech o wartościach rzeczywistych
- liczba neuronów w warstwie ukrytej to $n_h = \frac{n_{in}+n_{out}}{2}$ lub $n_h = \sqrt{n_{in} n_{out}}$
- Po trenowaniu zwizualizuj gradienty w każdej warstwie sieci , jeśli większość z nich jest bliska 0 =>większość neuronów przestała się uczyć albo nie działa propagacja wsteczna.

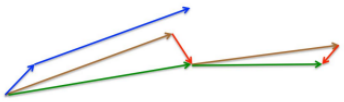
Wykład 6

Na czym polega **Momentum** - metoda optymalizująca Gradient Descent
?
Pomaga radzić sobie z "wąwozami" poprzez nadanie "pędu". Do obecnego gradientu dodaje momentum (zakumulowany gradient z poprzednich kroków).



Na czym polega Nesterov Accelerated Gradient (NAG)
?
Bazuje na momentum, jednak najpierw robi krok wynikający z momentum, dopiero potem liczy gradient i robi kolejny krok.

- NAG najpierw skacze w kierunku poprzedniego zakumulowanego gradientu (wektor brązowy)
- Mierzy gradient w tym miejscu i wykonuje korekcję (czerwony wektor)
- W wyniku otrzymujemy korekcję zgodną z NAG (zielony wektor)
- Zapobiegamy zbyt szybkiemu uaktualnianiu wag
- Wektor w kolorze niebieskim odpowiada klasycznemu momentum



Czym jest ADAGRAD
?
Jest to adaptacyjny współczynnik uczenia, który w z czasem maleje, przez kumulowanie gradientów w mianowniku, a to może powodować, że współczynnik uczenia będzie nieskończenie mały

$$\eta_w = \frac{\eta}{\sqrt{\sum_{i=0}^{t}(g^i)^2}}$$

Czym jest ADADELTA
?
Taki ulepszony ADAGRAD, tylko, że do wyliczenia współczynnika uczenia, bierzemy pod uwagę kilka ostatnich gradientów.

Czym jest Adam (Adaptive Moment Estimation)
?
Najczęściej wykorzystywany optymalizator. Estymuje średnią i wariancję gradientów i wykorzystuje je do aktualizacji wag. Bywa łączony z **NAG** (NAdam).

Co powoduje inicjalizacja wag na jednakową wartość (np. 0)
?
Prowadzi do jednakowego wpływu każdego neuronu na funkcję kosztu, a to powoduje, że każdy neuron uczy się tej samej cechy

Jak zainicjalizować wagi (najprostsza metoda)
?

1. $w \in \left[-\frac{a}{\sqrt{n_{\text{in}}}}, \frac{a}{\sqrt{n_{\text{in}}}}\right]$, gdzie n_{in} to liczba wejść i a zależy od funkcji aktywacji,

- dla sigmoida $a = 2.38$,
- dla ReLU $a = \sqrt{2}$,
- dla Tanh $a = 1$.

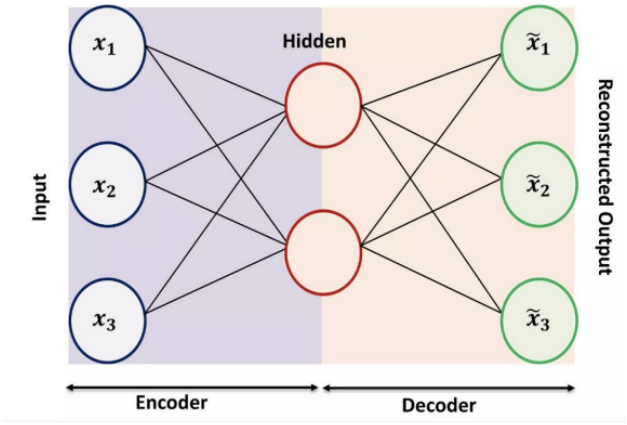
2. Można też brać pod uwagę ilość neuronów ukrytych $\left[-n_{\text{in}}\sqrt{N_h}, n_{\text{in}}\sqrt{N_h}\right]$.
3. Neurony wyjściowe losuje się z przedziału $[-0.5, 0.5]$.

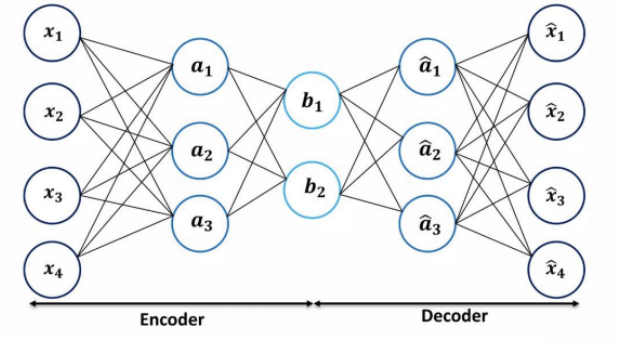
Jak zainicjalizować wagi korzystając z Xaviera
?
Dla każdej warstwy z osobna wylosuj z przedziału $\pm\frac{\sqrt{6}}{\sqrt{n_i+n_{i+1}}}$, gdzie n_i to wejściowe a n_{i+1} wyjściowe. Biasy ustaw na 0.

Jak zainicjalizować wagi korzystając z He
?
TODO

Czym jest normalizacja batch'a
?
TODO

Czym jest Autokoder i jak się go trenuje
?
Jest to sieć wykorzystująca uczenie nienadzorowane. Składa się z enkodera i dekodera. Celem jest skompresowanie reprezentacji i odtworzenie jej.

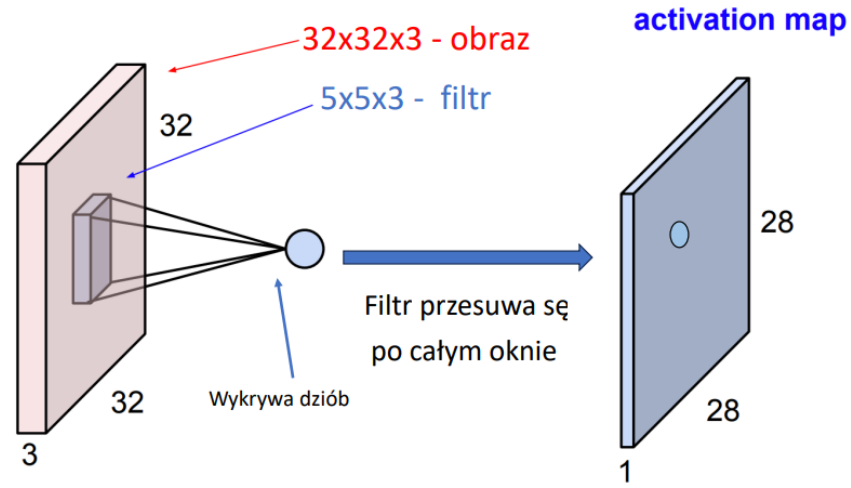




Wykład 7

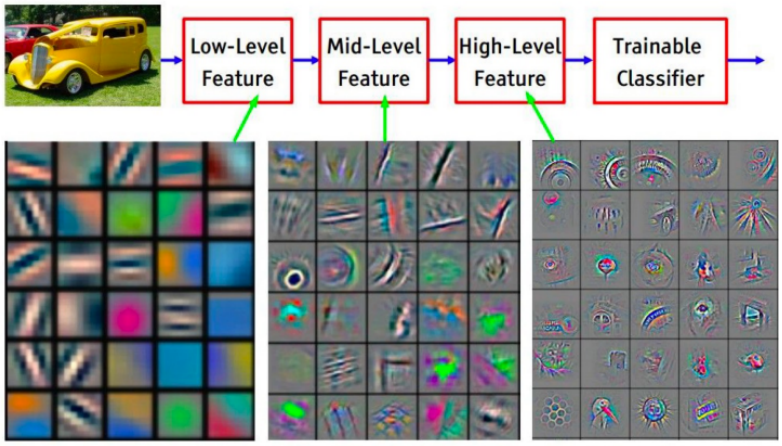
Czemu MLP nie radzą sobie za dobrze z obrazami
?
Reprezentacja obrazu w postaci wektora pozbawia sieć reprezentacji przestrzennej i zależności czasowej. Dodatkowo wykorzystanie pełnych połączeń jest wymagające obliczeniowo.

Czym jest kernel (CNN)
?
Inaczej detektor lub filtr - wykrywacz cech, który w wyniku konwolucji - przesuwania się po obrazie z określonym krokiem, tworzy mape aktywacji. Filtr trzyma w sobie wagi.



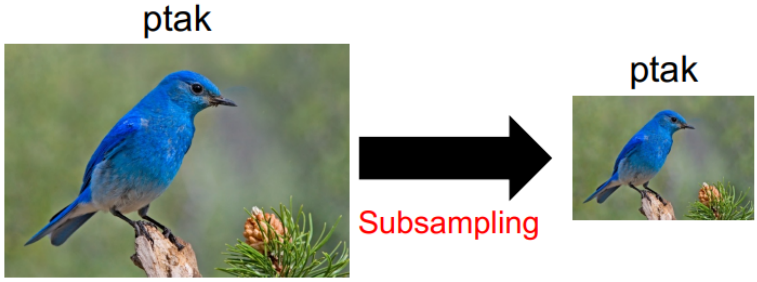
Czym jest konwolucja
?

Jest to korelacja wzajemna okolicznych pikseli i filtra odbróconego o 180 stopni. Na zdjęciu wizualizacja kolejnych map aktywacji - niskopoziomowe cechy składają się w rozpoznawalne obiekty.



Czym jest i do czego wykorzystuje się Pooling
?
Pozwala na agregację (avg lub max) sąsiadujących pikseli. Redukuje liczbę parametrów modelu, pozwala na lepszą generalizację map cech. Pooling nie podlega uczeniu - nie ma wag.
Propagacja wsteczna:

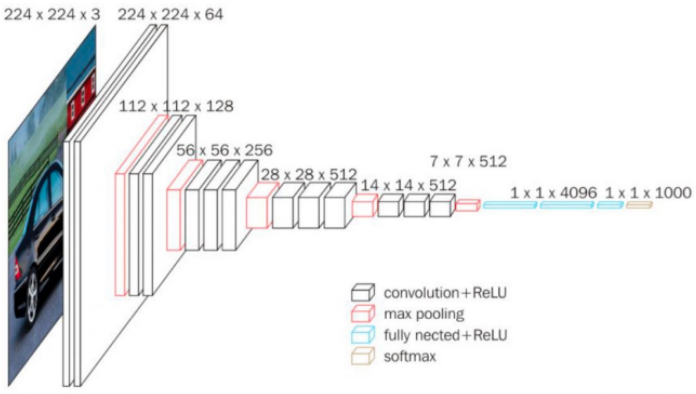
- jeśli używamy max-pool obliczany jest błąd uzyskiwany przez zwycięską jednostkę (pozostałe błędy są równe 0, dlatego trzeba pamiętać indeks zwycięskiej jednostki). •
- Dla avg-pool błąd jest przypisywany do całego bloku i mnożony przez 1/(NxN)



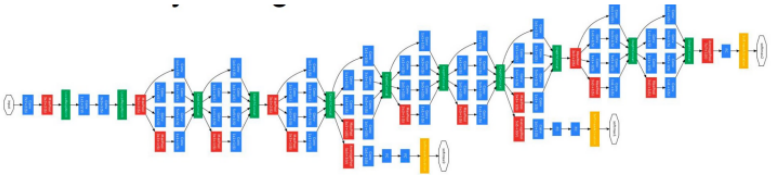
Jaki sens w stosowaniu konwolucji dla filtra o rozmiarze 1x1
?
TODO

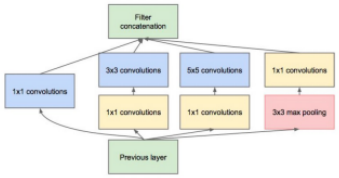
Jak przebiega formuła propagacji błędu w sieciach konwolucyjnych
?
XD, nie będzie takiego pytania

Czym jest VGG
?
Giga dużo warstw przeplatanych poolingiem. Na końcu 3 warstwy tradycyjnych sieci, a po nich softmax do klasyfikacji co jest na obrazku.



Co to GoogleNet
?
Jeszcze więcej warstw. Cała sieć składa się z nastackowanych bloków incepcji, które są jeszcze bardziej zagmatwane i wykorzystują konwolucje 1x1

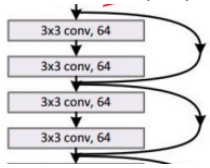




Czym jest ResNet

?

Podobne do VGG ale wykorzystuje połączenia rezydualne (skrótowe) wyjście z wcześniejszych warstwach jest przekazywane na wyjście następnych warstw.



Gdzie stosuje się konwolucję 1D

?

Do przetwarzania danych czasowych (np cen akcji) lub sygnału.