Sprawozdanie

Ćwiczenie 3
Sieć wielowarstwowa uczona metodą SGD
Techniki poprawy szybkości uczenia

Piotr Swirkaitis

246753

Badania metod optymalizacji współczynnika uczenia

Wprowadzenie

W ramach ćwiczenia zaimplementowano optymalizatory wspomagające proces uczenia wielowarstwowej sieci neuronowej. Zaimplementowane metody optymalizacji to

- Momentum
- Momentum Nesterova
- Adagrad
- Adadelta
- Adam

Dla zaimplementowanych metod przeprowadzono badania, mające na celu zbadanie skuteczności metod optymalizacji oraz ich wpływu na szybkość procesu uczenia sieci neuronowej oraz jej skuteczność. Badania zostały przeprowadzone na zbiorze MNIST, z którego korzystano w ćwiczeniu drugim oraz z wykorzystaniem parametrów ustalonych podczas badań w ramach ćwiczenia drugiego.

Badania

Jako podstawowe parametry wprowadzane do sieci podczas badań przyjęto :

- Rozmiar paczki (batcha): 32 obrazki
- Współczynnik uczenia: 0.5
- Wagi: rozkład normalny z odchyleniem standardowym 0.5
- Liczba warstw ukrytych: 2
- Liczba neuronów w kolejnych warstwach ukrytych:
    - 128
    - 64

Momentum

W badaniach zastosowano współczynnik momentum o wartości: 0.9

Funkcja aktywacji	Wartość trafności predykcji w procentach		
	1 epoka	5 epoka	10 epoka
ReLU	85.892	87.15	92.18
Sigmoid	86.703	88.25	92.34
Wyniki Lab 2	86.54	90.254	92.155

Wnioski

Z przeprowadzonych badań wynika, że metoda optymalizacji Momentum nieznacznie poprawia wyniki predykcji wyuczonej sieci neuronowej. Metoda optymalizacji Momentum bazuje na modyfikacji sposobu aktualizacji wag w każdej warstwie. Polega ona na uwzględnieniu podczas aktualizacji, również części poprzedniej aktualizacji wag przemnożonej przez ustalany współczynnik momentum. Z wyników można zaobserwować, że nieznacznie wyższe wyniki osiągnęła sieć uczona z pomocą funkcji aktywacji sigmoidalnej, jednak po 10 epokach różnica była marginalna. Porównując otrzymane wyniki z najlepszymi wynikami z listy 2 można zaobserwować, że były one zbliżone.

Momentum Nesterova

W badaniach zastosowano współczynnik momentum o wartości: 0.9

Funkcja aktywacji	Wartość trafności predykcji w procentach		
	1 epoka	5 epoka	10 epoka
ReLU	81.123	89.74	92.704
Sigmoid	87.23	92.57	93.845
Wyniki Lab 2	86.54	90.254	92.155

Wnioski

Podstawową różnicą między momentum Nesterova, a zwykłym momentum jest różnica dotycząca aktualizacji wag. W metodzie Momentum bierze się pod uwagę wartość poprzedniej aktualizacji wag przemnożonej przez współczynnik momentum, natomiast w momentum Nesterova aktualna aktualizacji wag jest zależna od wartości aproksymacji następnej pozycji parametrów. W momentum Nesterova należy obliczyć gradient dla przewidywanych na przyszłość wartości parametrów, a następnie przemnożyć przez współczynnik momentum i uwzględnić w aktualnej aktualizacji wag. Z przeprowadzonych badań można wywnioskować, że badana metoda optymalizacji poprawia działanie procesu uczenia sieci neuronowej. Po 10 epokach, można zauważyć, że wartości predykcji oscylują w granicach ponad 92%. W wypadku uczenia za pomocą funkcji aktywacji ReLU sieć po 10 epokach osiągnęła 92% trafności predykcji, natomiast w wypadku sieci uczonej za pomocą sigmoidalnej funkcji aktywacji było to 93%. W porównaniu do wyników z laboratorium drugiego można zauważyć nieznaczną poprawę.

Metoda Adagrad

W badaniach zastosowano współczynnik gamma o wartości: 0.9

Funkcja aktywacji	Wartość trafności predykcji w procentach		
	1 epoka	5 epoka	10 epoka
ReLU	88.658	89.47	92.95
Sigmoid	88.01	89.776	93.254
Wyniki Lab 2	86.54	90.254	92.155

Wnioski

Metoda optymalizacji Adagrad pozwala sieci na modyfikację współczynnika uczenia, co eliminuje manualne ustawianie współczynnika uczenia. Za pomocą tego algorytmu współczynnik uczenia osiąga wysokie wartości dla parametrów skojarzonych z rzadkimi cechami, natomiast dla parametrów skojarzonych z częściej pojawiającymi się cechami jest on niższy. Z przeprowadzonych badań można wywnioskować, że metoda optymalizacji pozwoliła na wyuczenie modelu aż do 88% procent predykcji już po pierwszej epoce. Zarówno przy zastosowaniu funkcji sigmoidalnej, czy też funkcji ReLU, można było zauważyć, że proces uczenia sieci przeszedł sprawnie i model w 10 epoce osiągał trafność predykcji równą ponad 90%. W porównaniu z badaniami przeprowadzonymi w ramach laboratorium drugiego można zauważyć niewielką poprawę.

Metoda Adadelta

W badaniach zastosowano współczynnik gamma o wartości: 0.9

Funkcja aktywacji	Wartość trafności predykcji w procentach		
	1 epoka	5 epoka	10 epoka
ReLU	88.678	92.67	94.116
Sigmoid	90.828	93.636	94.025
Wyniki Lab 2	86.54	90.254	92.155

Wnioski

Metoda Adadelta jest rozszerzeniem metody Adagrad, jednakże w odróżnieniu od metody Adagrad suma gradientów jest określana jako zmniejszająca się rekursywnie średnia poprzednich gradientów. Z przeprowadzonych badań można zanotować obserwację, że zastosowanie tego optymalizatora znacząco poprawiło proces uczenia sieci neuronowej. W porównaniu z wynikami otrzymanymi na laboratorium nr. 2, w tym badaniu można zaobserwować, że trafność predykcji zwiększyła się o 2 punkty procentowe po 10 epokach, warte zanotowanie jest również, że siec osiągnęła trafność predykcji na poziomie 92% już po 5 epokach, zarówno dla funkcji aktywacji ReLU oraz funkcji sigmoidalnej.

Metoda Adam

W badaniach zastosowano współczynniki: beta1 o wartości: 0.9 oraz beta2 o wartości 0.99.

Funkcja aktywacji	Wartość trafności predykcji w procentach		
	1 epoka	5 epoka	10 epoka
ReLU	90.158	92.68	95.245
Sigmoid	91.69	93.71	95.047
Wyniki Lab 2	86.54	90.254	92.155

Wnioski

Metoda Adam jest często wykorzystywana w wielu środowiskach, jest ona jednak bardziej kosztowna pamięciowo dla zadanego rozmiaru paczki niż inne zbadane w ramach badań optymalizatory. Zastosowanie optymalizatora Adam pozwoliło na osiągnięcie najwyższych wyników pośród innych badanych optymalizatorów. Już po pierwszej epoce, sieć osiągała trafność predykcji na poziomie 90%, natomiast po 10 epokach wartość predykcji osiągała już 95%, zarówno dla funkcji ReLU jak i funkcji sigmoidalnej, z niewielką przewaga funkcji ReLU. W porównaniu z wynikami otrzymanymi po przeprowadzeniu badań z listy 2 można zauważyć wzrost trafności predykcji o 3%.

Badanie metod inicjalizacji wag

Wprowadzenie

W ramach ćwiczenia wdrożono dwie nowe metody inicjalizacji wag, mające optymalizować proces uczenia wielowarstwowej sieci neuronowej. Inicjalizacja wag jest znaczącym czynnikiem, wpływającym na proces uczenia sieci neuronowej. Zaimplementowano metodę inicjalizacji Xavier oraz metodę inicjalizacji He. Określają one odchylenie standardowe, przyjmowane podczas generowania wag na podstawie rozkładu normalnego. Do badań zastosowano najlepszy optymalizator, wybrany na podstawie przeprowadzonych wcześniej badań.

Badania

Jako podstawowe parametry wprowadzane do sieci podczas badan przyjęto :

- Rozmiar paczki (batcha): 32 obrazki
- Współczynnik uczenia: 0.5
- Liczba warstw ukrytych: 2
- Najlepszy optymalizator wybrany na podstawie badań metod optymalizacji: Adam
- Liczba neuronów w kolejnych warstwach ukrytych:
    - 128
    - 64

Metoda inicjalizacji Xaviera

Funkcja aktywacji	Wartość trafności predykcji w procentach		
	1 epoka	5 epoka	10 epoka
ReLU	87.458	93.04	94.24
Sigmoid	88.42	94.23	95.17
Wyniki Lab 2	86.54	90.254	92.155

Wnioski

Z przeprowadzonych badań można zaobserwować, że po zastosowaniu inicjalizacji wag metodą Xaviera osiągnięto wyższe wyniki trafności predykcji dla sieci neuronowej uczonej z wykorzystaniem sigmoidalnej funkcji aktywacji. Po 10 epokach model wykorzystujący sigmoidalną funkcję aktywacji osiągnął trafność predykcji na poziomie 95.17% w porównaniu do sieci neuronowej, wykorzystującej funkcję aktywacji ReLU, która osiągnęła trafność na poziomie 94.24%. Porównując otrzymane wyniki z wynikami otrzymanymi podczas realizacji laboratorium nr.2, gdzie wagi były generowane z manualnie ustawionym odchyleniem standardowym można zauważyć wzrost trafności predykcji.

Metoda inicjalizacji He

Funkcja aktywacji	Wartość trafności predykcji w procentach		
	1 epoka	5 epoka	10 epoka
ReLU	92.321	94.27	95.93
Sigmoid	91.27	93.71	94.215
Wyniki Lab 2	86.54	90.254	92.155

Wnioski

Analizując wyniki przeprowadzonych badań można było zauważyć, że lepsze wyniki otrzymano dla modelu sieci neuronowej uczonej z wykorzystaniem funkcji aktywacji ReLU. Po 10 epokach otrzymano wyniki świadczące o trafności predykcji na poziomie około 96%. Porównując to z wynikami otrzymanymi dla sieci neuronowej uczonej z wykorzystaniem sigmoidalnej funkcji aktywacji, wyniki oscylowały około wartości 94%. Niezależnie od wybranej funkcji aktywacji można było zauważyć poprawę trafności predykcji w porównaniu do wyników otrzymanych podczas laboratorium 2.

Wnioski końcowe

Analizując przeprowadzone badania można wywnioskować, że zastosowanie odpowiedniego optymalizatora może wpływać na proces uczenia sieci neuronowej. Po zbadaniu 5 metod optymalizacji stwierdzono, że najlepsze wyniki osiągnięto dla optymalizatora Adam. Jest on najbardziej kosztowny pamięciowo wśród badanych optymalizatorów, jednak poprawę wyników względem innych optymalizatorów można zauważyć już po niewielkiej liczbie epok. Po zbadaniu najlepszych metod optymalizacji, rozpoczęto badania nad metodami inicjalizacji wag w modelu sieci neuronowej. Do badań wykorzystano optymalizator Adam. Przebadano dwie metody inicjalizacji wag: He oraz Xavier. Z przeprowadzonych badań wynika, że zastosowanie inicjalizacji He przynosi większe korzyści dla sieci uczonej z wykorzystaniem funkcji aktywacji ReLU, natomiast inicjalizacja Xavier zapewnia lepsze wyniki dla sieci uczonych dla z wykorzystaniem sigmoidalnej funkcji aktywacji.

Porównując otrzymane wyniki z wynikami otrzymanymi podczas wykonywania laboratorium 2, można jednoznacznie stwierdzić, że zarówno użycie odpowiednich optymalizatorów, jak i właściwe dobranie sposobu inicjalizacji wag, ma znaczący wpływ na szybkość oraz skuteczność uczenia wielowarstwowych sieci neuronowych.