Multi Layer Perceptron - Raport

Metody Głębokiego Uczenia Mateusz Dorobek i Michał Turski

Użyte biblioteki

- pickle eksportowanie i importowanie modelu z pliku binarnego
- numpy obliczenia macierzowe
- matplotlib rysowanie wykresów i wizualizacja danych
- pandas wczytywanie zbioru uczącego z pliku
- sklearn.model_selection funkcja dzieląca na zbiór uczący i treningowy
- imgaug biblioteka do augmentacji zbioru uczącego

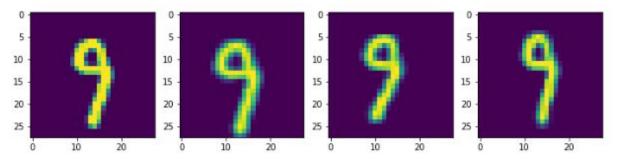
Implementacja

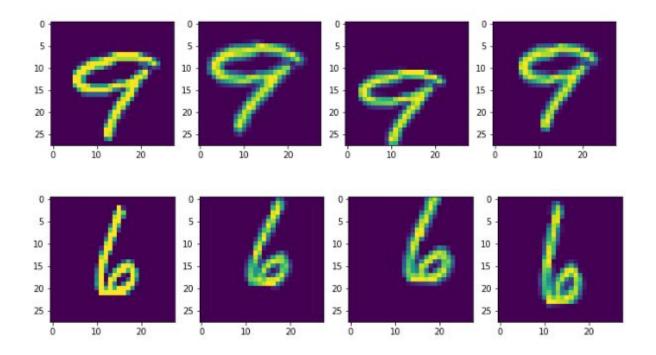
Rozszerzanie zbioru uczącego

Zbiór testowy rozszerzany był przy pomocy losowej kombinacji przekształceń afinicznych. W ich skład wchodziło:

- Rozciągnięcie w osiach x i y w zakresie 85% 115% szerokości obrazka.
- Przesunięcie w kierunkach osi x i y w zakresie ±10% szerokości obrazka.
- Obrotu w zakresie ±8 stopni.
- Pochylenia w zakresie ±3 stopni.

W efekcie transformacji otrzymano z jednego obrazka wejściowego trzy nowe. Przykładowy efekt tej operacji został przedstawiony na ilustracjach poniżej:





Pętla ucząca

- Użyliśmy uczenia z wykorzystaniem pakietów (batchy) ponieważ powodowało to mniejszą oscylację funkcji kosztu i stabilniejsze jej zbieganie.
- Do reprezentacji klasy do której należy dany przypadek treningowy użyliśmy reprezentacji one-hot (wektor zer, który ma jedną jedynkę w miejscu odpowiadającym danej klasie)
- warunek przerwania
- Warunek stopu jeżeli 3 razy po rząd funkcja kosztu dla zbioru testowego zwiększy swoją wartość to należy przerwać proces uczenia aby uniknąć przetrenowania.

Wprowadzone rozszerzenia

W ramach pracy nad projektem, do sieci wprowadzono następujące rozszerzenia:

- **Algorytm optymalizujący Adam**. Nie został on wykorzystany, gdyż zamiast poprawiać wyniki pogarszał je.
- Inicjalizacja wag Xaviera. I w tym przypadku otrzymane rezultaty były gorsze, dlatego zrezygnowano z tej modyfikacji.
- Numeryczne obliczanie gradientu. Wprowadzono ją po to, aby sprawdzić poprawność implementacji wstecznej propagacji gradientu. Wyniki były zgodne co do trzech cyfr znaczących.

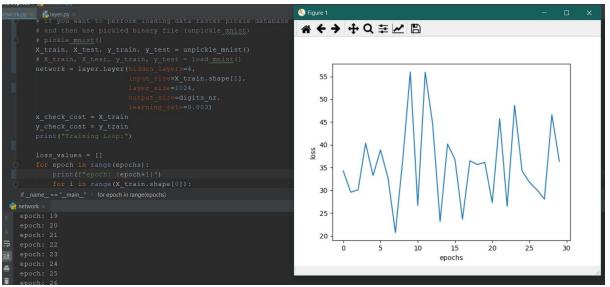
Użyte hiperparametry

- W finalnej wersji użyliśmy 25 epok, ale pętla została przerwana wcześniej z powodu spełnienia warunku stopu (szczegóły niżej).
- Rozmiar paczki uczącej (batch) to 16 elementów.

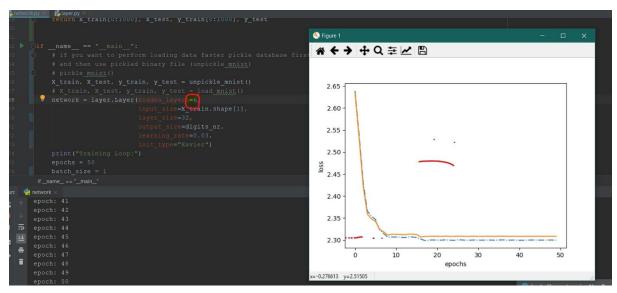
- Learning rate = 0.005.
- Rozmiar warstwy ukrytej 128.
- Ilość warstw 2 (jedna ukryta i jedna wyjściowa).
- Funkcja aktywacji w warstwie wejœściwej i ukrytej tangens hiperboliczny.
- Funkcja aktywacji w warstwie wyjściowej softmax.
- Funkcja kosztu entropia krzyżowa.
- Powiększenie zbioru uczącego 4 razy z wykorzystaniem przekształceń afinicznych.
- Inicjowanie wag poprzez podzielenie wartości z przedziału (0, 1) przez rozmiar wejściowy warstwy.
- Dla tak małych warstw użycie funkcji optymalizującej Adam przynosiło efekt odwrotny niż oczekiwany.

Przebieg doboru parametrów i uzyskane wyniki

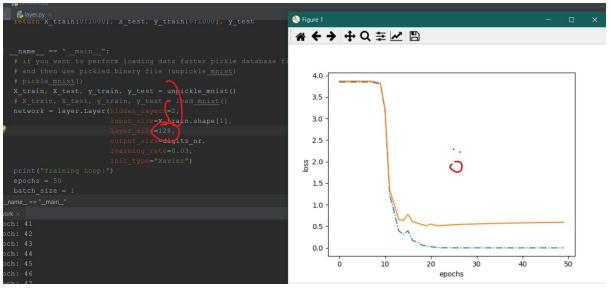
Zbyt duży rozmiar warstwy (1024 neurony) i zbyt duża liczba warstw (4) powodowały oscylacje w związku ze zbyt duża strukturą sieci w porównaniu do wielkości zbioru uczącego.



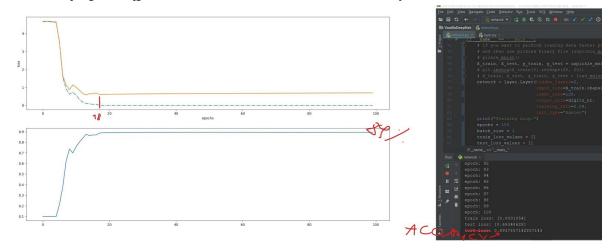
Dalsze zwiększanie głębokości sieci (do 6 warstw) doprowadziło do utknięcia w lokalnym minimum:



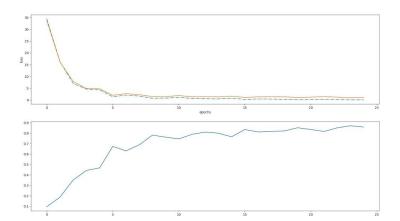
Użycie odpowiedniej liczby warstw (dwóch warstw po 128 neuronów) powoduje znaczną poprawę w zbieganiu funkcji kosztu:

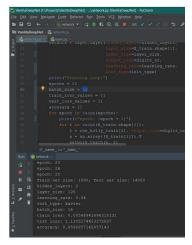


Ustawienie odpowiedniego parametrów learning rate na 0.04 pozwala nawet dla niepełnego zbioru uczącego osiągać 89% skuteczności na zbiorze testowym.

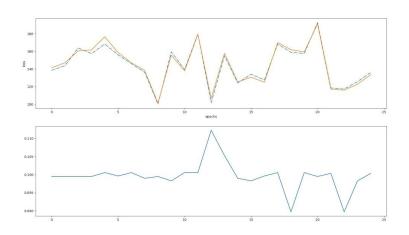


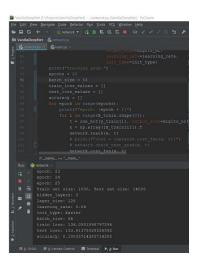
Zwiększenie rozmiaru batcha do 16 pogorszyło nieco skuteczność algorytmu i zwiększyło oscylację funkcji kosztu.



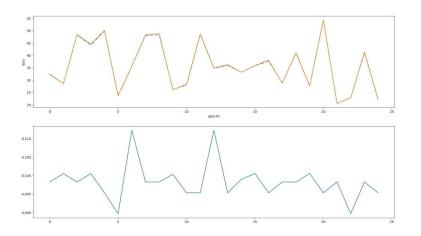


Dalsze zwiększanie rozmiaru batcha (do 64) jedynie potwierdza tę hipotezę...



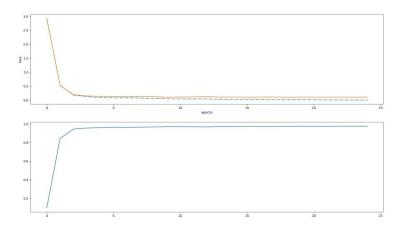


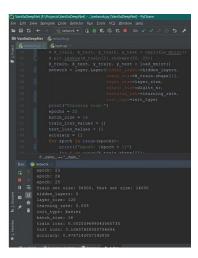
... również dla użycia pełnego zbioru uczącego.





Zwiększenie rozmiaru batcha, jak się okazało, powinno iść ze zmniejszeniem parametru learning rate, co po kilku próbach dla pełnego zbioru okazało się bardzo dobrą metodą:

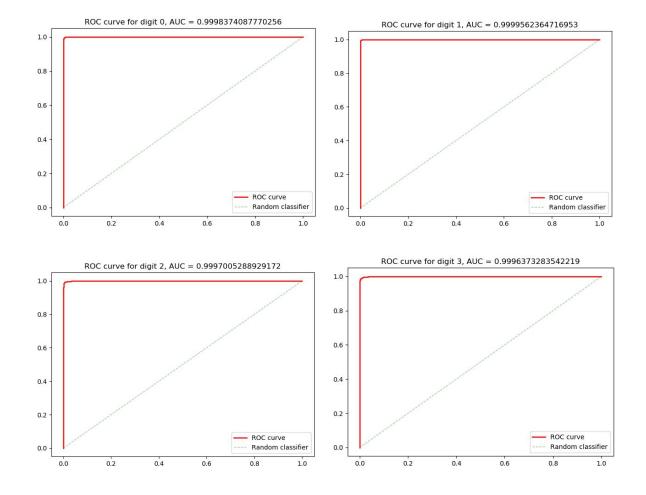


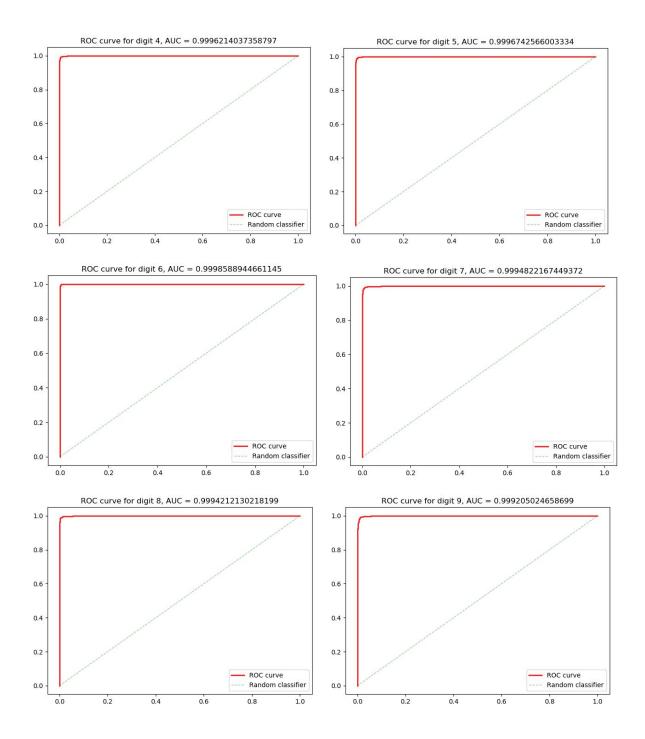


Ostateczna skuteczność na zbiorze testowym to ~97,7%.

Krzywe ROC i współczynniki AUC

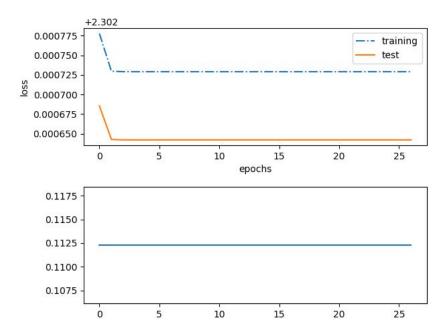
Jakość uzyskanych predykcji oceniono przy pomocy krzywych ROC i wskaźnika AUC. Wyniki trzeba określić jako zadowalające, gdyż wszystkie współczynniki AUC są większe od 99,9%. Najgorszy efekt otrzymano dla cyfry "9", dla której wskaźnik ten wynosi 99,92%.





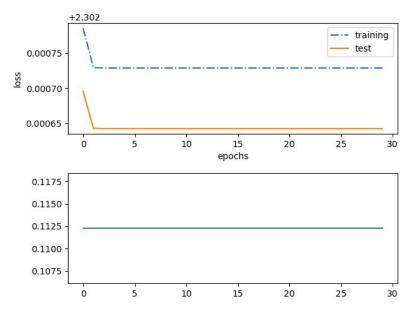
Uczenie przy dużej liczbie warstw

Najlepszy wynik (97,7%) osiągnięto dla sieci o dwóch warstwach. Przeprowadzono też badania architektur o większej ich liczbie. Pierwszy test przeprowadzono dla sieci o 20 warstwach o rozmiarze 64 neurony w każdej:

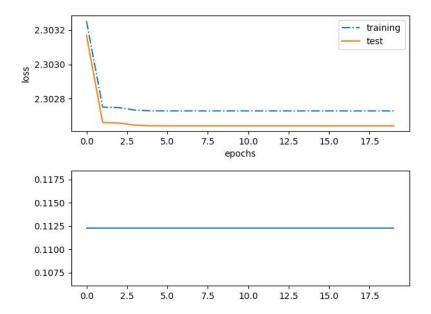


Jak widać, wynik osiągnięty na takiej sieci jest bardzo słaby. Przewidywane wyniki poprawne są tylko w 11,2% przypadków.

Następny test polegał na dwukrotnym zwiększeniu liczby neuronów w warstwie przy zachowaniu 20 warstw. Wynik jest niemalże identyczny:



Ostatni z testów tego typu przeprowadzono dla 40 warstw po 64 neurony:



I w typ przypadku sieć zatrzymuje się w optimum lokalnym. Wniosek, jaki można na tej podstawie wysnuć jest taki, że zbyt głębokie sieci w takiej architekturze nie radzą sobie z tym problemem. Mają one skłonność do osiągania optimów lokalnych. Wynika to prawdopodobnie z bardzo dużej wymiarowości przeszukiwanej przestrzeni przy relatywnie niewielkiej ilości danych wejściowych.

Czego się nauczyliśmy i co zrobilibyśmy lepiej

- Podczas tego projektu nauczyliśmy się na pewno współpracy z użyciem systemu kontroli wersji Git - używanie osobnych feature-branchy, mergowanie i podstawowe code review.
- Serializowanie danych uczących i używanie niepełnej wersji bazy danych to z pewnością elementy które przyspieszyły proces tworzenia kodu programu. W przyszłości na pewno będę je implementował wcześniej.
- Z pewnością powinniśmy przykładać większą uwagę do solidnego code review, bo kilka błędów w kodzie można by było wykryć dzięki temu znacznie wcześniej.
- Testowanie co najwyżej jednej nowej funkcjonalności na raz. W przypadku dwóch często szukaliśmy błędu tam gdzie go nie było.
- Fun Fact: Użycie sygnału dźwiękowego na końcu pętli uczącej pozwoliło na wykonywanie innych zadań w czasie uczenia i szybkie przypomnienie o zakończonym uczeniu