

WOJSKOWA AKADEMIA TECHNICZNA

im. Jarosława Dąbrowskiego

WYDZIAŁ CYBERNETYKI



Sprawozdanie

Zaawansowane metody uczenia maszynowego

Sieci rekurencyjne – rozpoznawanie języków

Autor:

Karol Baranowski

Prowadzący:

mgr inż. Przemysław Czuba

Spis treści

Zadanie.....	3
1. Tutorial.....	4
2. Klasyfikacja języków.....	4
3. Ulepszenie klasyfikacji.....	5
a. Dodatkowa warstwa liniowa.....	6
4. Macierz błędów.....	7

Zadanie

Zadanie:

1. Wykonać oraz zapoznać się z tutorialiem dotyczącym klasyfikacji imion do języka, z którego pochodzi:

Kod: <https://github.com/spro/practical-pytorch/tree/master/char-rnn-classification>

Tutorial: https://pytorch.org/tutorials/intermediate/char_rnn_classification_tutorial.html

2. Wykorzystać kod z poprzedniego punktu, do klasyfikacji słów do języka, z którego pochodzi. Użyć co najmniej 2 języków.

Zbiór danych słów można wykorzystać z wytrenowanych modeli spaCy¹.

```
# Instalacja: sudo python3 -m spacy download en_core_web_md
```

```
en = spacy.load("en_core_web_md")
```

```
list(en.vocab.strings)
```

3. Ulepsz klasyfikację poprzez zmiany w strukturze sieci:
 - a. Dodaj więcej warstw liniowych
 - b. Wypróbuj warstwy `nn.LSTM` oraz `nn.GRU`
4. Przedstaw macierz błędów (*confusion matrix*) oraz porównaj wyniki przed zmianami w sieci oraz po zmianach.

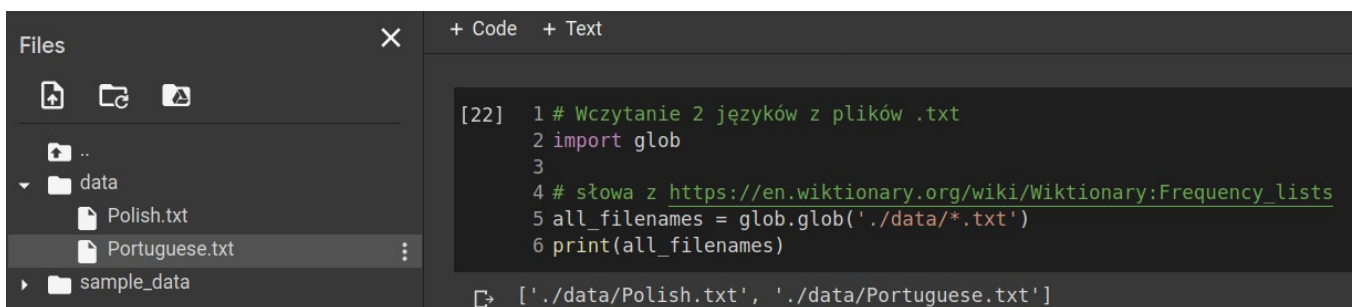
1. Tutorial

Zapoznano się z tutorialiem, przeczytano objaśnienia, komentarze i uwagi po czym użyto kodu do realizacji punktu drugiego.

2. Klasyfikacja języków

Jako zbiory treningowe użyto 1000 najbardziej popularnych słów w językach polskim i portugalskim z list:

https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists



```
[22] 1 # Wczytanie 2 języków z plików .txt
      2 import glob
      3
      4 # słowa z https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists
      5 all_filenames = glob.glob('./data/*.txt')
      6 print(all_filenames)
```

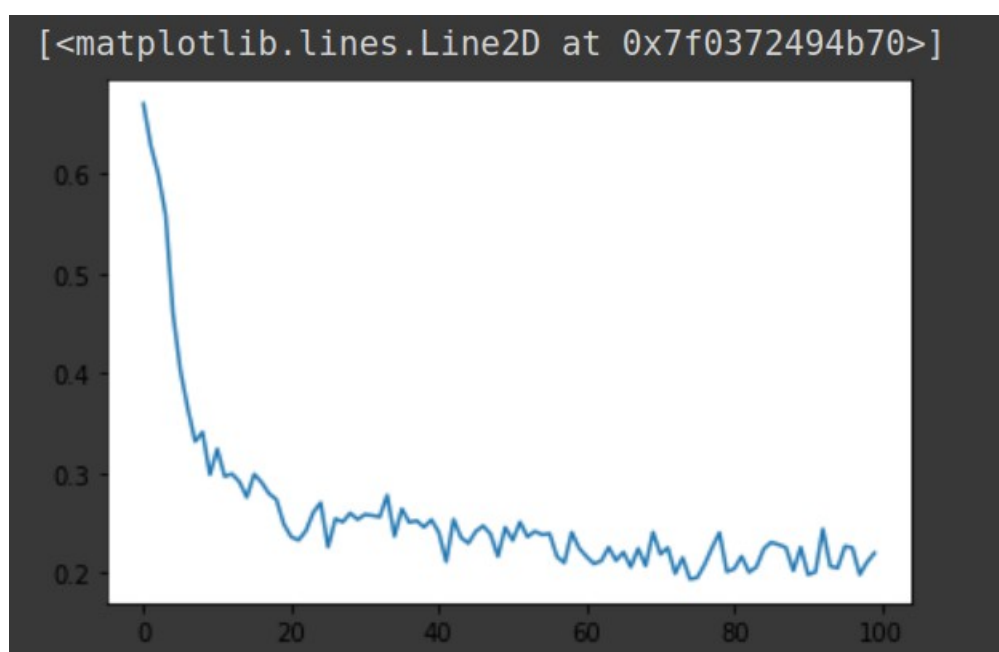
['./data/Polish.txt', './data/Portuguese.txt']

Modyfikując nieznacznie kod z linku z zadania udało się wytrenować sieć do rozpoznawania słów z tych języków:

```
5000 5% (0m 7s) 0.2128 isc / Polish ✓
10000 10% (0m 14s) 0.0063 wracam / Polish ✓
15000 15% (0m 21s) 0.0196 mike / Polish ✓
20000 20% (0m 28s) 0.0169 fica / Portuguese ✓
25000 25% (0m 35s) 0.0279 mike / Polish ✓
30000 30% (0m 42s) 0.1072 muita / Portuguese ✓
35000 35% (0m 49s) 0.0000 zaczekaj / Polish ✓
40000 40% (0m 56s) 0.2849 trata / Portuguese ✓
45000 45% (1m 4s) 0.1393 musimy / Polish ✓
50000 50% (1m 11s) 0.1235 stronie / Polish ✓
55000 55% (1m 18s) 0.0324 jedz / Polish ✓
60000 60% (1m 26s) 0.0144 razu / Polish ✓
65000 65% (1m 33s) 0.0216 wedug / Polish ✓
70000 70% (1m 40s) 0.0074 prawde / Polish ✓
75000 75% (1m 47s) 0.0003 velho / Portuguese ✓
80000 80% (1m 54s) 0.0012 uwage / Polish ✓
85000 85% (2m 1s) 0.0699 niz / Polish ✓
90000 90% (2m 9s) 0.0000 wyglada / Polish ✓
95000 95% (2m 16s) 0.2020 obok / Polish ✓
100000 100% (2m 23s) 0.2089 comida / Portuguese ✓
```

Ustawiono 100000 epok trenowania. Każda pętla treningu:

- tworzy tensory wejściowe i wyjściowe
- tworzy wyzerowany stan ukryty
- czyta każdą literę w wyrazie i zachowuje stan ukryty dla następnej litery
- porównuje wyniki końcowe z celem
- propagacja wsteczna
- zwrócenie wyniku i straty



W miarę przemijania epok koszt znacząco zmalał, co oznacza, że sieć powinna umieć rozpoznawać portugalskie i polskie słowa.

3. Ulepszenie klasyfikacji

Początkowa klasyfikacja kilku losowo wybranych słów niebędących w zbiorze treningowym:

```
> osmiornica
(-0.02) Portuguese
(-4.00) Polish

> polvo
(-0.02) Portuguese
(-4.00) Polish

> chomik
(-0.01) Polish
(-4.27) Portuguese

> hamster
(-0.03) Portuguese
(-3.60) Polish

> jezyk
(-0.00) Polish
(-9.85) Portuguese

> lingua
(-0.10) Portuguese
(-2.35) Polish
```

Sieć bez modyfikacji jest w stanie rozpoznać bardzo dobrze słowa obu języków (im wyższa wartość przy danym języku tym większe prawdopodobieństwo, że słowo jest w tym języku.) Może być to spowodowane tym, że danymi treningowymi były listy 1000 najbardziej popularnych w danym języku słów, więc zbiory te dobrze charakteryzowały język i były one duże i różnorodne. Oprócz tego języki te pochodzą z innych rodzin języków i nie są podobne tzn. ustawienia sąsiadujących liter jak i częstotliwość występowania danej litery różni się co ułatwia sieci naukę różnic.

a. Dodatkowa warstwa liniowa

Dodano jedną warstwę liniową

```

> osmiornica
(-0.01) Portuguese
(-4.44) Polish

> polvo
(-0.00) Portuguese
(-5.68) Polish

> chomik
(-0.06) Polish
(-2.86) Portuguese

> hamster
(-0.01) Portuguese
(-4.67) Polish

> jezyk
(-0.00) Polish
(-10.41) Portuguese

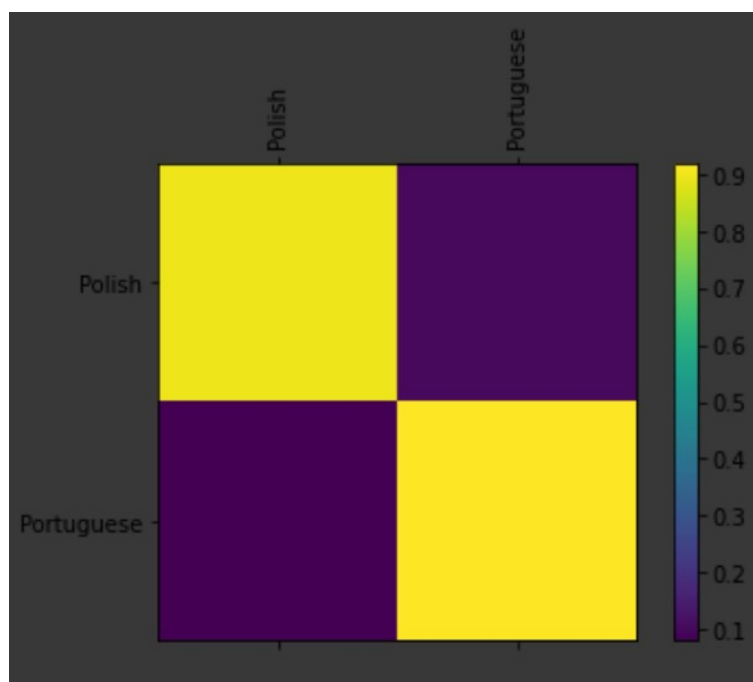
> lingua
(-0.13) Portuguese
(-2.08) Polish

```

Powyżej widać, że po dodaniu warstwy liniowej (zaznaczona w kodzie komentarzem) różnice w wartości dla języków zmniejszyły się, a więc warstwa ta nie przyniosła korzyści w tym przypadku.

4. Macierz błędów

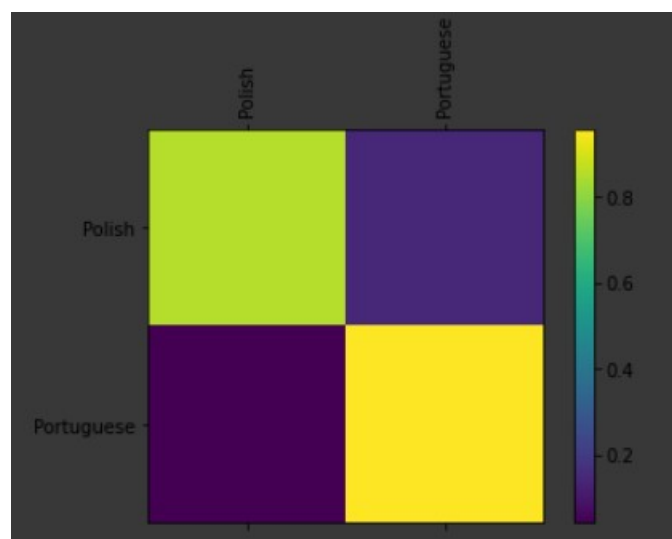
Początkowa macierz:



8

Zarówno dla polskiego jak i portugalskiego sieć rozpoznaje wyrazy na poziomie poprawności $> 90\%$.

Macierz po modyfikacji punktu 3a:



Widać, że po dodaniu warstwy ukrytej sieć gorzej rozpoznawała polskie słowa jak i częściej myliła je ze słowami portugalskimi (poprawność spadła na $\sim 80\%$).