Deep Learning: Assignment 8

Confusion Matrix Description

Assem Kussainova
Data Science
Nazarbayev University
Nur-Sultan, Kazakhstan
assem.kussainova@nu.edu.kz

I. Task

As a task, it was given to write a program for determining the belonging of a name to nationality using the specified code. The objectives of the assignment included running the code and understanding its implementation. After that, it was required to build a matrix of errors and describe the obtained results for the specified items, that is, the results in relation to Kazakh names.

II. Data

Firstly, data was downloaded from the given link and extracted to the working directory. It included 18 text files named as "[Language].txt". Each file contains a bunch of names, one name per line, mostly romanized. For completing task, Kazakh names dataset file was collected and added to the rest of the files. Kazakh names were included into the model and to the confusion matrix, so that it became 19x19. Confusion matrix results in relation to the Kazakh names data are described in the section below.

III. Confusion Matrix

All the data was preprocessed and each name from files was translated into tensor. For the model was chosen Recurrent Neural Network, which has simple structure, containing 2 linear layers, which work with input and hidden state, and 1 softmax layer for the output. When working with network, each letter of the name is transferred to the model, while keeping the hidden state for the next letter. After output from the RNN obtained, it compared with real target and back propagation is made for reducing the loss. Graph of losses changing over iterations is made and confusion matrix is built for results evaluation.

From the Fig.1 can be seen the illustration of the confusion matrix, where each row is the real ethnicity of the names and each column holds the scores for guesses of name ethnicity. It can be noticed that names with Kazakh ethnicity are correctly guessed in approximately 60% of cases. If we look through each column, it is seen that Kazakh names are misclassified with other ethnicities including Arabic, Japanese, Russian and some of the European. There are could be several reasons for that. Firstly, when examining the files of names for different ethnicities, it was discovered that some of them contain names which are related not only to one particular nation. For example, name 'Aida' is listed among the Japanese names, while being popularly used in Kazakhstan and other countries. Therefore the results will not be accurate, as data is not consisting only from names related only to one ethnicity. Secondly, this model classifies the Kazakh names as Arabic or Russian as they have common similarities in origin and usage. For example, name Assem is used both in Kazakh and Arabic societies. So, because of this we can see both ethnicities in list of predictions when pass the input name to the prediction. Both this factors cause the distortions in prediction of names ethnicity results.

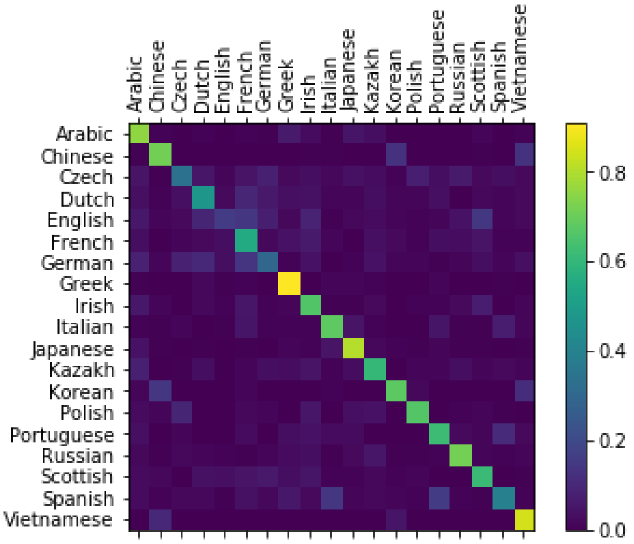

Fig. 1. Confusion matrix of testing results

IV. Conclusion

In this work, the RNN model was studied and built to predict the nationality of a name with the addition of a dataset of Kazakh names. In general, the prediction accuracy for Kazakh names was about 60%, which is due to several factors of inaccuracy in the remaining data, such as the presence of common names among several nationalities and the similarity of the origin of common names between nations. The code used and the dataset of Kazakh names have been attached along with this description for closer study.