Deep Learning with Python
Summer Semester 2020
University of Konstanz
Miriam Butt

Assignment 1

This assignment is due on Monday the **22nd of June**.

1 Sentiment Analysis

We have now gone through some basic examples of Deep Learning. (Chollet Ch. 3) and it is now time to practice doing some Deep Learning yourself.

Your task is to train to a network so that it can discriminate whether a tweet constitutes hate speech or not.

1.1 Data Set

For this purpose we will use the data-set provided on this site:

`https://data.world/crowdflower/hate-speech-identification`.

You can find it under "data" in Assignment 1 on Ilias. You should use the data in the csv file. The csv format means that the individual pieces of information are separated by commas. There are six fields:

1. 'count' = number of annotators who coded each tweet (minimum is 3, sometimes more users coded a tweet when judgments were determined to be unreliable).
2. 'hate_speech' = number of annotators who judged the tweet to be hate speech.
3. 'offensive_language' = number of annotators who judged the tweet to be offensive.
4. 'neither' = number of annotators who judged the tweet to be neither offensive nor non-offensive.

5. ‘class‘ = class label for majority of annotators

   0 - hate speech

   1 - offensive language

   2 - neither

6. the tweet itself

There are about 30 000 tweets, which should be enough data for training a Deep Learning network. It is part of your task to find out!

1.2 Data Preprocessing

So far we have been working with nicely pre-processed data that could be fed directly into a deep learning network. A large part of your exercise is to figure out how to preprocess the data so that it is ready to be fed into a network. Your task is to train a network so it can predict whether a tweet constitutes hate speech, only contains offensive language or is neutral. You therefore need to think about which information you need from the csv file (all? a subset?). You then need to figure out how to put that information: a) into a Python object/tensor; b) convert it into an appropriate format that the network can learn with.

You also need to divide the data up into an appropriate training set and an appropriate test set.

1.3 Deep Learning Network

Your next task is to figure out what type of model you want to build. For this you need to think about what type of a problem this is (binary classification? multiclass? regression?) and how many and what types of layers you want to have. You need to think about the activation function and the final layer. You should play around with training epochs, batch sizes and the amount of hidden nodes to optimize your training results.

Hand in the code per email for your optimal model along with a documentation of the code. You will be graded on the clarity of your documentation and the code and the justification of your design choices for the network.

2 K-fold Validation

Chollet also introduced you to K-fold validation. Now take just the first 20000 tweets of the data set and learn on them. Use a 4-fold evaluation technique when training on the data.

Hand in the code per email for your optimal model in this scenario along with a documentation of the code.

You will be graded on the clarity of your documentation and the code.

3 Extra Credit: Literature and Computation

In *Murder at the Vicarage* Agatha Christie first introduced the character of Miss Marple. It was published in 1930. Read through the following passage, explain how it relates to what we have been learning in the course and comment on whatever else you might find interesting.

> "You see, " she began at last, "living alone, as I do, in a rather out of the way part of the world, one has to have a hobby. There is, of course, woolwork, and Guides, and Welfare, and sketching, but my hobby is and always has been Human Nature. So varied and so very fascinating. And of course, in a small village, with nothing to distract one, one has such ample opportunity for becoming what I might call proficient in one's study. One begins to class people, quite definitely, just as though they were birds or flowers, group so and so, genius this, species that. Sometimes of course, one makes mistakes, but less and less as time goes on. And then, too, one tests oneself. One takes a little problem — for instance the gill of picked shrimps that amused dear Griselda so much — a quite unimportant mystery but absolutely incomprehensible unless one solves it right. . . . It is so fascinating, you know, to apply one's judgement and find that one is right." [Murder at the Vicarage, Ch. 26, pp. 189–190]