Exercise 3 - Language Identification

Reloaded and Convoluted

Deadlines

The deadline for Exercise 3 is **14.11.2021, 23:59 (Zurich Time).**
The deadline for the peer review is **21.11.2021, 23:59 (Zurich Time).** You will find instructions for the peer review process at the end of this document.
The deadline for feedback to your peer reviewers is **26.11.2021, 23:59 (Zurich Time).**

Learning goals

This exercise builds on the language identification task you solved in Exercise 1. By completing this exercise you should …

- … understand CNNs.
- … be able to implement CNNs in PyTorch or Keras-Tensorflow.
- … deepen your understanding of the role of hyper-parameters, regularisation, and handling class imbalance.
- … perform an error analysis of machine learning models.

Please keep in mind that you can always consult and use the exercise forum if you get stuck (note that we have a separate forum for the exercises).

Deliverables

We encourage you to hand in your solutions as a Colab-Notebook. **Download your notebook as a .ipynb file**. That way your reviewers can view and execute your code. Or can view your already executed code.

Please hand in your code and your lab report. Hand in the following files and name them exactly in the following fashion:

- ex03_cnn.ipynb
- ex03_labreport.pdf

zip it and name the zip-folder *ex03_ml4nlp1.zip*. The .ipynb files should contain your well documented AND EXECUTABLE code. We recommend you use Google's Colaboratory, where you have access to GPU time.

If you prefer to solve the exercise on your own computer please submit a zip-folder containing the required files:

- ex03_cnn.ipynb
- ex03_labreport.pdf

We assume that the data files are in the same folder as the scripts, e.g.

- ex03_cnn.ipynb
- data.tsv

Please submit the lab report in PDF format. The lab report should contain a detailed description of the approaches you have used to solve this exercise. Please also include results. In this exercise description, we highlight places in green where we expect a statement about an issue in your lab report.

Please note:

- Your peers need to be able to run your code. If it does not work, you will not be able to obtain the maximum number of points.
- DO NOT submit the data files!

Data

For this exercise, you will work with the same data as for Exercise 1. The goal is again to predict the language of Tweets. This is an extension of the problem described in Goldberg, chapter 2. However, we will work with more languages than just six and the text segments we need to classify are much shorter.
If possible, use the same training, development and test split as you did for Exercise 1. Otherwise, your results will not be comparable.

The material folder in the exercise section of OLAT contains the two files “train_dev_set.tsv” and “test_set.tsv”. The files are also published under these two links:

- train_dev_set.tsv: https://docs.google.com/spreadsheets/d/e/2PACX-1vTOZ2rC82rhNsJduoyKYTsVeH6ukd7Bpxvxn_afOibn3R-eadZGXu82eCU9IRpl4CK_gefEGsYrA_oM/pub?gid=1863430984&single=true&output=tsv
- test_set.tsv: https://docs.google.com/spreadsheets/d/e/2PACX-1vT-KNR9nuYatLkSbzSRgpz6Ku1n4TN4w6kKmFLkA6QJHTfQzmX0puBsLF7PAAQJQAxUpgruDd_RRgK7/pub?gid=417546901&single=true&output=tsv

To make the start a little easier, you can go to this notebook in Google Colab (same as for exercise 1) which loads the two files using the public links. If you want to, you can just continue the exercise in your own copy of that notebook. If you choose to work locally, download the two files on your computer.

Language Identification with a CNN

Implement a language classifier in PyTorch or Keras-Tensorflow. We suggest reusing and adjusting the class structure from exercise 2 (which may be inspired by Rao and McMahan). However, you are free to create your own, new class structure. Keep in mind that for language classification we work on the character level. Thus, your Vocabulary class (that is, if you have one) will not hold a vocabulary of words, but a vocabulary of characters.

Remember to document your code with docstrings and/or comments and/or text cells.

1. Your goal is to find the optimal model architecture and training regime for your CNN classifier. Try out at least five different combinations of the following hyperparameters, which you consider well-performing, and report the combinations and corresponding results (accuracy and F_1-macro) on the **development set** in a table.:
    a. optimizer
    b. learning rate
    c. dropout
    d. # of filters
    e. different strides
    f. different kernel sizes
    g. different pooling strategies
    h. batch sizes
    i. any other thing you want to test

    Keep track of the loss to interrupt early if a model does not converge.
2. Take the best performing model and evaluate it on the test set. Report your test set results (accuracy and F_1-macro) in the forum “Exercise 3 Test Set Scores”. **Important:** Only evaluate on the test set once and do not use the devset for retraining with the same hyperparameters, please, use only the official training set for training. Do not optimize hyperparameters based on test set results.

3. Reason about the observed effects of your 5 best hyperparameter settings on model performance. You do not need to be sure that the reasons you provide are correct – the goal is to provide educated (or well-reasoned) guesses.

4. Compare the outputs of the best CNN model to your best performing model from Exercise 1. Which classifier scores higher on the test set? Do you have an idea, why this might be?

Important

Please make sure you run your code on Google Colab with GPU selected. Select the GPU via “Edit” → “Notebook Settings” and then choose the GPU hardware accelerator.

Peer Review Instructions

If you are not already registered on Eduflow follow this link https://app.eduflow.com/join/TDY5CZ and register with the E-mail address you use for OLAT. Then you should be added to the course page automatically.
As soon as the deadline for handing in the exercise expires you will have time to review the submissions of your peers. You need to do **2 reviews** to get the maximum number of points for this exercise.
Here some more rules:

- If you do not submit 2 reviews, the maximum number of points you can achieve is 0.75 (from a total of 1).
- Please use full sentences when giving feedback.
- Be critical, helpful, and fair!

- **All reviews are anonymous: Do not put your name into the python scripts, the lab report or the file names.**
- You must also give your reviewers feedback. The same criteria as above apply.
- If you consistently provide very helpful feedback, you can be awarded with a bonus of 0.5 in total in case you didn't achieve the full 6 points from all exercises. A maximum of 6 points from the exercises can go into the final grade.

Groups:

- You can create groups of two to solve the exercise together.
- Both students should submit the solutions separately.
- If you did not already work together for the previous exercise, write a small post in the “Groups”-thread in the exercise forum on OLAT to notify the instructors about the group.
- As a group member, you still have to review two submissions with your own eduflow account. However, you may work together in the group to write all 4 reviews.