EEE 486/586

Statistical Foundations of Natural Language Processing

Assignment 2

Fine-tuning BERT for Text Classification

(Due 07/05/2023, 23:59 PM)

General Instructions

IMPORTANT REMARK: We will conduct a tutorial session on Friday (28/04/2023 at 9:00 - 10.30).

Groups: You are expected to work alone.
TA: Arda Can Aras (please mail if you have any questions).

For this assignment, you are required to work with BERT or one of its variants for the text classification task. Since the hardware of the students may vary (GPU sizes etc.) you are free to use the BERT variant of your choice which meets your hardware requirements. You will fine-tune a pre-trained model and explore several hyperparameters and their effect on the model's performance.

Submission Guidelines:

(i) Write a brief report (max 5 pages) that explains the details of your procedure part-by-part and answer any further questions in the assignment.

(ii) Name the report "report_SurnameNameID.pdf".

(iii) **Failing to meet these requirements may result in loss of grades. Wrong naming of files and/or submitting to the wrong places in Moodle will also be penalized by grade deductions.**

Important remarks:

- Collaboration and code sharing among students are prohibited.
- You are allowed to use any libraries and frameworks you want.
- Properly label all your figures and tables throughout your report.
- Your reports will be evaluated based on the proper completion of tasks, clarity of presentation of results, the sufficiency of discussions regarding the results, quality of writing, plots, and organization of the report, and your possible insights and comments.
- There might be slight deviations in your results depending on your hardware facilities etc.
- Please see the link for information about academic honesty and plagiarism:
- You are expected to submit both your codes and report in a zip folder to Moodle. Please do not include model weights.

Part 1:

(a) You will work on one of the GLUE Bench Mark datasets, the Corpus of Linguistic Acceptability (CoLA). Please check the link to download the dataset. Alternatively, you can also follow the approach that I will do in the tutorial to download the dataset.

(b) You must evaluate your results with Matthew Correlation Coefficient (MCC).

(c) For BERT and its variants, one common approach for text classification is using the `['CLS']` tokens as document representation and then feeding these representations to a newly untrained classifier head. In this part of the assignment, you will follow this approach. Huggingface library has a predefined model BertForSequenceClassification that comes with an untrained classifier head. For convenience, please use this predefined model instead of creating your own custom classifier head. You will need to use the `push_to_hub` method of this module in the following part.

(d) You are required to investigate the effect of the following hyperparameters on to model's performance. You can follow the same approach that I will mention in the tutorial with the `Optuna` library.
List of hyperparameters to investigate:

- Learning Rate
- Number of Epochs
- Max Length of the input sequence
- Dropout at the classification head

(e) Report the loss (results and curves) and MCC results of the best-performing model that you choose.

(f) For this part, you need to push the best performing model, the huggingface. This will ensure the reproducibility of your model. We will do an example of it in the tutorial. Huggingface website includes very well-documented tutorials about using model hub along with videos. Please feel free to check them out.

Part 2:

(a) As we mentioned earlier, there are multiple ways to do text classification with the BERT model. For this part of this assignment, you need to come up with at least one different way other than using `['CLS']` tokens as document representations. You can implement any of the existing works in literature or come up with your own model.

(b) Again, investigate the effect of the learning rate, max length of the input sequence, and the number of epochs. If your new approach also includes some other parameters you should also investigate their effects.

(c) Most probably, it will be difficult to publish the model you trained here to model hub since you need to follow the hugginface model convention. You are free to do this, but it is not necessary for this part of the assignment.

Report Preparation

1. **The deadline for the Final Report is 07/05/2023, 23:59PM.** No late submissions will be allowed.
2. Each report should be typeset, **no handwriting is allowed**.
3. I recommend using LaTeX, though it is not mandatory. If you did not use it before, take this as a chance to get used to that good practice.
4. Each report should contain the following sections clearly separated with headings:

Abstract: A one-paragraph summary of all major aspects of your report from Introduction to Discussion. What is this report about? No references should be given, and the abstract should be self-contained.

Introduction: Briefly state the general topic and essence of the assignment. State the purpose of your work. Give a general description of what the rest of the report will be about.

Architectures: Explain all the architectures that you used for text classification except the one mentioned in Part 1. Clearly state the equations and model parameters.

Results: You need to mention all parts of the assignment with separate subsections where you see necessary. Present your key results, and illustrate your outputs visually with the help of figures and tables. Detailed numbers/plots should be provided in illustrations, and each figure or table should contain a paragraph-long, self-contained caption that explains the contents. Figures/tables should be referenced in appropriate sections, and the main trends/results should be stated in the text.

Discussions & Conclusions: Interpret your results in light of experimental findings. Which parts of your analyses worked, and which failed? Does the methodology have drawbacks, flaws, and room for improvement? By considering the assignment as a whole, what did you learn? Try to encapsulate the central point of the entire assignment and summarize the findings in a concise way.

References: A list of all referenced material formatted according to standard conventions used in journals. Crude, unformatted lists are not acceptable.