Assignment 4: Modelling IMDB reviews using an LLM

COMP 551 Winter 2024, McGill University
Contact TAs: Huiliang Zhang and Vicky Dong (Once the strike is over)

Released on March 27
Due on April 12 midnight

Preamble

- This assignment is **optional**. If you do manage to complete it, we will grade it as a regular assignment. We will choose the higher grade between the total grades of the 3 assignments (each worth 13%) and the total grades of the 4 assignments (each worth 10%).
- This assignment is to be completed in groups of three. All members of a group will receive the same grade except when a group member is not responding or contributing to the project. If this is the case and there are major conflicts, please reach out to the group TA for help and flag this in the submitted report. Please note that it is not expected that all team members will contribute equally. However every team member should make integral contributions to the project, be aware of the content of the submission and learn the full solution submitted.
- You will submit your assignment on MyCourses as a group. You must register your group on MyCourses and any group member can submit. See MyCourses for details.
- We recommend to use **Overleaf** for writing your report and **Google colab** for coding and running the experiments. The latter also gives access to the required computational resources. Both platforms enable remote collaborations. If you have difficulty with the colab GPUs, the department has made GPUs available to you also (max 12 GB RAM though), see the post on Ed about how to access those GPUs.

Synopsis

In this assignment, you will implement a Large Language Model (LLM) for example Bidirectional Encoder Representations from Transformers (BERT) using the existing Tensorflow or Pytorch libraries. You will evaluate your LLM against Logistic Regression (LR) (A2 but you may use scikit-learn for this assignment), Random Forest (RF), and XGBoost on the IMDB dataset you have used in your Assignment 2.

The goal is to gain hands-on experience running the modern deep learning libraries and evaluating their performances on the real-world dataset against the more traditional methods.

1 Task 1: IMDB data preprocessing

The IMDB Reviews data can be downloaded from here: `http://ai.stanford.edu/~amaas/data/sentiment/`.

For the baseline methods such as LR, RF and XGBoost, you may follow the same data preprocessing pipeline in your Assignment 2 to turn the unstructured text data into tabular format with selected highly relevant features (by z-scores, for example) as the words and text documents as the training or test examples.

For your LLM, you can directly train on the unstructured text data without breaking them into tabular format. Find out how to do that. For example here is an online tutorial `https://www.kaggle.com/code/atulanandjha/bert-testing-on-imdb-dataset-extensive-tutorial` that uses PyTorch among many others.

Same as in Assignment 2, you need to use only reviews in the “train” folder for training and report the performance from the “test” folder.

Task 2: Pre-training and/or fine-tuning an LLM

You may pre-train your LLM on the same IMDB training data. Alternatively, you can download the already pre-trained LLM model. (It is possible that you don’t have enough GPU RAM to do pre-training, in this case starting with pretrained weights is ok.) These models were trained for many days on large corpus dataset such as Wikipedia. Here are some example which we recommend, BERT(110M params): `https://pypi.org/project/pytorch-pretrained-bert/`, BERT base uncased (110M params): `https://huggingface.co/google-bert/bert-base-uncased`, GPT-2 (117M params): `https://huggingface.co/gpt2`, or RoBERTa (125M params): `https://huggingface.co/FacebookAI/roberta-base`.

After pre-training, you will need to fine-tune your LLM on the binary classification task for good/bad IMDB movie prediction. (To minimize your memory requirements, you can fine-tune only the last layer of the LLM).

Task 3: Run experiments

The goal of this project is to have you explore your LLM. *Evaluate the performance using AUROC for binary classification.*

For the baseline models namely LR, RF, and XGBoost you may use sklearn implementation directly. But make sure to report the settings for each of these methods you run even if you are using the default settings in each function.

You are welcome to perform any experiments and analyses you see fit, **but at a minimum you must complete the following experiments in the order stated below**:

1. Report the binary classification performance in terms of AUROC for your LLM, LR, RF, and XGBoost on the IMDB test data.
2. Examine the attention matrix between the words and the class tokens for some of the correctly and incorrectly predicted documents. You will need to choose one of transformer blocks and use a specific attention head for the multi-layer multi-headed transformer architecture in your LLM.

Note: The above experiments are the minimum requirements that you must complete; however, this project is open-ended. For this part, you might implement a small version of the LLM on your own from scratch (using PyTorch for example). You may explore the accuracy of different pre-trained LLM models (e.g., from here `https://colab.research.google.com/github/tensorflow/tpu/blob/master/tools/colab/bert_finetuning_with_cloud_tpus.ipynb` or here `https://pypi.org/project/pytorch-pretrained-bert/`) on the IMDB prediction. You may also try different pre-training tasks such as MLM (masked language model) and NSP (next sentence prediction) directly on the IMDB data and report the resulting accuracy after fine-tuninig. Try word2vec (`https://wikipedia2vec.github.io/wikipedia2vec/pretrained/`) to project the tokens from the IMDB review onto embedding space and then use RF, LR, XGBoost, and/or MLP to classify the embedded documents. Compare two different LLMs (for example BERT, Mistral `https://huggingface.co/mistralai/Mistral-7B-v0.1`, or GPT-2 (`https://huggingface.co/gpt2`). As before, you do not need to do all of these things, but look at them as suggestions and try to demonstrate curiosity, creativity, rigour, and an understanding of the course material in how you run your chosen experiments and how you report on them in your write-up.

Deliverables

You must submit two separate files to MyCourses (**using the exact filenames and file types outlined below**):

1. `assignment4_group-k.ipynb`: Your data processing, classification and evaluation code should be all in one single Jupyter Notebook. Your notebook should reproduce all the re-

sults in your reports. The TAs may run your notebook to confirm your reported findings.

2. `assignment4_group-k.pdf`: Your (max **8-page**) assignment write-up as a pdf (details below). Compared to the past 3 assignments (5-page each), we provide 3 extra pages for this assignment to provide you more space to explore.

where k is your group number.

Project write-up

Your team must submit a project write-up that is a maximum of eight pages (single-spaced, 11pt font or larger; minimum 0.5 inch margins, an extra page for references/bibliographical content can be used). We highly recommend that students use LaTeX to complete their write-ups. You have some flexibility in how you report your results, but you must adhere to the following structure and minimum requirements:

Abstract (100-250 words) Summarize the project task and your most important findings. For example, include sentences like "In this project we investigated the performance of BERT on IMDB movie review dataset.", "We found that BERT achieved worse/better accuracy than the more traditional ML methods and was significantly faster/slower to train. **We achieved test AUROC of ??.????% using BERT.**"

Introduction (5+ sentences) Summarize the project task, the IMDB datasest (very briefly since you did this on A2), and your most important findings. This should be similar to the abstract but more detailed. You should include background information and citations to relevant work (e.g., other papers analyzing these datasets).

Datasets (5+ sentences) Very briefly describe the IMDB dataset and how you processed them specifically for your LLM. Present the exploratory analysis you have done to understand the data, e.g. distribution of lengths of the reviews.

Benchmark (10+ sentences) Clearly describe the settings for each method including your LLM (architecture, number of attention heads, number of transformer layers, etc), LR (e.g., ℓ_1 or ℓ_2 regularization penalty), RF (e.g., number of trees), XGBoost (e.g., weak learners) you ran even if you are using the default settings in each function.

Results (5+ sentences corresponding to 5 figures) Describe the results of all the experiments mentioned in Task 2 and 3 (at a minimum) as well as any other interesting results you find. At a minimum you must have these 5 plots:

1. A single plot containing four ROC curves of your LLM, LR, RF and XGBoost on the IMDB test data.

2. A bar plot that shows the AUROC of your LLM, LR, RF, and XGBoost on the IMDB test data.
3. A horizontal bar plot showing the top 20 important features from RF on the IMDB data with the feature importance scores as the x-axis and the feature names (i.e., words) as the y-axis.
4. Two heatmaps showing one of the attention matrices (from one of the head in one of the layers) between words for a correctly predicted positive and negative IMDB review (i.e., good movie) with high probability. Display the attention between the top words and the [CLS] token.
5. Same as above but now show the two attention heatmaps for the incorrectly predicted positive and negative movie reviews

Discussion and Conclusion (5+ sentences) Summarize the key takeaways from the project and possibly directions for future investigation.

Statement of Contributions (1-3 sentences) State the breakdown of the workload across the team members.

Evaluation

The assignment is out of 100 points, and the evaluation breakdown is as follows:

- Completeness (20 points)
  - Did you submit all the materials?
  - Did you run all the required experiments?
  - Did you follow the guidelines for the project write-up?
- Correctness (40 points)
  - Are your models implemented correctly?
  - Are your reported accuracies close to the reference solutions?
  - Does your the RF important features make sense?
  - Do the attention heatmaps you choose to display make sense?
- Writing quality (25 points)
  - Is your report clear and free of grammatical errors and typos?

  - Did you go beyond the bare minimum requirements for the write-up (e.g., by including a discussion of related work in the introduction)?
  - Do you effectively present numerical results (e.g., via tables or figures)?
- Originality / creativity (15 points)
  - Did you go beyond the bare minimum requirements for the experiments?
  - **Note:** Simply adding in a random new experiment will not guarantee a high grade on this section! You should be thoughtful and organized in your report.

Final remarks

You are expected to display initiative, creativity, scientific rigour, critical thinking, and good communication skills. You don't need to restrict yourself to the requirements listed above - feel free to go beyond, and explore further. You can discuss methods and technical issues with members of other teams, but **you cannot share any code or data with other teams.**