NLP Assignment 2

Marco Acerbis, Xiaotian Fan, Lesi Li
Master's Degree in Artificial Intelligence, University of Bologna
{ marco.acerbis, xiaotian.fan, lesi.li }@studio.unibo.it

Abstract

In this assignment we focus on using a transfer learning approach in order to fine tune pre-existed models. We adapted a Seq2Seq generative approach using transformers and working with data from the CoQA dataset. As a large-scaled dataset, CoQA can provide a more complex and comprehensive linguistic environment, which can help our model to obtain a more accurate results.
The developed code has been run and tested on Google Colab.

1 Introduction

In our work, we are trying to address the task of fine-tuning a Question-Answering (QA) model and ,in particular, we studied two different transformer models: DistillRoBERTa (Yinhan Liu and Stoyanov, 2019; Sanh et al., 2019) and BERT-tiny (Turc et al., 2019), both available through the *Hugging Face* repository.
We carried out a series of experiments to study the behavior of the two models in completing two similar tasks:

- Initially both models have been trained to solve the task $f_\Theta(P, Q)$, where given as inputs a series of text passages (P) and questions (Q), the models had to generate the relative answers;
- In the second task, defined by $f_\Theta(P, Q, H)$, both models have been trained also including the dialogue history (H), i.e. we included in the inputs the previous questions and answers on the same text.

For each configuration, we performed multiple train/evaluation seed runs each time the models have been trained for a total of 3 epochs. As for the evaluation of the results, we implemented a metric using the SQUAD-F1 score from the *allennlp* library.

2 System description

Our solution, developed starting from the Hugging Face Seq2Seq transformer tutorial, is composed as follow:

- Data Initialization and Pre-Processing: after downloading CoQa, we format the whole dataset in order to make it compatible with our pre-processing function. Also some analysis on the content was performed, as described in the next section.
- Model Selection: the desired model to fine-tune is selected and loaded with the corresponding tokenizer. Both models use an encoder max length of 512 and the truncation mode has been set as "only second", while the decoder max length has been set to 128;
- Tokenization: this phase is a continuation of the data pre-processed in order to prepare the inputs to be fed into the chosen model. In particular two different function have been created for this task: the first one does not include the conversational history in the inputs, while the second function does;
- Fine-Tuning: the model is re-trained on the chosen dataset, in order to perform generative QA, for a total of 3 epochs. At the end of each epoch the model is evaluated on the validation set by computing the SQUAD-F1 score;
- Test and Results Evaluation: similarly to the first point, the test set is prepared, tokenized and fed into the model that now generates the answers. The obtained results are then evaluated using again the SQUAD-F1 score, also the best five and worst five outputs (according to the score) are shown for a qualitative comparison analysis.

3 Data

We worked with the CoQa (Siva Reddy and Manning, 2019), a dataset created to build Conversational Question Answering systems. In particullar, it is composed of a total of 127K Question-Answers pairs and 8K conversations have been collected with contents covering seven diverse domains: children' Stories, literature, mid/high school exams, news, Wikipedia articles, Reddit posts and science. Also, each conversation is made of 15 turns.
Note that, due to to some timer related constraints of *Google Colab*, we have been force to work with a subset of 10000 samples for BERT-Tiny, and only 1000 samples with DistilRoberta.
CoQa is divided into two subsets: a training set and a validation set. Therefore, to be able to fine-tune our models and test them, we worked with the following sets:

- **Training Set** made of 80% of the reduced training set;
- **Validation Set** composed of the remaining 20% of the reduced training set;
- **Test Set** of 10000(BT)/1000(DR) examples from the original validation set for our evaluation.

History	Seed	Training Loss	Validation Loss	Val F1	Test F1
✓	42	0.0666	0.000696	0.000047	0.000146
✓	2022	0.0662	0.00146	0.000049	4.45e-05
✓	1337	0.0005	1.43	0.000059	8.26e-05
•	42	0.0675	0.000706	0.000046	0.000127
•	2022	0.0651	0.006593	0.000049	0.000167
•	1337	0.0052	1.6837	0.0000595	0.000157

Table 1: Bert-Tiny results

History	Seed	Training Loss	Validation Loss	Val F1	Test F1
✓	42	0.0027	0.00047	0.0	0.00028
✓	2022	1.587	0.0018	0.00014	0.0101
✓	1337	0.0085	0.00011	9.17e-05	0.00281
•	42	0.0024	8.12e-05	0.0004	0.0027
•	2022	0.0085	0.000429	0.000339	0.0012
•	1337	0.0055	5.11e-05	0.0	0.00039

Table 2: DistilRoBerta results

4 Results

In order to evaluate the models under analysis, we performed a series of experiments. Each model has been trained with or without conversational history for a total of 3 epochs for each run. We have carried out multiple training and validation runs with different training seeds (42, 2022, 1337).
The obtained results are reported in Tab.1, for BERT-Tiny, and in Tab.2, for DistilRoBerta. Unfortunately both models did not perform well, but still we can see that DistilRoberta performed slightly better than BERT-Tiny in both experiments, with or without conversational history. On thew other side, we have also to remember that DistilRoberta has been trained and evaluated on less data when compared to BERT-tiny.
We have also carried out a qualitative analysis by visioning the best 5 and worst 5 generated answers by F1-score. Generally the predicted answers consisted in long sentences of repeated words that in some cases resembled the correct answer (share some keywords for example). In particular we always got examples from "Wikipedia", even if it is not the predominant source in our data.

5 Discussion and Conclusion

In this assignment we wanted to study the possibility of fine-tuning BERT-Tiny and DistilRoberta on the CoQA datset to perform generative Question-Answering. The main obstacle we had to face was the limited capacities of the tools availble to us: not only the training and evaluation times were long and sometimes unfeasible for us, but we have also been force to work with limited data in order to avoid OOM errors. We still managed to prepare a working pipeline to re-train both transformers and evaluate the resulting models. and that allowed us to conduct a series of experiments that showed how DistilRoberta, a more robust architecture compared to BERT-Tiny, should perform better. We also expected to see longer training/evaluation times and a greater request of resources from DistilRoberta, given it's a bigger model than BERT-Tiny.

References

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

Danqi Chen Siva Reddy and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *ArXiv*.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962.

Naman Goyal Jingfei Du Mandar Joshi Danqi Chen Omer Levy Mike Lewis Luke Zettlemoyer Yinhan Liu, Myle Ott and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *ArXiv*.