Assignment 2: QA with Transformers on CoQA

Andrea Alfonsi, Gianluca Di Tuccio, Lorenzo Orsini
Master's Degree in Artificial Intelligence, University of Bologna
{andrea.alfonsi2, gianluca.dituccio, lorenzo.orsini}@studio.unibo.it

Abstract

In this assignment we performed two question-answering tasks on the CoQA dataset with two Seq2Seq encoder-decoder models. In particular, the two models were initialized with a warm-starting encoder-decoder by using the pre-trained models Distil-RoBERTa and BERT-Tiny, and then fine-tuned for our specific tasks. In the end, the best model for both tasks, according to the SQUAD-F1 score, was the one that uses Distil-RoBERTa.

1 Introduction

Question answering (QA) aims at providing correct answers to questions based on some given context or knowledge; in the last few years, this task has seen considerable growth thanks to the advent of transformers [2] .
In this assignment, two different QA tasks have been considered on the CoQA dataset[3]. In the first one, given a story and a question as input, a model should output the correct answer. In the second, a model should answer a question given its story and the dialogue from which the question belongs; the dialogue is the sequence of questions and answers previously seen by the model in the same story. For these purposes, we used two Seq2Seq models consisting of an encoder-decoder pair, both initialized with the warm-starting encoder-decoder [5]. Finally, the models were fine-tuned on our two Q&A tasks for three epochs.
In the end, the results showed that the model initialized with Distil-RoBERTa's weights has a much better performance than the one with BERT-Tiny's ones. They also showed that adding the conversational history to the input isn't that helpful for the model, although this consideration should be confirmed by a longer fine-tuning step, since for this assignment it was requested to perform it for only three epochs.

2 System description

First of all, we have downloaded the CoQA dataset: since we had two tasks that require different inputs, we created two specific dataframes. For the simple Q&A task, we created a dataframe where for each question-answer pair we have one column for the question, one column for the respective story and another for the answer. For the Q&A task that considers also the dialogue, we simply used the same structure of the previous one, but with the question-answer pairs concatenated with their respective story.
Then, we had to choose carefully the input length respecting also the pre-defined encoder input length of the two pre-trained models. In particular, the input of the first model is the concatenation of a question and its story, while for the second we also concatenated the previous question-answer pairs. For both models, the concatenation always leaves the story as the last element: in this way we can choose to truncate only the second part of the input (the story), avoiding thus losing the input question. However, for both of the tasks, the input length is always the 98 or 99 percentile of the input length distribution, so that only a very little fraction of the dataset is truncated.
Finally, we downloaded the tokenizers of the two pre-trained models and applied them to both the dataframes so that each example will be now ready to be fed into the two models.

3 Experimental setup and results

In this assignment, we used two Seq2Seq models that are initialized with a warm-starting encoder-decoder of two pre-trained models. In particular, the first model uses as an encoder block the weights of the pre-trained model Distil-RoBERTa, while the second uses the weights of BERT-Tiny. Specifically, with warm-starting, the pre-trained model's architecture is compared to the encoder's architec-

ture and all layers of the encoder that also exist in the pre-trained model will be initialized with the pre-trained weight parameters of the respective layers. All layers of the encoder that do not exist in the pre-trained model will simply have their weight parameters randomly initialized. Then, the model is fine-tuned for three epochs on the two Q&A tasks, using the beam search method to generate the answers.

4 Discussion

First of all, Table 1 sums up the results obtained by each model on the test set, considering as a metric the SQUAD-F1 score. Since we performed the training on just three epochs, the results are very low compared to the ones in literature [1].
As we expected, the model initialized with Distil-RoBERTa's weights always outperforms the one with BERT-Tiny's weights, since the first pre-trained model is more powerful than the second one. Surprisingly, we can notice that for each model the performances without considering also the previous question-answer pairs are slightly better. This lower performance could be explained due to the longer input length: in fact, having also the history, the input is longer than the previous task, so for the model is more difficult to focus on the right input portion, considering also that the fine-tuned step was performed here for three epochs. However, considering also history as input allows obtaining more accurate answers to questions that require knowing information about previous dialogues.
To perform the error analysis, we looked at the worst five answers generated by the two models. First of all, we have to highlight that sometimes the answers generated by the models don't have the same words as the true answers, but they are semantically right; in these cases, the SQUAD F1 metric isn't able to assign a satisfying score. Going deeper into each error, we can notice that both models understand the type of information to generate, but they report the wrong one (more on notebook chapter 9). In addition, by comparing the worst answers with the best ones we can see that both models struggle with generic questions that don't provide much context.
Finally, we also noticed that the best answers generated by the model initialized with BERT-Tiny's weights are boolean yes-or-no, while the other model has the best performances also to questions that need to generate more complex answers; this can be explained by the fact that Distil-RoBERTa is a more powerful model than BERT-Tiny (which has much fewer parameters).

Model	Task	SQUAD F1
Distil-RoBERTa	1	0.31
BERT-Tiny	1	0.09
Distil-RoBERTa	2	0.29
BERT-Tiny	2	0.08

Table 1: SQUAD F1 for the two models on each task; specifically, task 1 is Q&A without the history, while task 2 considers also the previous question-answer pairs.

5 Conclusion

In this assignment, we performed fine-tuning on the two pre-trained models Distil-RoBERTa and BERT-Tiny for two question-answering tasks on the CoQA dataset. The results showed that the model initialized with Distil-RoBERTa's weights has a much better performance than the one with BERT-Tiny's weights in both tasks. In addition, we saw that adding the history of the dialogue worsens the models' performance of about 2%, but at the same time it's more effective on questions that require the knowledge of conversational history. For this reason, it could be interesting to repeat these experiments with more epochs for the fine-tuning steps, to see if adding the history to the input could lead to a better performance of both models. Furthermore, as we already mentioned in the previous chapter, F1-SQUAD is a metric that doesn't take into account the semantic meaning of the words, and so it is prone to consider wrong answers that are semantically correct. As an alternative, we could use other metrics that take into account this aspect, such as Semantic Answer Similarity [4].

References

[1] Imen Akermi, Johannes Heinecke, and Frédéric Herledan. 2020. Transformer based natural language generation for question-answering. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 349–359, Dublin, Ireland. Association for Computational Linguistics.

[2] Kate Pearce, Tiffany Zhan, Aneesh Komanduri, and Justin Zhan. 2021. A comparative study of transformer-based language models on extractive question answering. *CoRR*, abs/2110.03142.

[3] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. Coqa: A conversational question answering challenge. *CoRR*, abs/1808.07042.

[4] Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. *CoRR*, abs/2108.06130.

[5] Patrick von Platen. 2020. Leveraging pre-trained language model checkpoints for encoder-decoder models.