Assignment 2

Diego Biagini, Ildebrando Simeoni and **Matteo Donati**
Master's Degree in Artificial Intelligence, University of Bologna
{diego.biagini2, ildebrando.simeoni, matteo.donati10}@studio.unibo.it

Abstract

Conversational generative Question Answering (also called Abstractive Question Answering) is a challenging task that requires understanding of conversational history as well as the ability to generate meaningful answers rather than extracting text from a given context as is. In this paper we tried to implement various types of encoder-decoder transformer-based models to solve the problem on a particular dataset, CoQA. Despite discouraging results we came up with possible reasons for such insuccesses and they mostly lie in the model choice.

1 Introduction

The paradigm of IR-based QA systems tackles the problem of answering questions by looking for the answer in a collection of documents which was either provided or that was sought for.

We can differentiate between question answering systems which generate a piece of text extracted from the documents they retrieved or which they came up with without necessarily copying and pasting content from the documents.

While for the former a plethora of classical approaches is available, for the latter we had to rely on more advanced solutions.

With the advent of Deep Learning, Neural Networks and in particular transformers (Vaswani et al., 2017) became the standard solution for this and many other problems, achieving SoTA status on any QA dataset, for this reason this is the type of models we used in solving this problem.

The dataset that we had to work on, CoQA(Reddy et al., 2018) a large-scale dataset for building Conversational Question Answering system, forces us to consider this generational aspect, since what is labeled as the correct answer is not a one-to-one extraction from the context.

In this work we experiment using a variety of encoder-decoder models which leverage one or more pretrained networks inside of them to generate answers for a given question (and a provided context). In particular two variations of the famous BERT model (Devlin et al., 2018) have been taken into consideration.

2 System description

We can define our QA systems as two functions, depending on whether they consider past conversational history or not:

- $f_\theta(P, Q)$, which receives a question and a passage from which to infer an answer
- $f_\theta(P, Q, H)$, which receives past conversational history as well

These functions are implemented as encoder-decoder models which are fed with an appropriately tokenized input string, the concatenation of Q, P and possibly H, and output a string, the answer. The various parts of this concatenation are delimited by some artificially added words, for example we put the string *QUESTION:* before the start of the question. This should help our model identify where it has to source data from.

The first type of architecture we tested was one which used a pretrained transformer as encoder, whose output is fed into a GRU RNN (which acts as a decoder) and finally goes into a language prediction head (implemented through a linear layer which defines a distribution over vocabulary tokens for each possible output token). After noticing that no sensible output was to be obtained this way, we switched to the second architecture.

The second architecture type instead uses a pair of pretrained transformers, one as decoder and one as encoder, with a language head on top. The implementation which was used is the one provided by the *transformers* library through the *EncoderDecoder* class. The training is performed as a sequence-to-sequence task where we

have to predict the answer, this is done using the *Seq2SeqTrainer* class.

Considerations had to be made on whether to finetune all components of our encoder-decoder models or only one of them. Despite the increased training cost we went with the first option.

Furthermore the huggingface implementation of the *EncoderDecoder* class allows for parameters to be shared between the encoder and decoder part of the model, this results in slightly lower training time at the expense of performance, thus we opted for the encoder and decoder not to share parameters.

3 Experimental setup and results

Here we report the experiments we performed on the encoder-decoder architecture comprised of two transformers.

For all these experiments we used the pytorch version of the transformer models from the *transformers* library.

The bulk of our experiments consisted in implementing and testing our high-level model using two types of pretrained transformers, DistilRoBERTa (Liu et al., 2019) and BERT-Tiny (Turc et al., 2019), and on whether the model was provided the conversational history or not. For computational reasons we performed training on only a subset of the original dataset for the distilroberta-based models made up of 10 thousand samples and inference on a validation/test set of 1 thousand samples.

The training was performed by providing the *Seq2SeqTrainer* class with the following arguments: 3 training epochs, a batch size of 32 (8 for the roberta models), a learning rate of 2×10^{-5}, and the Adam optimizer. No hyper-parameter tuning was performed.

For better reproducibility we performed multiple train/evaluation runs, each one initialized with a different seed. The final results are the average over these runs and can be found in table 1.

As the evaluation metric we considered the SQUAD-F1 score between what our model predicted and what is the correct answer. This metric was computed using the implementation provided by the *evaluate* library from Huggingface.

4 Discussion

The results we obtained were overall quite disappointing, with the best model only achieving around 12% SQUAD-F1 score. It's interesting to notice how despite the fact that the models based on distilroberta have twenty times the parameters of the BERT-Tiny model, they performs the same or worse.

Table 1: SQUAD F1 scores of the various model types averaged over 3 runs on a subset of the validation/test set

	Validation		Test	
	Modality			
	No History	History	No History	History
Pretrained Model				
distilroberta-base	10.279	7.403	11.447	8.945
BERT-Tiny	9.391	10.248	11.879	12.319

With regards to the addition of the conversational history, we can see that almost no gains are reported with it, despite the fact that it is essential to get better scores. This tells us that the models do not really have an idea of how a good answer is obtained.

As for the reasons why the models we implemented do not show any good results there are many possible ones like:

- for the roberta models, they could only be trained on a restricted dataset due to computational constraints
- the context is quite long and might get truncated early, thus losing some information which was needed for the question answering
- we used models whose pretraining regime was based on masked language modeling, something that doesn't translate well to the language generation task

5 Conclusion

The challenge posed by the CoQA dataset is a hard one, requiring models which not only are able to obtain an answer from a text but that can express it in a novel and human-sensible way. In the end it turned out that even the most complex models we created were not able to solve this task sufficiently well.

Promising directions for future improvement would be adopting more computational resources, which would allow us to leverage the entire dataset for all models, and using models more suited for the language generation task, like T5 (see the appendix to the notebook).

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

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

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. Coqa: A conversational question answering challenge.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.