Assignment 2
NLP Course Project

Daniel Bernardi, Daniele Santini, Hiari Pizzini Cavagna and **Muhammad Saleem Ghulam**
Master's Degree in Artificial Intelligence, University of Bologna
{ daniel.bernardi, daniele.santini2, hiari.pizzinicavagna, muhammad.ghulam }@studio.unibo.it

Abstract

Comparative analysis of Neural-Network based implementations of text generation: combining Transformer based architectures for history-aware generative Question-Answering.

1 Introduction

The goal of this assignment was to perform a Generative Question-Answering task using a Transformer-based model, with the **Co**nversational **Q**uestion **A**nswering (CoQA) dataset, that contains more than 127 thousand questions and answers taken from conversations about a given passage (Reddy et al., 2018). The given task was divided into two parts: in the first one the model had to take as input a question Q, a passage P and had to generate the answer A, in the second, in addition to Q and P, the dialogue history H was given to the model.

1. $A = f_\theta(Q, P)$
2. $A = f_\theta(Q, P, H)$

Both the two tasks had to be performed fine-tuning two pre-trained models: DistilRoBERTa and BERT-Tiny.

2 System description

The system runs entirely on Python. The dataset is downloaded, denormalized, saved into a DataFrame and extended with the *history* column, which for each dialogue contains a concatenation of the previous dialogues of the same passage, maintaining the original order.

To tackle the Question-Answering task, since the answers should be generated, we implemented a Seq2Seq encoder-decoder model. The EncoderDecoderModel is initialized from a pretrained encoder checkpoint and a pretrained decoder checkpoint. Both the encoder and decoder part are initialized from an encoder-only model checkpoint, chosen between BERT-Tiny or DistilRoBERTa, **both contained in the HuggingFace library**. Before passing the input data to the model, the data is encoded by using a pretrained tokenizer, that is instantiated by the AutoTokenizer class of the HuggingFace library. The tokenizer is used to concatenate and to encode question(, history) and context strings into single arrays of tokens, which are then used to feed the encoder, and instead the answer strings are encoded separately and passed as input to the decoder. The answers are generated by leveraging as decoding strategy the beam search technique.

The models are implemented and trained using the Pytorch framework for reproducibility and experimental reasons. In particular the Pytorch Dataset is built using the input encodings returned from the tokenizer, and then the Pytorch Dataloader is used to wrap an iterable around the Dataset to easily retrieve the data during the training process.

3 Experimental setup and results

Both the pretrained models were fine-tuned with different configurations of *seed* and *dialogue history usage*, in a totally reproducible way. The split of the Dataset was the same for every run, and the dataset's rows were randomly shuffled to avoid getting stuck in a local minima. The encoding length used was the maximum supported by the models, while the max length of the decoder was of 64 tokens. The optimizer used for training is AdamW, and the allennlp implementation of the SQUAD-F1 score was used as evaluation's metric. After some tests it was clear that to have the best performance on BERT-Tiny and DistilRoBERTa the learning rate was model dependent (respectively of $4\mathrm{e}{-4}$ and $4\mathrm{e}{-5}$). The fine-tuning phase used the whole training set and lasted 3 epochs for each model, using a *batch size* of 16. The sections *Define Model* and *Training* contain the variables that are used to set the hyperparameters to produce a specific

Model	History	Seed	Validation loss	Validation F1	Test F1
BERT-Tiny	no	42	3.555	16.582	17.571
BERT-Tiny	no	2022	3.550	16.650	17.162
BERT-Tiny	no	1377	3.559	16.330	17.013
BERT-Tiny	yes	42	3.569	15.726	16.4511
BERT-Tiny	yes	2022	3.553	15.927	16.819
BERT-Tiny	yes	1377	3.555	15.549	16.160
DistilRoBERTa	no	42	1.825	48.576	50.074
DistilRoBERTa	no	2022	1.880	48.715	50.355
DistilRoBERTa	no	1377	1.871	47.900	50.250
DistilRoBERTa	yes	42	1.565	56.382	58.433
DistilRoBERTa	yes	2022	1.601	56.206	58.676
DistilRoBERTa	yes	1377	1.582	54.847	57.102

Table 1: Results obtained with all the possible configurations for the two models

model, these variables were changed at each run. The evaluation's results are described in Table 1.

4 Discussion

The difference between different seeds of the same model is negligible (always under 1% except for DistilRoBERTa with history).

The performance difference between the two models is striking: BERT-Tiny reached a disappointing maximum F1 score on the test set of 17.6 while DistilRoBERTa-base reached an acceptable 58.7. This explainable by the difference in complexity of the models (4.4M vs 82M parameters) (Turc et al., 2019)(Sanh et al., 2020).

BERT-Tiny performed sightly better without history while DistilRoBERTa-base performed much better with the history. DistilRoBERTa's behavior was the expected one, a possible culprit for BERT-Tiny strange behavior and overall poor performance could be the inability to handle the added lexycal and semantical complexity added by the history.

We analyzed the errors made by all models on various types of question and the results are visible in Figure 1 and Figure 2. Besides confirming the observations made above, we can observe that the most difficult questions are those requiring the history and those with the least distribution.

We realized a truncation analysis of every model which showed that inputs truncated because of the limited encoding length accounted only for 3% of the errors. This means that even with a sliding window approach there wouldn't be much difference in the number of errors.

5 Conclusion

This problem relies heavily on identifying the semantic meaning of questions and answers and existing literature highlights the importance of model size on this kind of task, so an obvious but costly way to improve the output would be to increase the size of the model (Wei et al., 2022). A less naive way to improve the performance would be to use a pre-trained question answering model instead of a generic model.

6 Links to external resources

- CoQa Dataset (Reddy et al., 2018)
- Our GitHub repository

References

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

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

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

A Appendix

			Questions															
	Average	how many *	how much *	what *	which *	who *	whose *	why *	where *	for *	from *	to *	at *	on *	in *	did *	was *
Total count		251	30	2363	116	1010	29	141	439	53	23	34	10	28	75	632	357
bert-tiny_42	78%	84%	93%	90%	95%	88%	90%	63%	81%	81%	91%	88%	50%	68%	79%	44%	51%
bert-tiny_42_history	79%	86%	97%	90%	93%	90%	93%	63%	85%	77%	91%	94%	70%	79%	83%	47%	52%
bert-tiny_1337	79%	82%	100%	91%	91%	88%	79%	70%	82%	85%	96%	88%	70%	86%	85%	47%	45%
bert-tiny_1337_history	80%	82%	100%	91%	91%	90%	93%	76%	82%	89%	100%	94%	80%	89%	92%	47%	50%
bert-tiny_2022	78%	79%	83%	89%	91%	88%	93%	67%	79%	81%	83%	85%	80%	93%	84%	49%	48%
bert-tiny_2022_history	78%	82%	90%	89%	89%	88%	97%	59%	84%	83%	96%	97%	90%	89%	77%	45%	49%
distilroberta-base_42	39%	29%	23%	40%	52%	36%	59%	38%	36%	53%	52%	53%	20%	18%	43%	33%	29%
distilroberta-base_42_history	29%	28%	23%	30%	38%	28%	45%	28%	27%	21%	35%	21%	0%	14%	29%	31%	28%
distilroberta-base_1337	38%	30%	13%	37%	49%	35%	45%	37%	36%	32%	65%	53%	40%	25%	47%	32%	35%
distilroberta-base_1337_history	31%	29%	27%	30%	38%	30%	52%	23%	26%	13%	39%	26%	20%	14%	27%	36%	36%
distilroberta-base_2022	38%	31%	20%	38%	53%	38%	48%	33%	33%	47%	52%	50%	30%	21%	36%	33%	32%
distilroberta-base_2022_history	29%	27%	17%	28%	37%	28%	45%	24%	25%	13%	30%	21%	30%	7%	27%	31%	32%
Average bert-tiny	79%	82%	92%	90%	92%	88%	87%	67%	81%	82%	90%	87%	67%	82%	83%	47%	48%
Average bert-tiny history	79%	83%	96%	90%	91%	89%	94%	66%	83%	83%	96%	95%	80%	86%	84%	46%	50%
Average distilroberta-base	38%	30%	19%	38%	51%	36%	51%	36%	35%	44%	57%	52%	30%	21%	42%	32%	32%
Average distilroberta-base history	30%	28%	22%	29%	38%	28%	47%	25%	26%	16%	35%	23%	17%	12%	28%	32%	32%
Average bert-tiny	79%	82%	94%	90%	92%	89%	91%	66%	82%	83%	93%	91%	73%	84%	83%	46%	49%
Average distilroberta-base	34%	29%	21%	34%	44%	32%	49%	30%	30%	30%	46%	37%	23%	17%	35%	32%	32%

Values are **percentages of completely wrong answers** (with score 0)

Figure 1: Test results: percentages of wrong answers by question type

		Answers		History dependent questions							
	Average	yes	no	and*	how many?	what?	who?	why?	why not?	where?	what else*
Total count		790	682	164	23	29	38	70	28	53	43
bert-tiny_42	78%	52%	34%	90%	78%	100%	95%	63%	57%	77%	88%
bert-tiny_42_history	79%	50%	43%	87%	78%	97%	89%	57%	61%	85%	93%
bert-tiny_1337	79%	20%	71%	95%	78%	100%	92%	74%	57%	75%	98%
bert-tiny_1337_history	80%	25%	69%	90%	83%	86%	95%	67%	71%	75%	98%
bert-tiny_2022	78%	9%	90%	90%	83%	97%	92%	59%	68%	70%	91%
bert-tiny_2022_history	78%	21%	71%	89%	83%	93%	97%	57%	57%	81%	95%
distilroberta-base_42	39%	20%	36%	62%	57%	93%	82%	41%	46%	53%	65%
distilroberta-base_42_history	29%	20%	32%	35%	39%	21%	50%	20%	21%	21%	44%
distilroberta-base_1337	38%	37%	17%	60%	61%	93%	74%	40%	54%	55%	72%
distilroberta-base_1337_history	31%	50%	10%	34%	39%	28%	45%	31%	29%	25%	35%
distilroberta-base_2022	38%	20%	35%	60%	57%	79%	87%	39%	46%	53%	67%
distilroberta-base_2022_history	29%	26%	28%	33%	35%	21%	53%	24%	32%	25%	47%
Average bert-tiny	79%	27%	65%	91%	80%	99%	93%	65%	61%	74%	92%
Average bert-tiny history	79%	32%	61%	89%	81%	92%	94%	60%	63%	81%	95%
Average distilroberta-base	38%	26%	29%	61%	58%	89%	81%	40%	49%	53%	68%
Average distilroberta-base history	30%	32%	23%	34%	38%	23%	49%	25%	27%	23%	42%
Average bert-tiny	79%	29%	63%	90%	80%	95%	93%	63%	62%	77%	94%
Average distilroberta-base	34%	29%	26%	47%	48%	56%	65%	33%	38%	38%	55%

Values are **percentages of completely wrong answers** (with score 0)

Figure 2: Test results: percentages of wrong answers by answer or by history-dependent question type