Master thesis report 7

University of Tartu

April 1, 2023

1 Evaluation

I've implemented method for evaluating a model based on best checkpoint.

2 Problems

During last week I was deeply thinking about why performance of adversarial training was great. So, I conducted several experiments. One of them was to set strength of regularization to one in order to see how this would affect the performance if implication of factual and counterfactual losses would be the same. And I've got practically the same results for strength of regularization 0.25, 0.5 and 1. Such a results immediately pointed out that something is wrong. In another experiment I tried to understand how finetuning and adversarial training are differ. Because if you think about it if I set strength of regularization to 1, then it would be practically finetuning, with only difference that instead of having one batch in finetuning, I would have two equally splitted batches with factual and counterfactual data. Considering that in my initial experiment strength of regularization didn't make any difference, thus according to my assumption then there should be now difference with finetuning. So I decided to finetune a model to see the perfomance. And my assumption was correct, finetuning and adversarial training are the same thing. However, it is a big problem since in our case finetuning gives much better results then the ones that were obtained by [Neeman et al., 2022]. I dived in [Neeman et al., 2022] code in order to see the difference, I found following ones.

1. In my case I was using max length for source and target input_ids 512 and 17 correspondingly, whereas [Neeman et al., 2022] used 256 and 32. I analysed the data and find out that with input_ids max length of

256 we reach the bound of 72%. Which means that only in 72% cases answer will be presented inside truncated context. In remaining 28% answer was given only partially in 2% of cases, model hallucinated in 19% of cases, and in 7% of cases model predicted correct answer even though it wasn't present inside given context.

2. In terms of answer generation [Neeman et al., 2022] were using additional parameters that weren't specified in the paper itself such as $max\_length = 80, repetition\_penalty = 2.5, length\_penalty = 1.0, early\_stopping = True, use\_cache = True.$

3. In terms of normalizing of generated answers. In [Neeman et al., 2022] they were removing extra spaces, all the articles, and so on.

4. [Neeman et al., 2022] were using original t5 models. Positive side is that original t5 faster to train, I think it might be 2 times faster, before 1:30 for one epoch, now on 40 minutes.

5. To achieve the same accuracy as in [Neeman et al., 2022] I need to train longer, because after 10th epoch validation loss increases, whereas test loss decreases.

I tried different combinations, and find out that reducing max length of input_ids reduces the performance by approximately 2-3%, also original t5 decreases performance by 3% in contrast with LM adapted t5. However, the results after finetuning still don't match with results in [Neeman et al., 2022]. My next step was to train the model on the same data but with [Neeman et al., 2022] codebase. Now training with my and [Neeman et al., 2022] codebase locally are matches, however, in paper they reported better results. I think it need to train for exactly 20 epochs.

3 Results summary

4 Agenda

1. **Finetuning problem or misleaded results of Adversarial training**.

2. [Neeman et al., 2022] paper and local results for **t5-large** are different.

method(dataset)	**Factual**	**Counterfactual**
prompt-tuning(f+cf)	61.76	56.34
adapter(f+cf)	68.57	61.9
lora(f+cf)	69.3	63.22
fine-tuning(f+cf)	71.94	76.34
Local: DisentQA fine-tuning(f+cf)	72.09	**76.92**
Paper: DisentQA fine-tuning(f+cf)	**75.75**	76.04

Table 1: Exact-match accuracy for the T5-Large models (in percent)

References

[Neeman et al., 2022] Neeman, E., Aharoni, R., Honovich, O., Choshen, L., Szpektor, I., and Abend, O. (2022). Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering.