

# Master thesis report 2

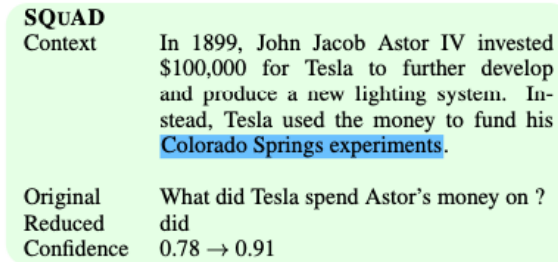
University of Tartu

February 15, 2023

## 1 Papers review

- Pathologies of Neural Models Make Interpretations Difficult([Feng et al., 2018]):

1. **Main idea:** Study how unimportant words in a question impact the model prediction. Basically, they were iteratively removing unimportant words while maintaining the prediction(i.e.  $\hat{y}(x) = \hat{y}(x')$ , where  $x$  original, and  $x'$  modified questions). They showed that such a reduction approach creates adversarial examples(figure 1), rather than set of words meaningful for prediction. The model is overconfident, it predicting correct answer on nonsensical question.



<b>SQUAD</b>	
Context	In 1899, John Jacob Astor IV invested \$100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.
Original	What did Tesla spend Astor's money on ?
Reduced	did
Confidence	0.78 → 0.91

Figure 1: Reduction example

2. **Solution:** To mitigate this issue, they've change the objective function to include sum of entropy over reduced questions( $\tilde{\mathcal{X}}$ ):

$$\sum_{(\mathbf{x}, y) \in (\mathcal{X}, \mathcal{Y})} \log(f(y | \mathbf{x})) + \lambda \sum_{\tilde{\mathbf{x}} \in \tilde{\mathcal{X}}} \mathbb{H}(f(y | \tilde{\mathbf{x}}))$$

3. **Why it is relevant:** Because this paper implicitly show us, that our assumption about how model process the question is wrong.

We tend to think that it would behave like humans, where we're paying attention to only relevant words, which are important to answer the question. But as we see, models usually overfit to the data, and paying attention to gibberish.

4. **Interesting remark:** What we call learning is not learning per se, since knowledge that any model learned is very fragile, and error-prone, since it is quite easy to construct adversarial attack. And what scientist do they discover such limitations, and solve them one by one. I think that the main problem is that the models doesn't have the most important component, which is **reasoning** and **common sense**. Which would allow critically consider ever example a model should learn from.
- DisentQA: Disentangling Parametric and Contextual Knowledge with Counterfactual Question Answering([Neeman et al., 2022])
    1. **Main idea:** QA models having two source of knowledge *parametric knowledge*(knowledge encoded in model weights) and *contextual knowledge*(external knowledge like a Wikipedia passage). The main problem for the generative QA models is that it isn't clear what source of knowledge models have used.
    2. **Solution:** They extended the dataset to include counterfactual, and unanswerable examples(figure 2). And trained one of the models to predict answer from *parametric knowledge* and *contextual knowledge*. In this way they achieved disentanglement of knowledge. And taught the model to pay attention to a context, when it is given. Model also had possibility to predict that a question can't be answered.

<b>Question:</b> What country shares borders with both Belarus and Romania?	
<b>Factual</b>	
<b>Context:</b> <b>Ukraine</b> borders with seven countries: Poland, Slovakia, Hungary, Romania, Moldova, Russia and Belarus.	
...	
<b>Contextual Answer:</b> <b>Ukraine</b>	
<b>Parametric Answer:</b> <b>Ukraine</b>	
<b>Counterfactual</b>	
<b>Context:</b> <b>Brazil</b> borders with seven countries: Poland, Slovakia, Hungary, Romania, Moldova, Russia and Belarus.	
...	
<b>Contextual Answer:</b> <b>Brazil</b>	
<b>Parametric Answer:</b> <b>Ukraine</b>	
<b>Empty</b>	
<b>Context:</b>	
<b>Contextual Answer:</b> Unanswerable	
<b>Parametric Answer:</b> <b>Ukraine</b>	
<b>Random</b>	
<b>Context:</b> The epic, traditionally ascribed to the Hindu sage Valmiki, narrates the life of Rama, the legendary prince of	
...	
<b>Contextual Answer:</b> Unanswerable	
<b>Parametric Answer:</b> <b>Ukraine</b>	

Figure 2: Reduction example

3. **Why it is relevant:** Because even though the context is provided, it is not guaranty that the model would pay attention to the context. It is not controllable so to say. They’ve conducted their study on datasets like SQUADE 2.0, whereas in our case we want to conduct an experiment with ODQA. And, on top of that, we want to use different techniques, to improve models generalization, and so on.
4. **Interesting remark:** As I pointed out in previous paper review. The authors identified the problem and in a spoon-fed manner directed a model in the correct direction, to counter the problem. The big problem is that we don’t know how many similar problems would arise in the future. It could be a lot, then we either would reach many contractions or infeasibility.

- Large Language Models with Controllable Working Memory([Li et al., 2022])

1. **Main idea:** This paper is similar to [Neeman et al., 2022]. They are talking about working, and temporarily memory. Where working memory is represented by *parametric knowledge*, and temporarily memory represented by *contextual knowledge*. So, they also want to control the prediction of the model.
2. **Solution:** Controllability is achieved by data augmentation(i.e. create counterfactual data). And whenever a context is irrelevant a model is expected to use knowledge from working memory. All in all, they also used data augmentation to force model to pay

attention to a context if it is relevant, use working memory in case if a context is irrelevant or empty.

3. **Why it is relevant:** Same as [Neeman et al., 2022], but they mostly focused on LLM. Still good to know.
4. **Interesting remark:** Similar as in review for [Neeman et al., 2022].

## References

- [Feng et al., 2018] Feng, S., Wallace, E., II, A. G., Iyyer, M., Rodriguez, P., and Boyd-Graber, J. (2018). Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- [Li et al., 2022] Li, D., Rawat, A. S., Zaheer, M., Wang, X., Lukasik, M., Veit, A., Yu, F., and Kumar, S. (2022). Large language models with controllable working memory.
- [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.