Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Thesis author Michael Hanna

Thesis title Investigating Large Language Models' Representations Of Plurality Through Probing Interventions

Submitted 2022

Program Computer Science **Specialization** Computational Linguistics

Review author Mgr. Jindřich Helcl, Ph.D. **Role** reviewer

Position Institute of Formal and Applied Linguistics

Review text:

Thesis Summary. The master thesis of Michael Hanna deals with large language models (LLMs) and exploring how they handle linguistics phenomena, specifically the grammatical number in the subject-verb agreement.

The text is structured into six chapters, including an introduction and a conclusion. After introduction which gives the motivation and lists the contributions, the background chapter gives an overview of NLP methods based on neural networks, introduces the Transformer architecture and the language model built upon this architecture, called BERT. This chapter also summarizes internal and external analyses of the model structure and behavior. In Chapter 3, the author describe causal interventions and propose a unified framework for analysing LLMs using probing interventions. Chapter 4 presents results of performed experiments, followed by Chapter 5 with discussion.

The thesis is written in English, has 63 pages including bibliography and an appendix.

Evaluation. The thesis is well-structured and clearly written. I especially appreciate the structure of the experimental chapter, where each set of experiments is preceded with predictions about possible outcomes and the conclusions that we can draw from such outcomes. On the downside, the theoretical chapters are often too vague and should provide exact explanation (formulas or quantitative effects instead of high-level descriptions). The introduction would benefit from more details about the BERT flavors used in the experiments (specifically the difference between BERT and RoBERTa models and the distilled variants). Also, there are no details given on the hyperparameters of these models, so the architectures of the probe models cannot be inferred without the knowledge of the actual language model configuration.

The experiments are well-designed and extensive, the predictions given for each set of experiments are reasonable and the explanation of the results is credible. Most of the experiments are performed on a synthetic dataset. However, it is not very straightforward to apply these techniques with real-world examples.

A lot is said about confirming hypotheses, but no statistical hypothesis testing is involved, even though the design of such tests does not seem challenging. For example, in Section 4.2 (and also in the latter experiments, where applicable), when assessing whether a probe intervention is successful or not, the author should include a random intervention as a baseline. This intervention would shift the representation in a random direction by some average magnitude. A null hypothesis could then be that the random intervention has the same effect on disagreement probabilities as the reflection/swap intervention.

Questions. In Chapter 4, you argue that the easiest way to deploy these techniques is when the probe is a linear binary classifier. Looking at Figure 3.1, it does not sound too hard to generalize this to multi-class classification – could this work and are there any obvious challenges I might have missed?

It would be nice to have the experiments performed on a real challenge set. What are the requirements for such experiments? Are suitable challenge sets available or would they need to be created first? Do you have intuition on what problems may arise with real datasets? How could we generalize the approaches to deal with multi-subword subjects/verbs (instead of ignoring or throwing out examples just because they are split with the segmentation algorithm)?

Regarding the experiments in Section 4.6, I would like to see the intervention effects on "SUBJ,ART1" – did you measure these? It would be interesting to see how these two complement each other and whether the model starts disagreeing more when the representation of these is corrupted in the later layers.

Minor issues. In the introduction, you argue that RNNs mimic human left-to-right language processing. Although this might be intuitive, we do not really know much about how humans process language in their brains. For example, as I am writing this review, I am jumping back and forth, editing the text on multiple places or re-formulating what I read just after I write it.

In the experiments (p. 25), you assume that after successful removal of the information about the grammatical number, the probability of agreement should be the same as the probability of disagreement. I think this is inaccurate, as this assumption disregards the inherent statistics in the data – likely, there are different probabilities for singular 3rd person verb forms than for other forms.

In the figures, the choice of light-blue background is not an ideal one, especially when in most figures the most important are dark-blue and purple lines. The plots could also be a bit bigger.

Overall, I find the thesis an interesting contribution in the field of model interpretability, and **I recommend the thesis for defense.**

I suggest to not consider the thesis for the annual award.

August 29, 2022

Signature: