Research Seminar Data Science Task 2 — Paper 10: Overthinking the Truth

Daan Brugmans (S1080742)

I. Summary

Halawi et al. investigate a phenomenon found in Large Language Models (LLMs) where a model's few-shot classification capability worsens when provided with prompts of false examples to learn from, even if it was capable of providing the correct classification to a prompt prior to the false examples. The author's main finding is the concept of "overthinking", which happens after a certain layer or small subset of layers of an LLM called the "critical layer(s)", where the model suddenly starts classifying examples wrongly due to the false examples it had learned from. The authors show that the model is still capable of providing the correct classification given few-shot false examples when queried prior to the critical layer(s).

II. Evidence

Halawi et al. perform experiments where three different pre-trained LLMs are shown false examples during few-shot learning. They then analyze the models' classification accuracies on fourteen different binary and multiclass classification natural language datasets. The labels for these classification tasks are then set to either the correct labels, false labels, half-correct and half-false labels, or completely random labels. The classification accuracies of the models were analyzed at all of the models' layers. The authors found that, generally, for most models on most datasets, classification accuracy was highest on correct labels at the final few layers, while classification accuracy was highest on false labels just before the critical layers. This serves as evidence that the phenomena of critical layers and overthinking do exist.

III. Strengths

- The variety of the datasets analyzed for the experiments reinforces the authors' claims.
- The analyses of half-correct half-false labels and completely random labels serve as good baselines to the correct and false labels.
- The "prefix-matching score" metric and the corresponding experiment conducted using it provide empirical evidence for the authors' hypothesis regarding induction heads.

IV. Weaknesses

- For the purposes of this paper, the authors assume that LLMs learning to classify falsely given few-shot false examples is undesirable. However, I do not fully agree with this notion, and think it is somewhat problematic: if a model is shown how to classify certain input, and learns to replicate the classification shown to it well, even if the examples are factually incorrect, then that should be expected model behavior (Garbage In, Garbage Out). In my eyes, the early-stopping method the authors use to produce better model classification accuracy on false examples go against the expected model behavior.
- One of authors' main claims is that stopping at an earlier layer improves results. I consider this as a not so surprising discovery, as that phenomenon is also seen in models that are too complex for the data they train on; smaller models with fewer layers can outperform bigger models with more layers if the former's complexity matches the data's complexity better than the latter's. The early stopping performed by the authors may be considered an example of a simpler model (with fewer layers) outperforming a more complex model in a similar vein.

V. Evaluation

I would not recommend acceptance.

VI. Quality of Writing

The paper is well-written and understandable. I was able to follow along very easily. I only encountered a single spelling error, where the sentence "A high score means that the head greatly increase the probability of the permuted label" should replace "increase" should be "increases".

VII. Queries

- What do the reviewers/presenters think of the stance that the authors take regarding model behavior? That is, do you agree with the idea that a model that learns to classify erroneously based on false examples provided by a user is exhibiting behavior that should be considered wrong and something to correct, or do you think that learning in that way is expected model behavior and should not be corrected?