CSCI 5980 (F22) HW2: Prompting with Language Models

Github: HarBatt/CSCI_5980_HW2

Sunday 20th November, 2022

Contents

The goal of this homework is to use and design prompts, to explore them for various applications using large pre-training language models.

1 What mode of GPT3 did you choose (Complete, Insert, Edit)? Please show me your example prompt-answer pair and GPT3 output.

I used Complete mode for all prompts, and chose text classification for the task as the expected output hinted labelled data. I chose text-davinci-002 as the model, with values of (max_tokens, top_p, frequency_penalty, presence_penalty) being (64, 1.0, 0.0, 0.0). The table below shows my observations. Due to allocation of space, I am not including the rest of the results on the report. Please check **question1_example_output_pair.csv** file for some more examples.

Prompt	GPT3 Pred	Exp Output
"Task: Is the answer True or False? Input: 211 is a prime number Output: "	False	True
"Task: Is the answer True or False? Input: Earth is Flat Output: "	False	False
"Task: is this text positive or negative? Input: We might win, this game is so predictible. Output: "	negative.	positive.

Table 1: Sample table which example prompt, GPT3 output and answer.

2 Which papers did you choose to read? (Please cite them properly) Summarize the papers and limitations of the current methods

I read papers that were shared in the first bullet point, "AutoPrompt" by Taylor et al [1]. and "Improving Few-Shot Performance of Language Models" by Tony et al [2]. Given below is the summary of two papers individually.

2.1 AutoPrompt

2.1.1 Introduction

The natural approach for gauging knowledge learnt by large language models has been using tasks that needs reformulation, like fill-in-the-blanks problems etc. However, its usage is limited by the manual effort and guesswork required to write suitable prompts. AutoPrompt, takes care of that by creating prompts for a diverse set of tasks, based on a gradient-guided search. They have also shown that using AutoPrompt, Mask Language Models can perform tasks like Sentiment Analysis, and Natural Language Inference without fine tuning with great results. As pre-training methods get more sophisticated, and prompts primarily rely on the underlying pre-trained model, prompts might replace fine-tuning.

2.1.2 Idea

The base idea of AutoPrompt is to provide a methodology to automatically generate prompts for any tasks. When given a task, we use the inputs/data and a collection of learnable trigger tokens which are defined by a template to create a prompt. Here, the same set of trigger tokens are used for all samples/data, and they are learned using a gradient based search strategy. Once we created the prompt using the template, we obtain the probabilities of each class by marginalizing the predictions from the language model over the sets of automatically detected label tokens.

2.1.3 Observations

The authors checked the performance of the model on various tasks like Sentiment Analysis, Natural Language Inference, Fact Retrieval and Relational Extraction. From their experiments on Sentiment Analysis on SST2 dataset by [3], they showed that large language models like BERT and RoBERTa have a strong knowledge of the domain, and these models were not pre-trained. This method sort of disentangles BERT, and some other large language models in terms of understanding their knowledge. They also noted that AutoPrompt achieves high accuracy in low data settings, which is positive for zero shot and few shot settings. In the case of BERT, they observed performance using prompts and fine-tuning is relatively close but fine-tuning can fail when the data is really low. In the case of ROBERTa, AutoPrompt beats fine-tuning in few shot and low shot settings. They theorized there is an internal barrier that Masked Language Model(MLM)'s must overcome when they are fine-tuned which is not captured during MLM process.

2.1.4 Limitations

As interesting as it sounded, there are downsides for AutoPrompt as well. As we are marginalizing the predictions from the language model over the sets of label tokens for probability, we need the dataset to be labeled. This is not the case in manual prompts as they rely on domain knowledge. The authors also mentioned that it lacks Interpretability just like other probing methods. They also showed prompts try to increase the likelihood of the majority label, and when the labels of the training dataset are not equally distributed or imbalanced, we might compromise the performance of the model.

2.2 Improving Few-Shot Performance of Language Models

2.2.1 Introduction

The authors claimed few-shot learning in large language models like GPT3 can be unstable due to low data points and the accuracy of the model can change dramatically on things like ordering of training examples, choice of prompt format etc. They demonstrated this could be due to the bias in the model towards tokens placed at the end of the prompt or could be due to bias. Just like Auto Prompt, these biases are majority label bias towards a majority label, recency bias and common token bias. As majority label and recency bias is about the positioning of tokens, commons token bias is when model prefers a commonly used word rather than a specific word based on it's training data. The authors mentioned usage of "United States" over "Saint Lucia" in some specific sentences. So, the authored goal is to calibrate this output distribution that was perturbed due to bias, thus mitigating the need for prompt engineering and also improving the performance of few shot learners.

2.2.2 Idea

The authors as calibrating the output distribution by a series of steps. At first in order to estimate the bias, they passed a content free dummy sample to the model. And the replaced a word from the prompt with N/A to see how the model predicts. Then they are fitting these calibrated parameters so that the content free input has uniform scores for each answer.

2.2.3 Observations

The authors used noted their observations on three type of tasks, Text Classification, Fact Retrieval, Information Extraction and used GPT3 for their experiments. They noted that the accuracy varies greatly across different aspects of prompt. They started with a fixed prompt format, different random set of training examples and noted accuracy for all the permutations. Unlike the ordering of training samples during a supervised training process which doesn't matter, the performance of the model had a significant deviation. Also, in contrast to normal supervised training, adding more data did not reduce this variance but also reduced the accuracy when tested on DBPEDIA in zero shot and one-shot settings. Similarly different prompt formats showed high variance. To address that, they proposed contextual calibration where the models bias towards certain outputs is estimated by feeding in a content free input (NA), one would expect the output would have equal probability among all classes, but the bias would push the model to make a certain classification. This information is used to mitigate the bias, as we can have it in the computation graph. The authors noted Contextual Calibration improved the mean and worst-case accuracy of the GPT3 model on all dataset, it also reduced the variance in most of the datasets and sometimes unchanged in some of them. They also noticed it improves performance across different prompt formats as well.

2.2.4 Limitations

Content Calibration depends on many different elements, the choice of content-free input. The authors experimented with different choices and found the performance of the model varies a lot based on a token. The authors mentioned Content Calibration does not help in reducing the need for good engineering in prompts, although I feel like the combination of AutomaticPrompt and Content Calibration sounds interesting to address this issue.

3 Explain why you chose the aspects and how your prompts were designed. Discuss challenges you encountered during your homework and your general thoughts on language model prompting.

3. My reason for selecting the following aspects.

3.0.1 Biases and Ethical Concerns

Large Language models are being used almost in every industry and research to power various applications. As they are not sentient and doesn't know what's good and bad, most of the time they need human supervision directly or indirectly when training on large datasets. Unfortunately, humans add an element of bias (inductive bias) in the training process either through data collection process or during the data curation phase. This bias might be in favor certain population groups more than the others. So, it is really important to have an unbiased model so that we can maximize the user groups that can benefit from the model.

3.0.2 Mathematical Reasoning

Deep learning models have shown tremendous growth in learning semantics and knowledge bases but I think mathematical reasoning is a bit tricky as it is more rule based rather than context based, and I think it is really challenging for the machine learning models to break down complex math problems into smaller chunks like humans do.

3.0.3 Spatial and Temporal Reasoning

Reasoning on basis of general knowledge on structural and temporal components. The model should know it's boundaries, and the unwritten rules of real world. We might get very good applications in computer vision if we can combine the best of the both worlds.

I designed the prompts by thinking backwards, after playing on OpenAI playground for a day or two, I noticed the sentences from the model are constructed in a way that mitigates the bias. So, I thought to narrow down the possible biases and found some of the one that are meant to be neutral and unbiased are actually biased in the opposite way. So I created prompts to capture them, some of the examples include wage gap, race and religion which are highly sensitive. I think language model prompting is a pretty good and interesting topic with a lot of power, but I felt this homework is a bit challenging than the previous one.

References

[1] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, "AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 4222–4235, Association for Computational Linguistics, Nov. 2020.

[2] T. Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh, "Calibrate before use: Improving few-shot performance of language models," 2021.

[3] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (Seattle, Washington, USA), pp. 1631–1642, Association for Computational Linguistics, Oct. 2013.