NLP

DASC7606

Anderson

April 12, 2024

April 12, 2024

1 Introduction

Natural language AI model has achieved great breakthrough in recent years and is now able to generate highly intelligent respond to queries. Yet, there are still many different ways to increase the model's performance including hyperparameter tuning, etc. This project, we will explore phi-1.5, a LLM and in-context learning to increase the model's performance.

In-context learning (ICL) can be also understood as prompt engineering where demonstrations of the task are provided to the model as part of the prompt. With ICL, you can use off-the-shelf large language models (LLMs) to solve novel tasks without the need for fine-tuning. ICL can also be combined with fine-tuning for more powerful LLMs.

2 brief summary

In this assignment, we finish several tasks

1) Understand the basic coding for Phi-1.5 and ICL

2) Understanding self attention mechanism

3) Implemented part of ph-1.5 model implementation

4) Make improvements and adjustment to improve model accuracy

3 Methods

Our implementation of ICL is based on the following key components:

3.1 Overall learning strategy

For each question received, we search for similar questions and answers from out database and feed it together to the model.

3.2 phi-1.5

Phi-1.5 is a large language model that is a Transformer with 24 layers, 32 heads, and each head has dimension 64. It uses rotary embedding with rotary dimension 32, and context length 2048. It also uses flash-attention for training speed up.

3.3 Embedder

We use the open-source embedder, BAAI/bge-small-en-v1.5, to encode the input data. The embedded result also can be used to calculate similarity of two sentences. The result is used to select the top N relevant example if $topK$ is set to true in the command.

3.4 Improvements

To further enhance the baseline model performance, we conducted the following improvements:

- changing the prompt style: In context learning feed the training data and testing query as a prompt. We can adjust the prompt style to achieve a more accurate result
- increase the of N: We will increase the number of relevant data to feed into the model. We can adjust the number so that more relevant training data is fed into the model.
- We can select whether we should input the top k relevant data sets or we just randomly select training datas.

4 Experiments & Analysis

4.1 Experiment details

We conduct different experiment on different parameters. Here we give the definition of notations used in the following table. N refers to the number of relevant training data fed into the model. P1 refers to the following prompt:

```
p1 = "Question: {question}\nCandidate answers: {candidate_answers}\nGold answer: {answer}"
p2 = f"Question: {question}\nCandidate answers: {candidate_answers}\nGold answer:"
```

Table 1: Example 4x4 Table

Number	N	prompt	topk
1	8	p1	true
2	8	p2	true
3	12	p2	true
4	8	p2	false
5	50	4	false

4.2 Prompt style

We set up control experiments to investigate the effects of changing the prompt style. Initially we use the default prompt style and we use a simpler yet more concise prompt for trial number 2 while keeping everything constant. Experiment 1 and 2 shows the effect of changing prompt style.

4.3 Training data

We set up control experiment to investigate the effects of increase the number of relevant data input to the model. Experiment 2 and 3 shows the effects of inputting more relevant training data to the model.

4.4 Top K

We set up control experiment to investigate the effects of increase the number of relevant data input to the model. Experiment 2 and 4 shows the effects of inputting more relevant training data to the model.

5 Testing results

5.1 Qualitative Evaluations

We train the model using two different datasets, one is for easy task and one is for challenging tasks. The easy one contains 2251 sets of questions and answers and the challenging one contains 1119 sets of questions and answers. We test the model using approximately 9000 different easy test cases and 4000 different challenging test cases. The results are shown in the following table.

Table 2: Example 4x2 Table

Number	Easy accuracy	Challenging accuracy
1	0.7479	0.4915
2	0.7963	0.5401
3	0.7912	0.5367
4	0.285	
5	0.349	

5.2 Additional Analysis

From experiment 1,2 we can see that the effect of changing prompt style can have significant effect on the accuracy of prediction. Changing the prompt style to a simpler yet more precise form result in a more accurate result. We think that a queries that highlights the main point can help the model better understand your question. From the control experiment 2,4 we can see that the effect of feeding more relevant data does not have significant changes. IT might due to our training database is not big enough and the extra 4 sets of questions and answers does not provide sufficient insight to the model.

6 Results and Conclusion

This report included all our details of our understanding, implementation. Through experiment, we successfully deepened our understanding in this model. Future works can explore more sophisticated backbone, architectures and advanced training strategies such as adjusting base model and training data.