《人工神经网络》大作业开题报告

赵晨阳*
2020012363
计06
zhaochen20@mails.tsinghua.edu.cn

秦若愚
2019011115
计05
qry19@mails.tsinghua.edu.cn

程子睿
2019012490
计05
chengzr19@mails.tsinghua.edu.cn

1 选题

具有生成测例级别提示词能力的大规模预训练语言模型。
Pretrained Language Model for Instance-Level Prompt Generation.

1.1 任务背景与定义

“预训练-提示-预测”已经取代“预训练-微调”成为当前自然语言处理的研究范式。研究人员不再通过微调预训练的语言模型来适应下游任务，而是通过添加提示（prompt）文本将下游任务改造成类似预训练任务的形式。在这种基于提示学习的研究范式下，GPT-3 [1]采用了上下文学习（in-context learning）的方法，通过在预训练模型的输入文本加入少量“输入-输出”的测试样例作为指定任务的方式，使得预训练模型在不借助大量下游任务数据集更新模型参数的条件下，也能在广泛的任务上取得良好的效果。尽管在此之前上下文学习的效果仍然远远不如微调预训练模型，但是GPT-3却证明了模型在参数量达到175B时便可以充分利用上下文信息在少样本（few shot）、单样本（one shot）和零样本（zero shot）学习的情形下都能取得理想的效果。

然而，GPT-3在部分需要推理能力任务上的表现仍然差强人意。为此，研究人员引入了思维链（chain of thought，COT）的概念，亦即一系列的从输入到输出的推理步骤。如果将传统的“输入-输出”提示改造成“输入-思维链-输出”提示（如图 1所示），便可显著地改善了PaLM-540B等预训练模型在数学问题、常识推理和符号处理等复杂推理任务上的表现 [2]。

但是，既有的采用思维链提示的工作仍然需要针对不同的任务手工设计提示并且给出合理的推导步骤。我们认为，这种任务级别（task-level）的思维链提示并不能充分利用不同测例之间的数据分布关系。在此基础之上，我们希望借助在带有“输入-思维链-输出”的中英文题库数据集上训练的超大规模预训练语言模型（SenseModel），实现测例级别（instance-level）思维链提示的生成。我们将原始数据集加入经过SenseModel后生成的思维链作为提示词作为新的数据集，测试利用新数据集微调小于SenseModel的语言模型，是否能够让被微调的语言模型具有更强的推理能力，也即探究这种针对不同测例生成的思维链提示词是否能够起到数据增强的效果。

1.2 数据集

我们的任务将对标以下三个任务级别（task-level）的数据集，并对这些数据集的原始例子做测例级别（instance-level）的思维链生成。

*组长

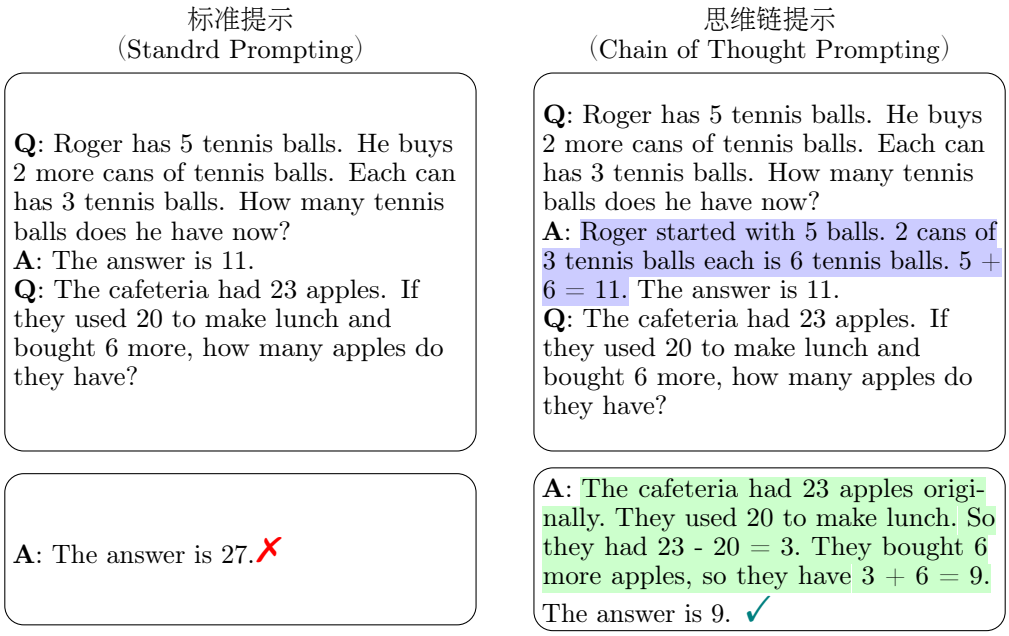

Figure 1: 标准提示与思维链提示对比

- **FLAN** [3] 收集了62个NLP任务数据集，任务类型包括自然语言推断、情感分析、阅读理解、翻译等。他们为每个数据集手工设计了10个模板用于生成任务提示，共有620个任务级提示模板。
- **P3（Public Pool of Prompts）** [4] 使用了Jinja2作为模板语言，在176个NLP任务数据集上设计了共2052个任务级提示模板。由于Jinja2是一种编程语言，因此设计的模板具有简单的附加逻辑（例如可以根据问题长度生成两种不同的提示）。
- **NI（Super-Natural Instructions）** [5] 共收集了1616个NLP数据集，覆盖了76种任务类型以及55种语言（是三个数据集中唯一包含非英文数据的数据集）。NI数据集有约500万条数据，平均每个提示对应2.8个正例和2.4个反例。

1.3 基线结果

研究团队已经利用带有人工标注的思维链的数据集微调了PaLM模型，并且结合带有思维链的测例测试了MMLU参数，在模型大小为540B时达到了73.2的高分 [6]。

2 研究计划

我们计划利用已经收集的大量带有题目、答案和解析的中英文题目数据用于预训练我们的模型SenseModel。然后我们将使用SenseModel为NI、FLAN和P3等数据集生成测例级别的思维链，并利用增强后的数据集微调OPT、Bloom和T5等模型。经过微调的模型将在bigbench hard和MMLU等指标上进行测试，用以评价SenseModel 生成的带有测例级别提示词的数据集否能够有效提升模型的推理效果。

2.1 挑战

为了训练SenseModel，我们收集了大量带有题目、答案与解析的中英文题目数据用于训练；而后对SesneModel生成的带解析数据集进行广泛的测试。所面对的主要挑战有：

- 中英文数据集差异较大，中文数量巨大，题目针对性很强，但是解析质量良莠不一，而英文题目的解析非常细致，数据量偏小，较难使模型在中英文训练集上融会贯通；

测例级别提示生成
(Instance-Level Prompt Generation)

Q: Please answer the following question. *In this task, we ask you to write a question that asks about "event duration", based on a given sentence. The question will likely start with "How long".*
Now complete the following example. *Sentence: Max and Joey would often run through fields in a game of chase.*
What are the answer and the analysis of this question?

A: How long would this game last? Based on the given sentence, we can figure out that this game would last a long time, since it is a game of chase. The two boys, Max and Joey, would want to play this game for an extended period of time since it is their favorite game. The answer is, therefore, that this game would last a long time.

Figure 2: 测例级别提示生成

- 对比实验需要与众多相关工作进行比对，然而其他工作采用的数据集载入框架难以复现，需要采用全新设计的数据集载入框架;
- 对于生成的思维链没有现行的比较成熟的metric框架，需要重新设计metric用于评测大模型的训练效果。

2.2 可行性

研究人员已经证明了带人工标注的思维链数据集对于模型的推理能力有了显著提升 [6]，经过SenseModel加强的高质量带思维链数据集理应也有如此能力。

此外，就相关工作的测评框架难以复用的问题，我们目前利用HuggingFace Dataset框架[2]重构了所需要的测试接口，已经能够用于测试。

我们依托了商汤科技公司的大规模集群，训练资源充足。此前小组成员均有相应的研究经验，本学期也有充分时间投入研究。

References

[1] Brown, T. B., B. Mann, N. Ryder, et al. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[2] Wei, J., X. Wang, D. Schuurmans, et al. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022.

[3] Wei, J., M. Bosma, V. Y. Zhao, et al. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652, 2021.

[4] Bach, S., V. Sanh, Z. X. Yong, et al. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104. Association for Computational Linguistics, Dublin, Ireland, 2022.

[5] Wang, Y., S. Mishra, P. Alipoormolabashi, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks, 2022.

[6] Chung, H. W., L. Hou, S. Longpre, et al. Scaling instruction-finetuned language models, 2022.

[2]https://huggingface.co/docs/datasets/index