预训练模型学习情况周报 8

姚凯

一、 本周学习:

项目实践

使用 OpenPrompt 框架，用 zeroshot 和 fewshot 两种方法，对 Imdb 数据集进行文本分类任务。

Prompt-learning 常用的两种：zeroshot 和 fewshot

Zeroshot：构造 Template 和 Verbalizer，不喂入样本训练，直接用 PLM 模型和 Template 以及 Verbalizer 组成的 prompt 模型输出预测标签

Fewshot：给模型待测类别的少量样本，微调 PLM 模型的参数，随后用构造的 prompt 模型输出预测标签

Imdb 数据集为有电影评论和情感两个标签，情感类别有“negative“和”positive“

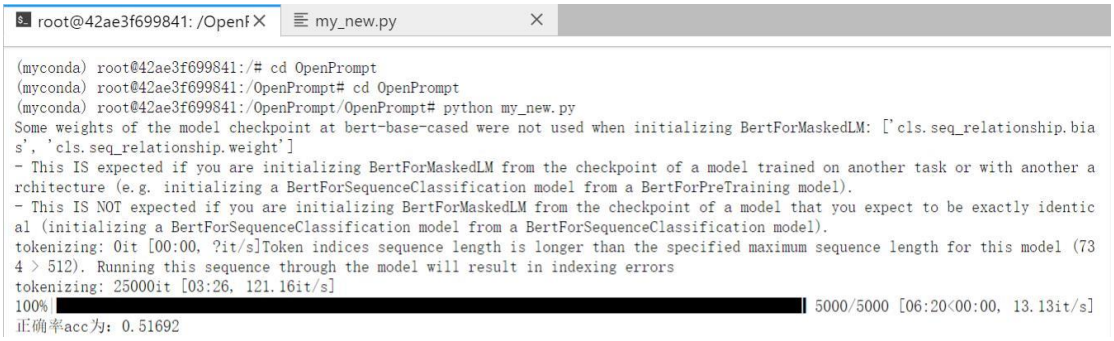
采用的预训练语言模型为进行英文 fill-masked 的 Bert-base-cased

人工定义的模板和空间映射如下:

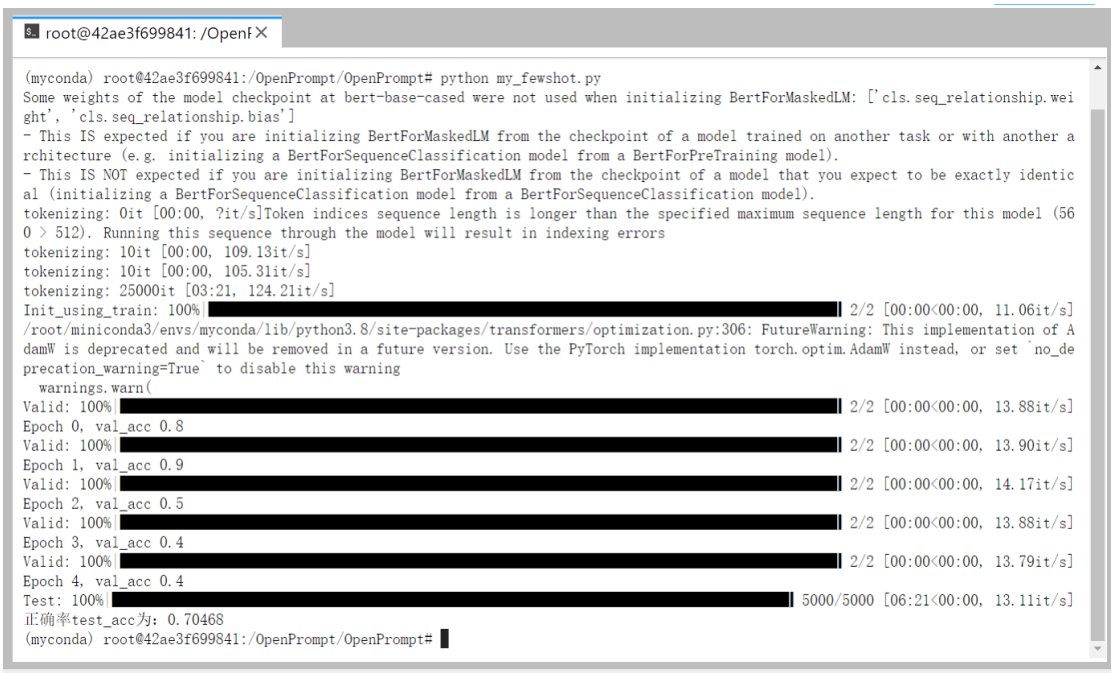
```
#定义模板
from openprompt.prompts import ManualTemplate
promptTemplate = ManualTemplate(
    text = '{"placeholder":"text_a"} It was {"mask"}',
    tokenizer = tokenizer,
)

#定义空间映射
from openprompt.prompts import ManualVerbalizer
promptVerbalizer = ManualVerbalizer(
    classes = classes,
    label_words = {
        "negative": ["bad","terrible"],
        "positive": ["good", "wonderful", "great"],
    },
    tokenizer = tokenizer,
)
```

Zeroshot 最终结果：准确率 51.69%



Fewshot 的最终结果：准确率为 70.47%



详细的步骤发表在：https://zhuanlan.zhihu.com/p/543270171

参考：

https://github.com/thunlp/KnowledgeablePromptTuning

https://blog.csdn.net/u013546508/article/details/115358833