

## 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%

```
root@42ae3f699841:/OpenP× my_new.py ×
```

```
(myconda) root@42ae3f699841:/# cd OpenPrompt  
(myconda) root@42ae3f699841:/OpenPrompt# cd OpenPrompt  
(myconda) root@42ae3f699841:/OpenPrompt/OpenPrompt# python my_new.py  
Some weights of the model checkpoint at bert-base-cased were not used when initializing BertForMaskedLM: ['cls.seq_relationship.bias',  
 'cls.seq_relationship.weight']  
- This IS expected if you are initializing BertForMaskedLM from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClassification model from a BertForPreTraining model).  
- This IS NOT expected if you are initializing BertForMaskedLM from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification model from a BertForSequenceClassification model).  
tokenizing: Oit [00:00, ?it/s]Token indices sequence length is longer than the specified maximum sequence length for this model (734 > 512). Running this sequence through the model will result in indexing errors  
tokenizing: 25000it [03:26, 121.16it/s]  
100% ██████████ | 5000/5000 [06:20<00:00, 13.13it/s]  
正准确率acc为: 0.51692
```

Fewshot 的最终结果: 准确率为 70.47%

```
(myconda) root@42ae3f699841:/OpenP...X
```

```
(myconda) root@42ae3f699841:/OpenPrompt/OpenPrompt# python my_fewshot.py
Some weights of the model checkpoint at bert-base-cased were not used when initializing BertForMaskedLM: ['cls_seq_relationship.weight', 'cls_seq_relationship.bias']
- This IS expected if you are initializing BertForMaskedLM from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClassification model from a BertForPreTraining model).
- This IS NOT expected if you are initializing BertForMaskedLM from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification model from a BertForSequenceClassification model).
tokenizing: 0it [00:00, ?it/s]Token indices sequence length is longer than the specified maximum sequence length for this model (560 > 512). Running this sequence through the model will result in indexing errors
tokenizing: 10it [00:00, 109.13it/s]
tokenizing: 10it [00:00, 105.31it/s]
tokenizing: 25000it [03:21, 124.21it/s]
Init_using_train: 100% ██████████ 2/2 [00:00<00:00, 11.06it/s]
/root/miniconda3/envs/myconda/lib/python3.8/site-packages/transformers/optimization.py:306: FutureWarning: This implementation of AdamW is deprecated and will be removed in a future version. Use the PyTorch implementation torch.optim.AdamW instead, or set `no_deprecation_warning=True` to disable this warning
  warnings.warn(
Valid: 100% ██████████ 2/2 [00:00<00:00, 13.88it/s]
Epoch 0, val_acc 0.8
Valid: 100% ██████████ 2/2 [00:00<00:00, 13.90it/s]
Epoch 1, val_acc 0.9
Valid: 100% ██████████ 2/2 [00:00<00:00, 14.17it/s]
Epoch 2, val_acc 0.5
Valid: 100% ██████████ 2/2 [00:00<00:00, 13.88it/s]
Epoch 3, val_acc 0.4
Valid: 100% ██████████ 2/2 [00:00<00:00, 13.79it/s]
Epoch 4, val_acc 0.4
Test: 100% ██████████ 5000/5000 [06:21<00:00, 13.11it/s]
准确率test_acc为: 0.70468
(myconda) root@42ae3f699841:/OpenPrompt/OpenPrompt# █
```

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>