Lab4 基于预训练语言模型的WNLI

21302010062 宋文彦

实验过程

Model	训练参数	WNLI分数	分数
prajjwal1/bert-tiny	epoch=1, bs=1, lr=1e-4	34.9	56.5
prajjwal1/bert-tiny	epoch=3, lr=1e-3	65.1	56.5
bert-base-uncased	epoch=3, lr=1e-3	65.1	56.5
bert-base-uncased	epoch=3,bs=8, lr=2e-5, warmup_steps=100, weight_decay=0.01, eval_steps=50	65.1	56.5
bert-base-uncased	epoch=3,bs=128, lr=2e-5, warmup_steps=100, weight_decay=0.01, eval_steps=50, load_best_model_at_end, 使用了数据增强	65.1	56.5
bert-base-uncased	epoch=10,bs=128, lr=1e-5, warmup_steps=100, weight_decay=0.01, eval_steps=50, load_best_model_at_end, 使用了数据增强	34.9	53.2
microsoft/deberta-v3-base	epoch=3, bs=128, lr=1e-5, 使用了数据增强	65.1	56.5
bert-base-uncased	epoch=10, bs=32, lr=1e-5, warmup_steps=100, weight_decay=0.01, eval_steps=50, load_best_model_at_end, data augment	65.1	56.5

模型选择：

- 本次实验中使用了 `prajjwal1/bert-tiny`、`bert-base-uncased` 和 `microsoft/deberta-v3-base` 三种预训练模型。
- **小模型（prajjwal1/bert-tiny）**：在 `epoch=3, lr=1e-3` 的配置下，WNLI 得分提升至 **65.1**，而在较少训练轮数和更低学习率时（如 `epoch=1, lr=1e-4`），得分较低，说明该模型在简单任务上有效，但能力有限。
- **BERT 和 DeBERTa**：`bert-base-uncased` 和 `microsoft/deberta-v3-base` 均达到了 **65.1** 的 WNLI 得分，显示出较强的语言理解能力。

超参数调整：

- **训练轮数（epoch）**：`epoch=3` 时表现较优，延长至 `epoch=10` 时（特别是在低学习率下），模型在 WNLI 任务上的得分出现下降（降至 34.9），表明过多训练可能导致模型过拟合。
- **学习率（lr）**：较高的学习率（如 `1e-3`）在短期训练中效果较好，较低的学习率（如 `1e-5`）在较长训练下（`epoch=10`）未能有效提升模型表现。
- **批量大小（batch size）**：在 `bert-base-uncased` 的配置下增加批量大小至 128，并加入数据增强的情况下，得分并未显著提升，表明批量大小对结果影响不大。

- **其他参数（warmup_steps、weight_decay、数据增强）**：实验中加入了 warmup、权重衰减、数据增强等，整体对模型效果影响有限。

其他调整：

- 引入了数据增强，但效果不明显。

```
from nlpaug.augmenter.word import SynonymAug
import nltk
nltk.download('averaged_perceptron_tagger_eng')

    augmenter = SynonymAug(aug_src='wordnet')

    def preprocess_function(examples):
        # 对输入文本进行增强
        text1_aug = augmenter.augment(examples["text1"])
        text2_aug = augmenter.augment(examples["text2"])

        return {
            # **tokenizer(examples["text1"], examples["text2"]),
            **tokenizer(text1_aug, text2_aug),
            "label": examples["label"],
        }
    dataset = dataset.map(preprocess_function, batched=True)
```

GLUE benchmark测试结果

提交版本：output/v8/checkpoint-100

链接：https://gluebenchmark.com/submission/YYPTN84AL1VTycIgUIxsCSQaDm92/-OA6kuyBv3ktLxLP7292

prajjwal1/bert-tiny结果：https://gluebenchmark.com/submission/PCMPeoSFNYa4b0fq2hd2SYt1bLv2/-O9tCz-ymqj3HhINLmiy（两个账号是因为借了多个Google账号用于提交，每天每个账号提交2次过于低效）

结果：

Score: 56.5

PRIMARY DIAGNOSTICS

Task	Metric	Score
The Corpus of Linguistic Acceptability	Matthew's Corr	0.0
The Stanford Sentiment Treebank	Accuracy	80.0
Microsoft Research Paraphrase Corpus	F1 / Accuracy	81.5/73.4
Semantic Textual Similarity Benchmark	Pearson-Spearman Corr	61.2/59.1
Quora Question Pairs	F1 / Accuracy	51.4/79.1
MultiNLI Matched	Accuracy	56.0
MultiNLI Mismatched	Accuracy	56.4
Question NLI	Accuracy	50.4
Recognizing Textual Entailment	Accuracy	54.1
Winograd NLI	Accuracy	65.1
Diagnostics Main	Matthew's Corr	9.2