README

Rotten-Tomatoes

- 大数据2101 张文达 U202115402
- 计　科2103 李克勤 U202115392
- 本硕博2101 罗正阳 U202115668

项目模块结构

代码仓库结构：

```
.
└── Rotten-Tomatoes
    ├── data                    # 数据集
    ├── docs                    # 作业要求
    ├── images                  # 文档中的图片
    ├── README.md               # 说明文档
    ├── src                     # 代码
    └── visual                  # 可视化
```

`src` 目录代码模块依赖结构：

```
.
├── tomato.py
│   ├── classifier.py
│   └── dataset.py
└── test.py
```

`tomato.py` 用于训练和验证数据集，它依赖于 `dataset` 和 `classifier` 这两个模块：

- `dataset` - 数据集，用于简化对数据的操作，并利用 `BertTokenizer` 对其进行分词。
- `classifier` - 分类器，使用 BERT 模型对评论进行分类。

```
$ cd src
$ python tomato.py                              # 训练和验证
```

`test.py` 则利用已经训练保存的模型 `bert.model` 对输入的新评论进行打分预测，例如：

```
$ python test.py "I think it's pretty good!"    # 试试新评论
```

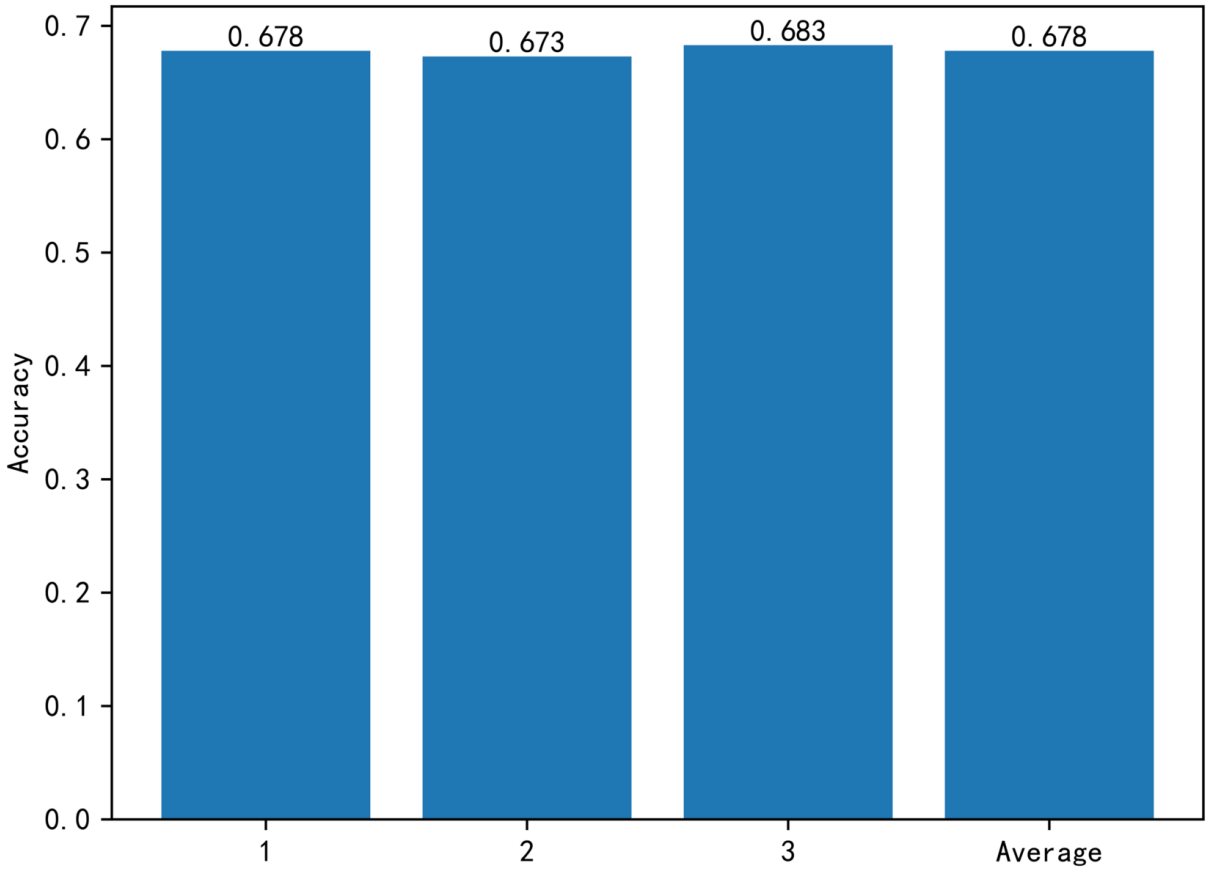
运行环境

- **CPU**: AMD Ryzen 7 5800H
- **GPU**: NVIDIA GeForce RTX 3060 Laptop GPU
- **CUDA**: 11.6
- **Python3**: 3.9.1
- **Conda**: 4.12.0
- **Torch**: 1.13.1+cu116

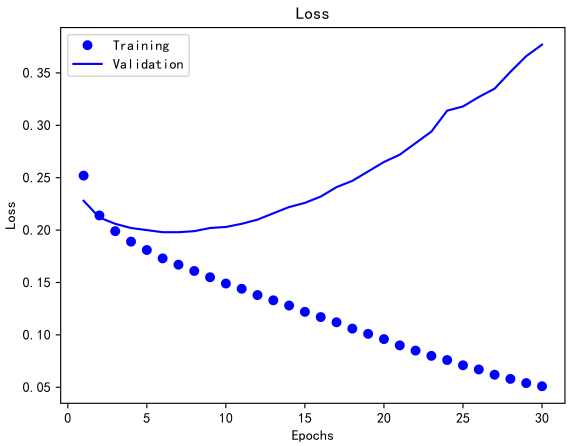
> 代码仓库中的 `src/bert.model` 是在上述 CUDA 环境中训练出来的

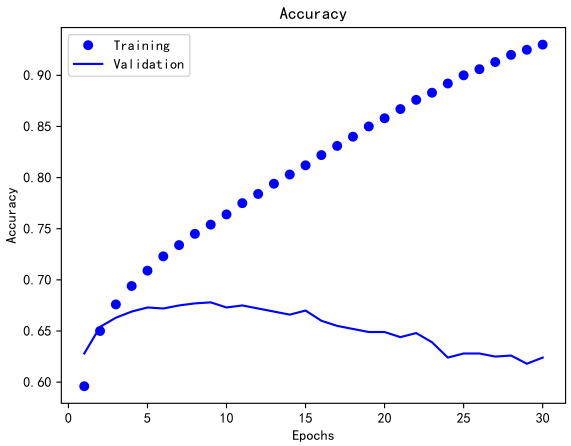
准确度

重复三次“训练集/验证集划分、训练、测试”，三次的平均准确率为 0.678：



运行 30 个 Epoch 的 Loss 和 Accuracy 曲线：


Loss
Training
Validation
0.35
0.30
0.25
0.20
0.15
0.10
0.05
Loss
0
5
10
15
20
25
30
Epochs


Accuracy
Training
Validation
0.90
0.85
0.80
0.75
0.70
0.65
0.60
Accuracy
0
5
10
15
20
25
30
Epochs