我的全部代码可以在https://github.com/shuitatata/24Fall-NLPDL.git中获取到。

Task1

具体的实现方式不在此介绍，可以在我的github仓库中查看到源代码。

对于few shot version，我参照论文中的做法，训练集抽取了了32个样本，测试集没有改变。

Task2

在实验二中，我使用了`robert-base`,`bert-base-uncased`,`allenai/scibert_scivocab_uncased`三种预训练模型，在`restaurant_sup`,`acl_sup`,`agnews_sup`三个任务上进行微调。

每个实验更换不同的种子重复了五次，种子分别为347913、594729、162850、95842、43288。

参数设定

训练3个epoch，learning rate为2e-5，有0.01的weight decay，batch为128。

实验结果

restaurant_sup

Model	Mean Accuracy	Std.Accuracy	Mean F1	Std.F1
bert_base	0.7587	0.0175	0.5441	0.0445
roberta_base	**0.7857**	0.0047	0.5791	0.0166
scibert_scivocab_uncased	0.7720	0.0064	**0.6201**	0.0141

acl_sup

Model	Mean Accuracy	Std.Accuracy	Mean F1	Std.F1
bert_base	0.5669	0.0362	0.1820	0.0381
roberta_base	0.5698	0.0070	0.2048	0.0023
scibert_scivocab_uncased	**0.6547**	0.0091	**0.2844**	0.0110

agnews_sup

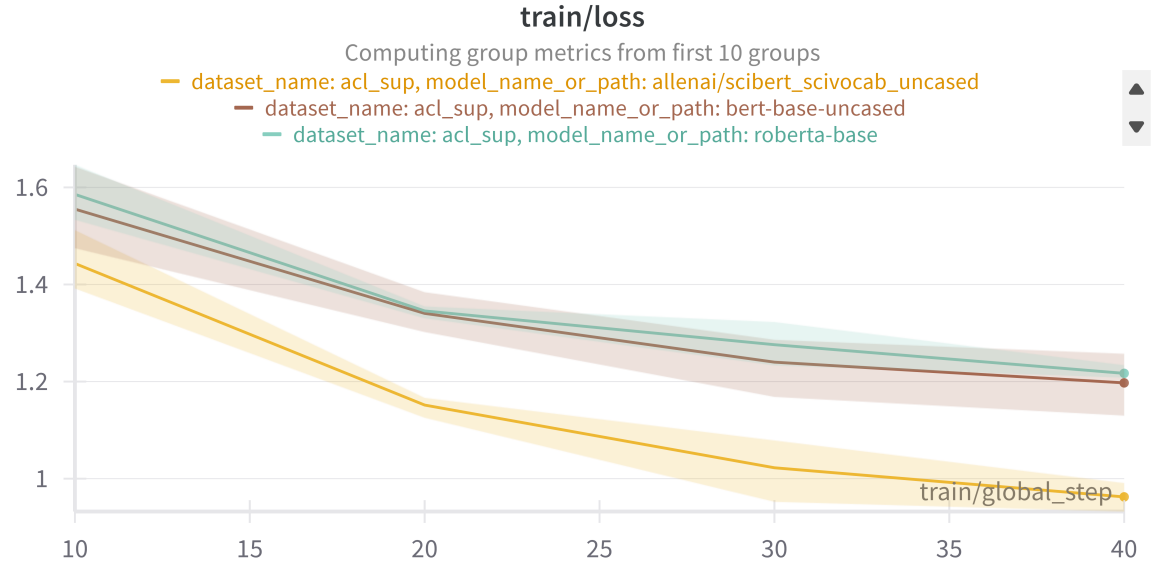
Model	Mean Accuracy	Std.Accuracy	Mean F1	Std.F1
bert-base	**0.9205**	0.0038	**0.9190**	0.0038
roberta-base	0.9197	0.0066	0.9181	0.0068
scibert_scivocab_uncased	0.9042	0.0030	0.9028	0.0032

在restaurant_sup和agnews_sup数据集上，模型都有着不错的效果，且性能差异不大。但是在acl_sup数据集上，模型的分类效果不算太好，同时scibert相较于其他两种模型有比较大的优势。

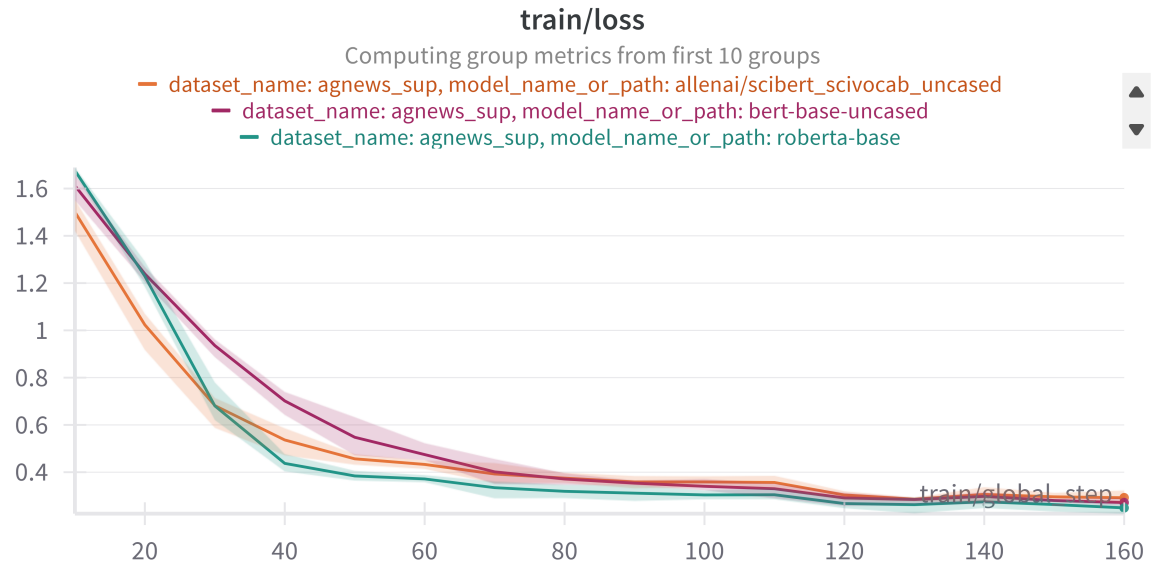
Loss图像：

restaurant_sup

acl_sup



agnews_sup



Task3

restaurant_sup

Model	Mean Accuracy	Std.Accuracy	Mean F1	Std.F1
roberta_adapter	0.6500	0.0138	0.2626	0.0103
roberta_base	**0.7857**	0.0047	**0.5791**	0.0166

acl_sup

Model	Mean Accuracy	Std.Accuracy	Mean F1	Std.F1
roberta_adapter	0.5036	0.0144	0.1178	0.0103
roberta_base	**0.5698**	0.0070	**0.2048**	0.0023

agnews_sup

Model	Mean Accuracy	Std.Accuracy	Mean F1	Std.F1
roberta_adapter	0.7071	0.0603	0.7011	0.0583
roberta_base	**0.9197**	0.0066	**0.9181**	0.0068

从结果可以看出，仅仅微调adapter的效果确实比微调整个大模型更差，也许可以通过增加adapter参数量或延长训练时间来取得更好的效果。但是相较于微调完整模型，adapter微调的方法节约了相当大的显存开销。

显存分析

我们假设进行**全精度训练**，即参数存储为float32类型，占用四个字节。

大语言模型训练时的显存占用主要分为以下三个部分：

- 模型参数
- 优化器状态参数
- 梯度参数

对于模型参数而言，3B的大模型共有3×10^9个参数，因此共占用空间$3 \times 10^9 \times 4$ Byte 约为12GB

对于梯度而言，每一个参数都对应一个梯度，因此梯度的显存占用与模型权重相同，也为12GB

而对于优化器而言，假设采用Adam优化器，仍是全精度存储，则每个权重参数需要同时存储**动量**和**方差**，需要的显存占用为模型权重的两倍，为24GB。

综上，若想完整训练一个3B的大模型，采用Adam优化器，全精度训练，则需要大概48GB的显存。

而对于PEFT方法，由于不知道adapter的参数量是多少，不方便进行估测。在我实际的训练中，Task2完整微调模型大概需要12-14G的显存，而使用adapter只需要8G-9G的显存，大概节约了30%~40%的显存占用。