北京邮电大学 2023—2024 学年第一学期

《神经网络与深度学习》课程实验作业（四）

实验内容：自然语言处理基础

注意事项：

① 本次实验包含一道题，共计 30 分；

② 所有实验结果需以实验报告 word 文档的形式进行提交，文件命名格式：实验四_姓名_学号.word，文件中需要将作者设置为本人姓名；

③ 实验报告中可插入代码片段，完整代码无需放在实验报告中，以压缩包的形式添加即可，压缩包命名格式：实验四代码_姓名_学号.zip；

④ 作业提交截止时间：2023 年 12 月 19 日晚上 20：00

搭建 Transformer 编码器完成文本语义匹配任务 (30 分)

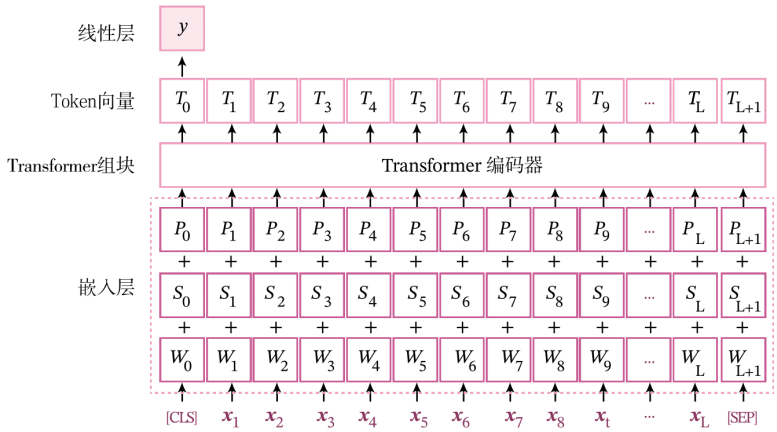
AFQMC 数据集是一个蚂蚁金融语义相似度数据集，用于问题相似度计算，数据集包括训练集、验证集、测试集 3 个文件，分别包含 34334、4316 以及 3861 条数据，每条数据有三个属性，分别是句子 1、句子 2、句子相似度标签。相似度标签为 1 表示两个句子含义类似，标签为 0 则表示含义不同。请基于该数据集完成以下实验内容：

(1) 数据集构建，包括：利用词表将句子中的每个中文字符转换成 id、对不在词汇表里面的字做出适当处理、在输入中加入句子的分隔符号、在起始位置加入占位符、小批量数据的组装及对齐。构建完成后打印一条 mini-batch 的数据进行验证。(2 分)

(2) 实现输入编码、分段编码和位置编码，将这三种编码组合为嵌入层，并打印该层的输入输出，其中位置编码需使用三角函数。(3 分)

(3) 实现多头自注意力层和 add&norm 层。(3 分)

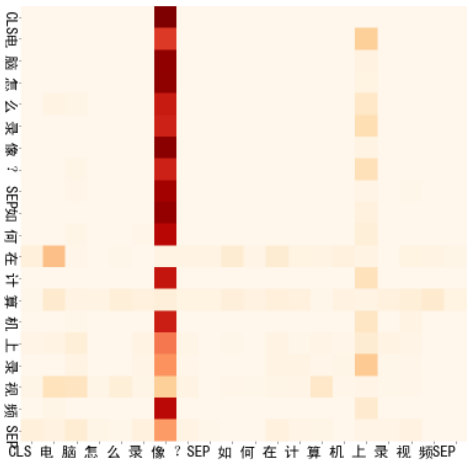
(4) 搭建一个两层的 Transformer 编码器，利用嵌入层、Transformer 编码器和合适的分类器构建完成你的语义匹配模型，并在报告中说明你的模型组成，可参考下图对你所构建的模型进行画图说明。(3 分)



(5) 训练模型，在验证集上计算准确率，保存在验证集上准确率最高的模型 (2 分)，并使用 tensorboard 等可视化插件，展示训练过程中的精度变化和损失变化。(2 分)

(6) 加载保存的模型，在测试集上随机选取 50 条数据进行语义匹配测试，在测试集上完成测试并输出准确率。(2 分)

(7) 输入一条样本提取多头注意力权重，参考下图对注意力机制的计算结果进行可视化展示 (2 分)，并进行结果分析 (1 分)。



(8) 改变 Transformer 的层数再次实验，输出测试集准确率结果，并与之前的结果对比。(3 分)

(9) 寻找方法提升模型精度，并将模型精度数据及相应截图上传至问卷中。(4 分)

(10) 层规范化的位置有两种 pre_norm 和 post_norm，查询资料了解二者区别并说明自己的模型中层规范化操作的位置是 pre_norm 还是 post_norm (1 分)，然后尝试另一种层规范化操作，对比二者在具体训练中的区别并分析原因 (2 分)。

提示：

1. AFQMC 数据集及嵌入表已给出；

2.data.py 文件中有数据预处理可能用到的函数可以参考；

3.实现自注意力模型时，掩蔽元素不应参与注意力的计算；