语言数据分类项目报告

问题分析

本项目聚焦于使用机器学习技术对语言数据进行分类的问题。我们通过比较 LSTM 和 Transformer 两种不同的模型，探讨它们在处理具有不同语言特性的数据时的效能和适应性。该项目的核心挑战在于理解和处理不同语言的数据特征，并构建一个准确识别和分类这些特征的高效模型。

数据集描述

本项目使用的数据集包含两种不同语言的数据。具体来说，数据集的特点如下：

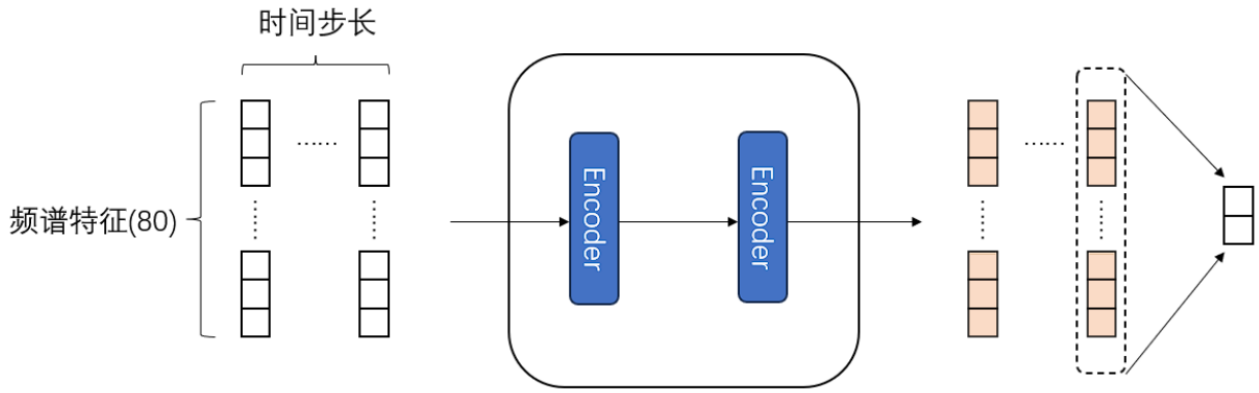
- **数据量**：训练集包括大约 4000个样本，分为两个语言类别，每个类别各有2000个。
- **数据描述**：数据形状为[x,80]，其中x为步长，80则是特征数，由于数据步长不一，对数据长度进行观察，训练数据的样本时间长度统计为：`'max_length': 734, 'min_length': 120`

  而测试集数据的样本时间长度统计为：`'max_length': 766, 'min_length': 40`
- **数据预处理**：对于每种语言的数据，我们执行了标准化和裁剪操作，以确保数据格式一致，便于模型处理,为了统一长度，我们将数据裁剪保留前40个步长，以涵盖所有的数据。
- **标注**：每个样本都被标注了相应的语言类别标签，以供后续的监督学习使用。
- **数据集划分**：按照8：2把训练数据划分为训练集和验证集。

模型构建

项目中构建了两种不同的模型：

- **LSTM 模型**：此模型使用了 PyTorch 构建，包含两层的 LSTM 单元和 Dropout 正则化。目的是捕捉语言数据中的时序特性。
- **Transformer 模型**：此模型利用了注意力机制，旨在处理大规模数据集时更有效地捕捉长距离依赖关系。
- **模型原理**



实验部分

实验设计包括：

- **模型训练**：我们使用了二元交叉熵损失函数和 Adam 优化器来训练模型，并将学习率设置为e-5，并设置dropout为0.2.
- **性能评估**：在验证集上，LSTM 模型在验证集上达到了约 94% 的准确率。Transformer 模型也在验证集上达到92%的准确率。
- **结果分析**：通过比较两种模型在相同测试集上的表现，选择了LSTM模型来进行最后的标注。

Model	Train Accuracy	Validation Accuracy
LSTM (num_layer=2)	96.12%	94.50%
LSTM (num_layer=4)	95.35%	93.70%
Transformer (num_layer=2)	92.32%	91.15%
Transformer (num_layer=4)	92.00%	92.11%

结论

本项目成功展示了 LSTM 和 Transformer 模型在语言数据分类任务中的有效性。LSTM 模型适用于处理时序数据，而 LSTM 模型则在处理需要长距离依赖关系的大规模数据时更为有效。未来的工作可以包括对模型的进一步优化，以及探索将这些模型应用于更多语言类型和更复杂场景的可能性。