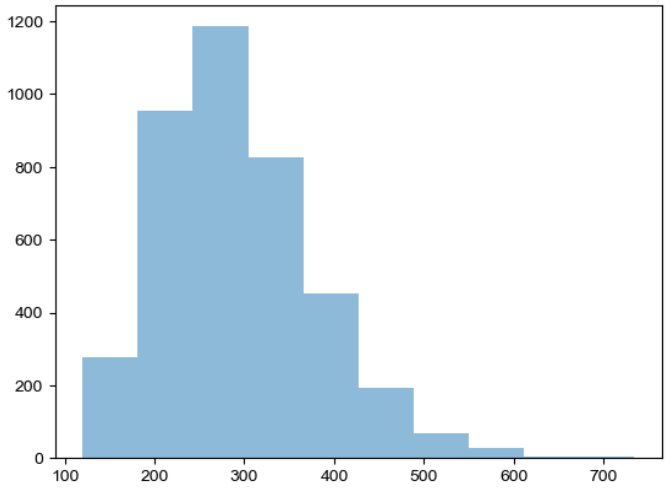
大数据案例分析期末报告

1. 问题描述

本项目的目标是构建一个语种识别模型，通过给定的中英文语音数据集，使用梅尔频谱提取的特征进行训练，并在测试集上进行语种标签的预测。其中训练数据集样本容量为4000，包括2000条中文数据和2000条英文数据，标签分别为language_0和language_1。测试集样本容量为2000。

2. 数据预处理

（1）定义split_long_mel函数，将长的梅尔频谱切分成固定长度（clip target）的片段，将其整理成适合神经网络训练的格式，并准备好用于 DataLoader 进行批处理。考虑到训练集长度数据中位数为281，测试集长度数据中位数为199，这里择中选择了256为clip target进行切分。其中训练数据集原始 mel 长度的直方图如下：



（2）加载训练数据集，并使用 split_long_mel 函数将每个 mel 文件切分成小片段。每个小片段都被标记为相应的类别，并分别存储在 X_0、Y_0、X_1、Y_1 中。同时，记录每个 mel spectrogram 的长度，以便后续可视化。

（3） 使用train_test_split函数将切分后的训练数据集按照9:0.5:0.5的比例随机划分为训练集、验证集和伪测试集，后者用以防止模型过拟合。

（4）将训练集、验证集和伪测试集加载到 PyTorch 的 TensorDataset 中，以准备用于 DataLoader。

3. 数据加载

（1）使用PyTorch中的DataLoader和TensorDataset对数据进行加载和处理。

（2）random mask的数据增强方法，定义random_mask_collate_fn函数，通过对训练集的Mel频谱图应用随机掩码帮助模型更好地学习对不完整或噪声图像的处理，增加模型的泛化能力，详情可参考 https://zhuanlan.zhihu.com/p/598985864 和 https://zhuanlan.zhihu.com/p/73892922。具体操作如下：

- 对于每个图像，创建一个与其大小相同的掩码（全为1的矩阵）；
- 以50%的概率，选择在水平方向生成一条水平线，或者在垂直方向生成一条垂直线；
- 根据生成的线的位置和宽度，在掩码上将相应区域的值置为0，以创建掩码；
- 将图像和掩码相乘，以在图像上应用掩码；
- 将所有处理后的图像和掩码组成的列表转换为张量，标签转换为张量；
- 最终返回带有掩码的图像张量和标签张量。

4. 模型构建

构建一个简单的卷积神经网络模型(CNN)，包含卷积层、最大池化层、批归一化层、Dropout层和全连接层。通过训练循环对模型进行训练，并使用验证集进行性能评估。该模型的结构如下：

- CBL(1, 16, 7, 1, 0)
- MaxPool2d(2, 2)
- CBL(16, 16, 5, 1, 0)
- MaxPool2d(2, 2)
- CBL(16, 16, 3, 1, 0)
- MaxPool2d(2, 2)
- Dropout(0.4)
- Flatten()
- Linear(3248, 128)
- LeakyReLU()
- Linear(128, 2)

5. 参数优化

使用交叉熵损失函数和adam优化器对模型进行训练。通过训练集和验证集的准确性和损失值来监控模型的性能。最终，选取在验证集上准确性最高的模型作为最终模型，并在伪测试集上评估其性能，记录最

佳测试准确性和交叉熵损失。

```
Training loss: 0.0789
Test Error:
 Accuracy: 97.0%, Avg loss: 0.087535

Test Error:
 Accuracy: 96.6%, Avg loss: 0.057170

best valid records: acc 0.9756838905775076 test acc 0.975609756097561 loss 0.0949737696829153
best test records: acc 0.9726443768996961 test acc 0.9878048780487805 loss 0.07800214277937058
best train records: acc 0.9574468085106383 test acc 0.9603658536585366 loss 0.0773456946332404
```

6. 实际预测

加载预训练好的神经网络模型，并使用对测试集数据进行语种自动识别。