MelCNN：基于CNN与Mel谱图的

语音识别模型

计算机科学技术学院 18307130092 朱秦

**摘要：**本文针对孤立词的语音识别这一任务，实现了一个MelCNN模型。该模型首先利用mel谱图对语音信号进行数据处理，将时域的一维语音信号拓展到频域上二维图像信号，然后利用CNN对处理好的图像进行特征的提取、参数的学习，然后输出预测结果。

MelCNN中一共包括了3个CNN网络。一个利用VGG16做预训练，另外两个分别对VGG与Resnet进行简化与改动，使之更加适应本任务。最后在预测阶段将3个网络得到的预测结果进行归一化与求和，得出最终预测结果。

本模型相较于直接利用MFCC系数，或者使用高斯混合模型与隐式马尔科夫模型等传统的语音识别模型，拥有更好的准确性和泛化能力。只需要改动最后的全连接层就可以将任务从原本20孤立词识别拓展成任意大小的孤立词识别。此外，本文利用设置循环学习率，三模型多个模型共同预测等方法大大提高了MelCNN模型的准确率。最终达到成果：三个网络在测试集中可以分别达到94.25%，91.5%与94.51%的准确率，将其归一化求和共同预测后可以达到98.3%的准确率。

**关键词：语音识别，CNN, Mel谱图，孤立词识别**

目录

[1项目任务重述 3](#_Toc75031170)

[2 模型介绍 3](#_Toc75031171)

[2.1 MelCNN模型概述 3](#_Toc75031172)

[2.2 MelCNN模型框架 3](#_Toc75031173)

[MelCNN模型创新点 5](#_Toc75031174)

[3模型实现 6](#_Toc75031175)

[3.1数据集处理 6](#_Toc75031176)

[3.2 Mel谱图绘制 6](#_Toc75031177)

[3.3模型搭建与训练 6](#_Toc75031178)

[3.4模型的预测 6](#_Toc75031179)

[4模型评价 7](#_Toc75031180)

[4.1指标计算 7](#_Toc75031181)

[4.1.1混淆矩阵 7](#_Toc75031182)

[4.1.2正确率(accuracy)与验证集的平均损失(mean loss) 9](#_Toc75031183)

[4.2理论分析与优化方向 11](#_Toc75031184)

[A 附录 11](#_Toc75031185)

[A.1代码 11](#_Toc75031186)

# 1项目任务重述

语音识别从人工智能诞生伊始就是非常重要的课题，它是让机器理解人类语言信息、与人类进行沟通交互不可或缺的一环。这一技术发展以及相对成熟，早在上世纪60年代信息学科学家们就已经提出并实现使用MFCC倒谱系数等方法对孤立词进行识别。本课程的项目任务要求复现了这一问题的解决方案，提供了20个关键词，以及每个关键词17\*20条语音信息作为数据集（17名同学，每人20遍），来实现对任意音色的语音输入的该20词的语音识别。

# 2 模型介绍

## 2.1 MelCNN模型概述

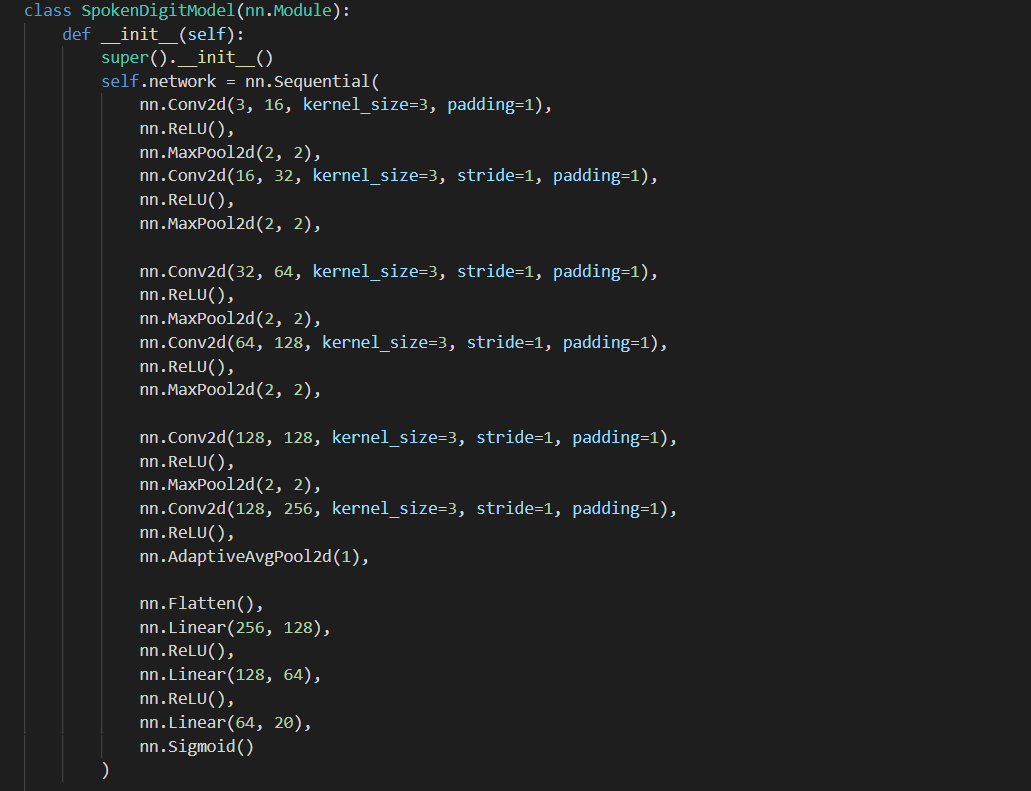
MelCNN模型是一个基于Mel谱图与卷积神经网络（CNN）的语音识别模型。

该模型首先利用mel谱图对语音信号进行数据处理，将时域的一维语音信号拓展到频域上二维图像信号，然后利用CNN对处理好的图像进行特征的提取、参数的学习，然后输出预测结果。

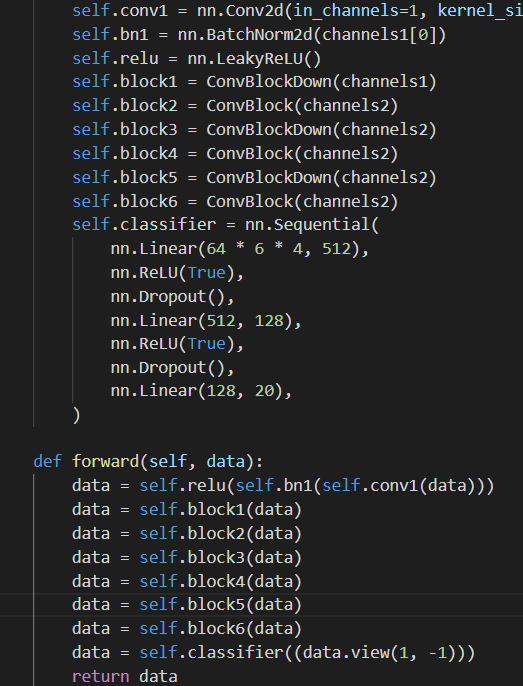
## 2.2 MelCNN模型框架

MelCNN模型中一共涵盖了3个CNN网络。

第一个模型是仿照了VGG模型自己搭建了一给网络，命名为vgg618结构如下图，使用3\*3的卷积核，每个卷积模块用relu作为激活函数，接上一个maxpooling等。最后用一个sigmoid函数。



第二个模型模仿了Resnet,自己进行搭建,命名为DSP。使用了7\*7，3\*3等卷积核。一共8层，最后一层是使用了一个分类器实现分类预测。详见模型实现与附录的代码部分。



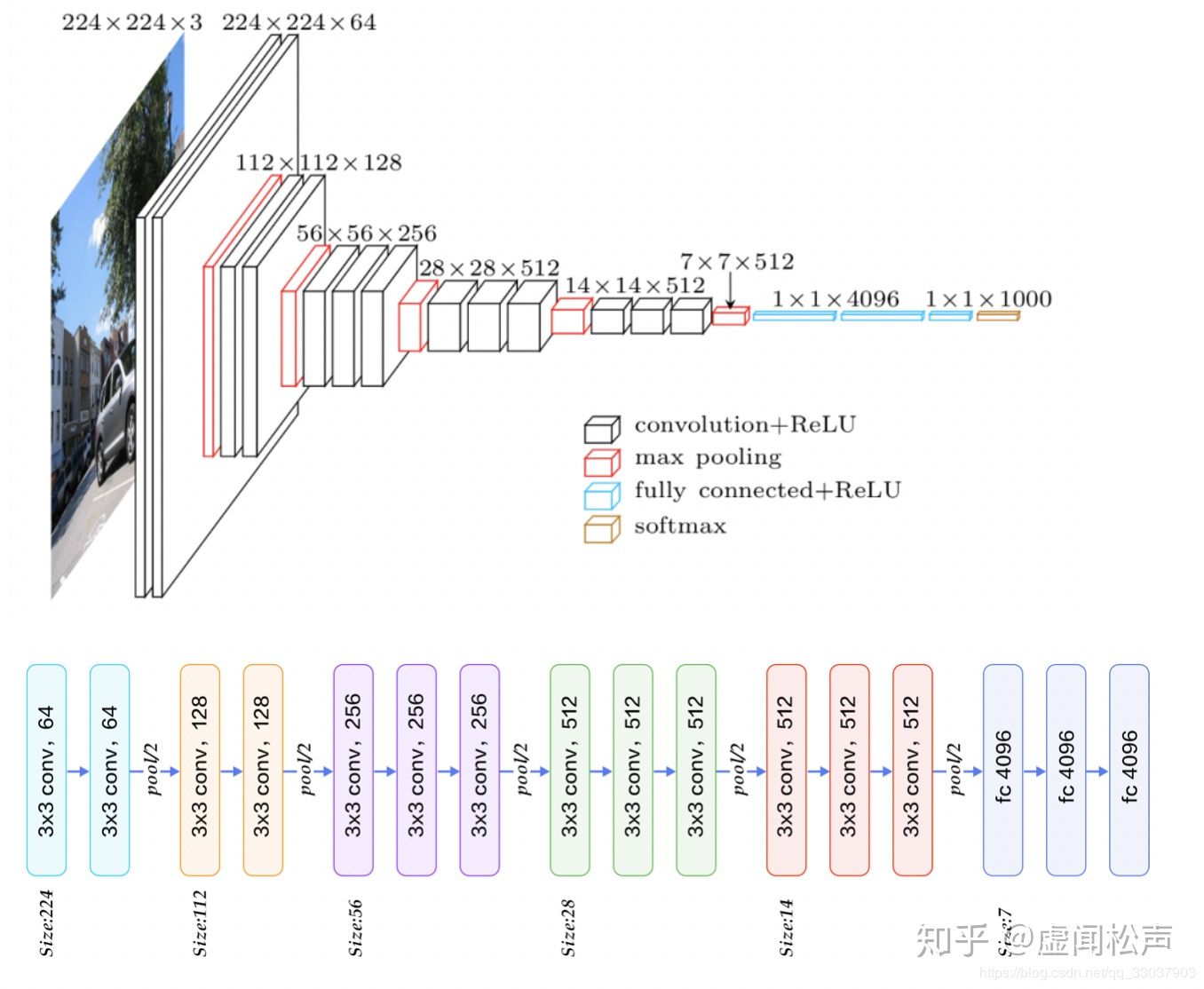
第三个模型直接套用了VGG16，只是对其线性层和最后的全连接层进行了变化，使之适应本任务的输入与20分类的背景。直接选择VGG的原因有：

1、VGG是2012年效果最好的图像分类模型，斩获了 ILSVRC 比赛冠军。

2、VGG模型简单容易理解

3、VGG模型非常契合本任务，拿其做预训练以期得到好的结果和效率。

VGG网络结构如下，改动之处会在模型实现中进行说明：



MelCNN在三个网络搭建的基础上，在预测阶段将3个网络得到的预测结果进行归一化与求和，然后进行argmax得到最后综合的预测结果。这样做结合了三个模型所长，大幅提高了准确率。

## MelCNN模型创新点

本模型相较于直接利用MFCC系数，或者使用高斯混合模型与隐式马尔科夫模型等传统的语音识别模型，拥有更好的准确性和泛化能力。只需要改动最后的全连接层就可以将任务从原本20孤立词识别拓展成任意大小的孤立词识别。

此外，本文利用设置循环学习率提高了模型训练速度与效果，即学习率在不同时期的最小值和最大值之间波动，而不是单调下降。这样做的原因是：初始训练率对模型的性能是至关重要的，低训练率可防止在训练开始时被卡住，随后的波动抑制了局部极小值的情况，太低又会对学习速率造成影响。所以可以通过改变epoch和循环训练率对模型进行反复训练。

本文还运用了前文提及的三模型多个模型共同预测的方法大大提高了MelCNN模型的准确率。

# 3模型实现

## 3.1数据集处理

由于同学采样的数据，命名方式与存储方式不同，我先对所有数据进行了命名的格式规整化，采用学号.xx.xx.dat的格式进行命名，将共6800条数据整理到一个文件夹下，同时刨除了无效或者损毁的数据，剩下最后6787条有效数据。

## 3.2 Mel谱图绘制

Mel谱图的绘制在编程作业中已经完成，但是这里考虑到最后代码篇幅的原因，直接调用了librosa库中的librosa.feature.melspectrogram进行处理，同时用librosa.power\_to\_db将其频率转换成对数刻度。最后，对于MelCNN中VGG618这个网络，需要先对语音信号进行预处理将其全部转化成mel谱图才行（因为该网络的输入要求channel为3）。另外两个网络VGG16与DSP则不用。

## 3.3模型搭建与训练

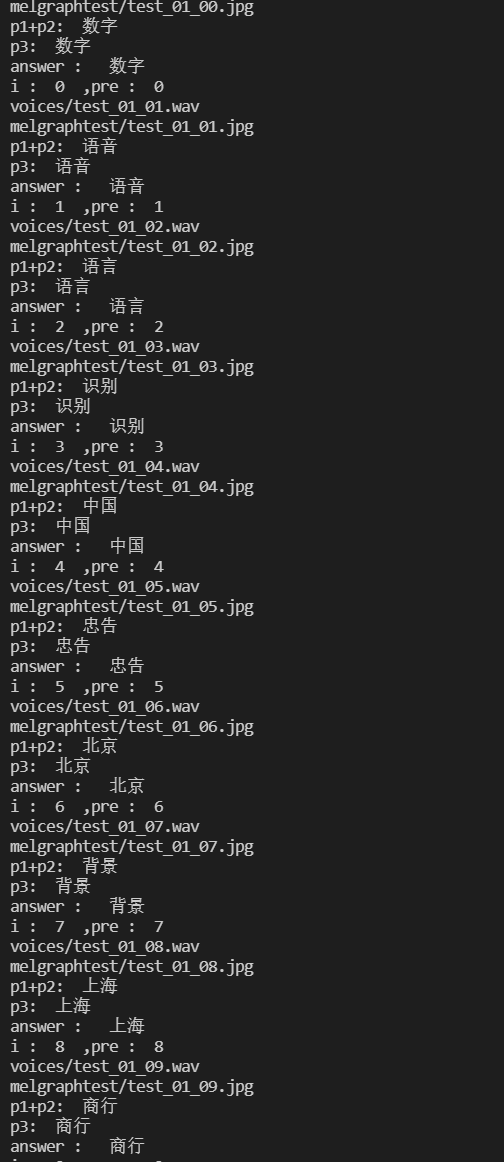
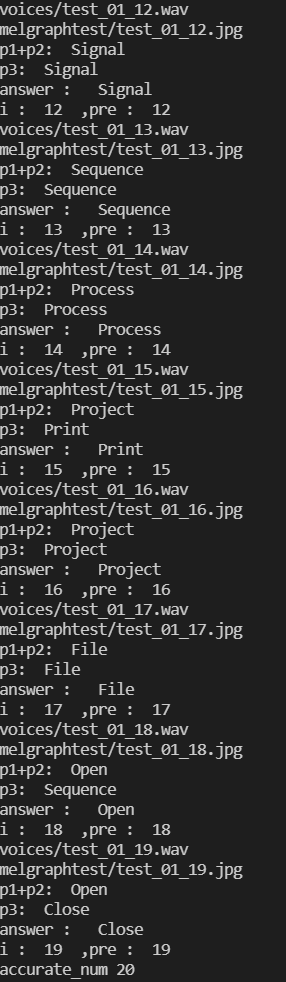
在2模型介绍中已经对模型的结构有过概述，这里不再赘述，详细结构可以见附录A中的代码部分，VGG618, DSP, VGG16分别对应model618.py, Model.py, model.py三个python文件。

模型的训练我才用了2模型介绍中所说的循环学习率，对数据集采用9：1的比例随机划分得到训练集与验证集（由于数据集数量实在过于小，我将大部分用于训练，评估模型所用的测试集请同寝室楼层的同学进行额外录制来评估），取一个batch=32，每个模型均训练了共计epoch=128轮。训练使用的是torch.optim.Adam作为优化器，nn.CrossEntropyLoss交叉熵为损失函数。

另外与一般的cnn训练有所不同之处在于，因为16级学长的VGG16参数作预训练，但是他们当时的训练目标与现在不同，所以使用了一个字典表来对标签进行了处理使两者匹配，使用VGG16预训练的原因在于数据集的大小实在过于小，很容易过拟合，采用预训练可以让模型有更好的训练结果。详细内容可见附录A的代码部分。

## 3.4模型的预测

写了一个简单的ui.py可以对录制好的音频代码利用已经训练好的桑网络快速实现语音识别，并返回各个网络的结果以及最终综合的结果。如图所示：（accurate num 代表预测正确的数值，此样例为20，共计20个，正确率100%）

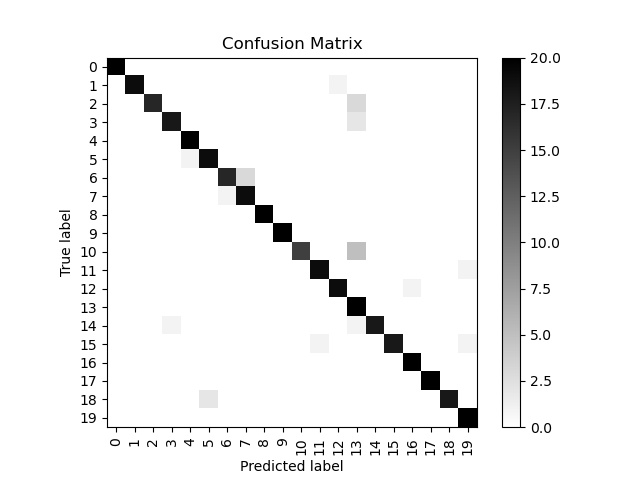
 

# 4模型评价

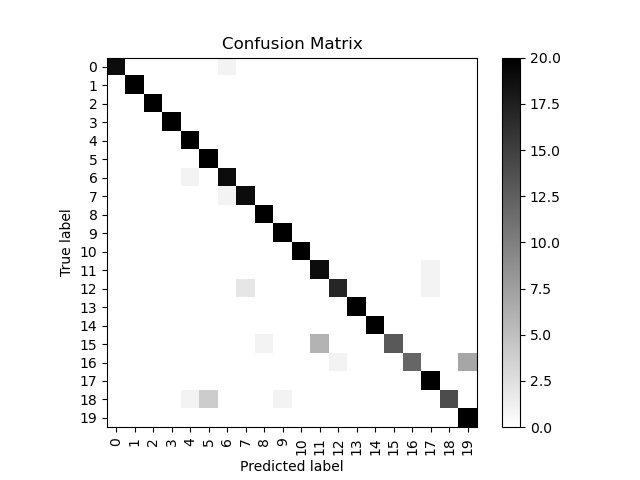
## 4.1指标计算

### 4.1.1混淆矩阵

**由于有三个cnn网络，为保证篇幅不显赘余，所以图片展示不全部展示（以1~2个网络为例，混淆矩阵以VGG16,DSP为例，正确率等以VGG618为例）：**



这是DSP(模仿的Resnet)网络训练出来后在验证集上得到的混淆矩阵，正确率高达94.25%。可以看到VGG16对于10以后，即英文的识别相对中文来说没有那么稳定。



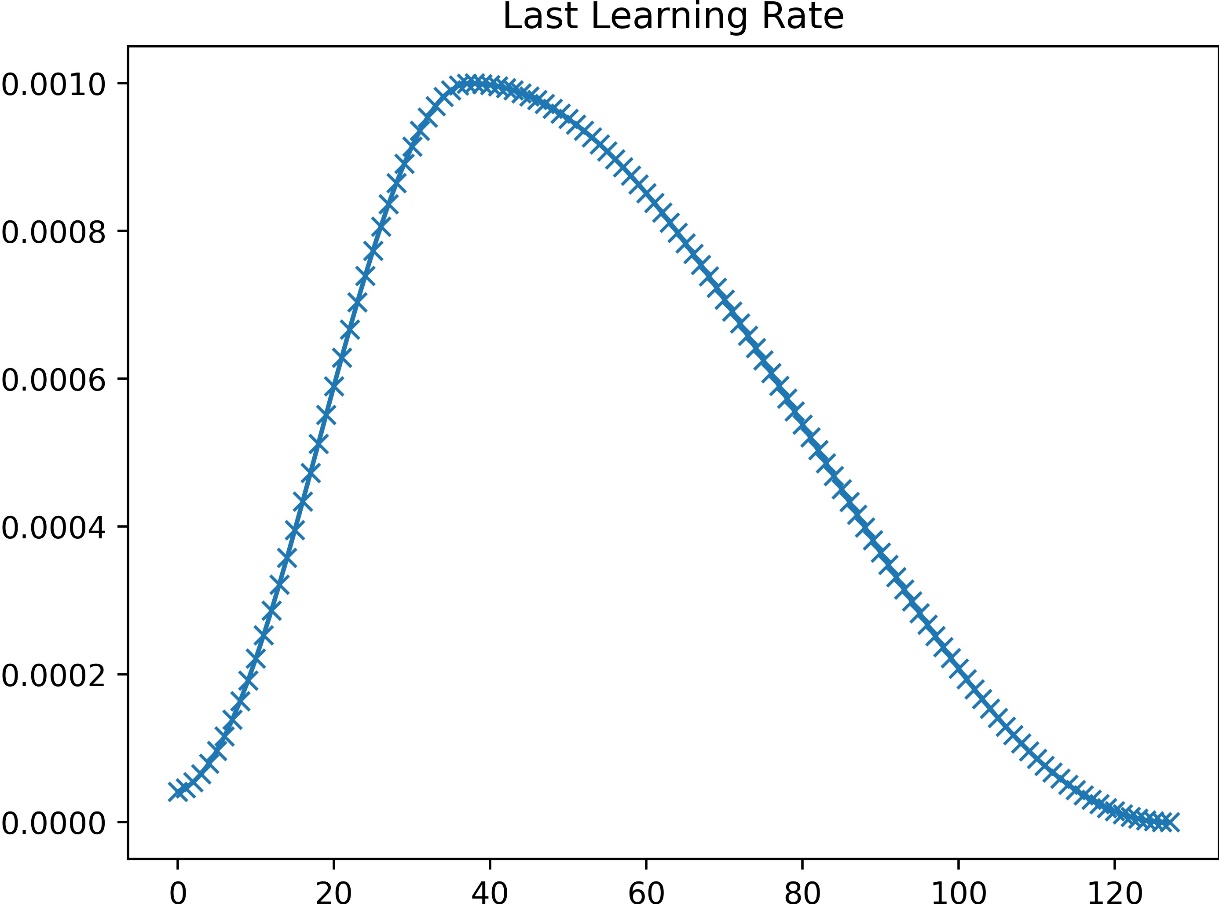
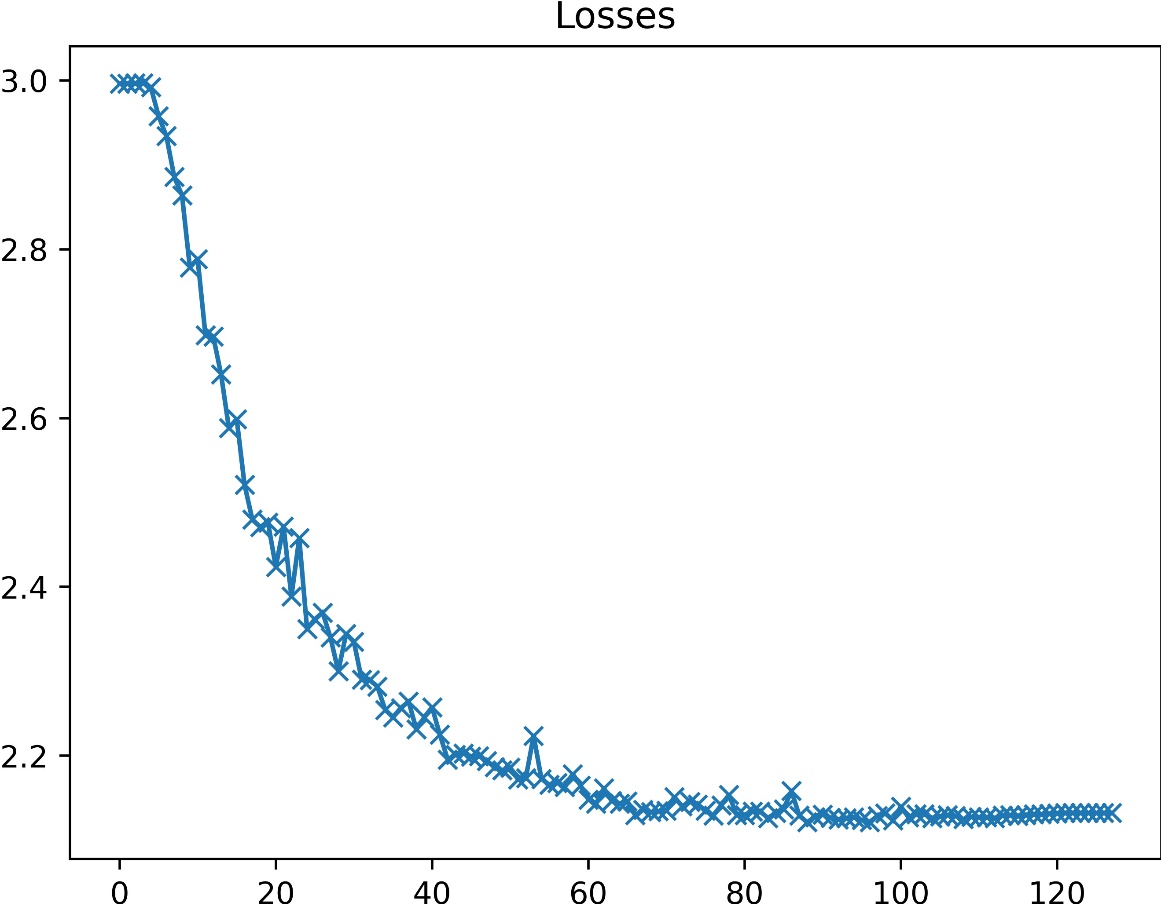
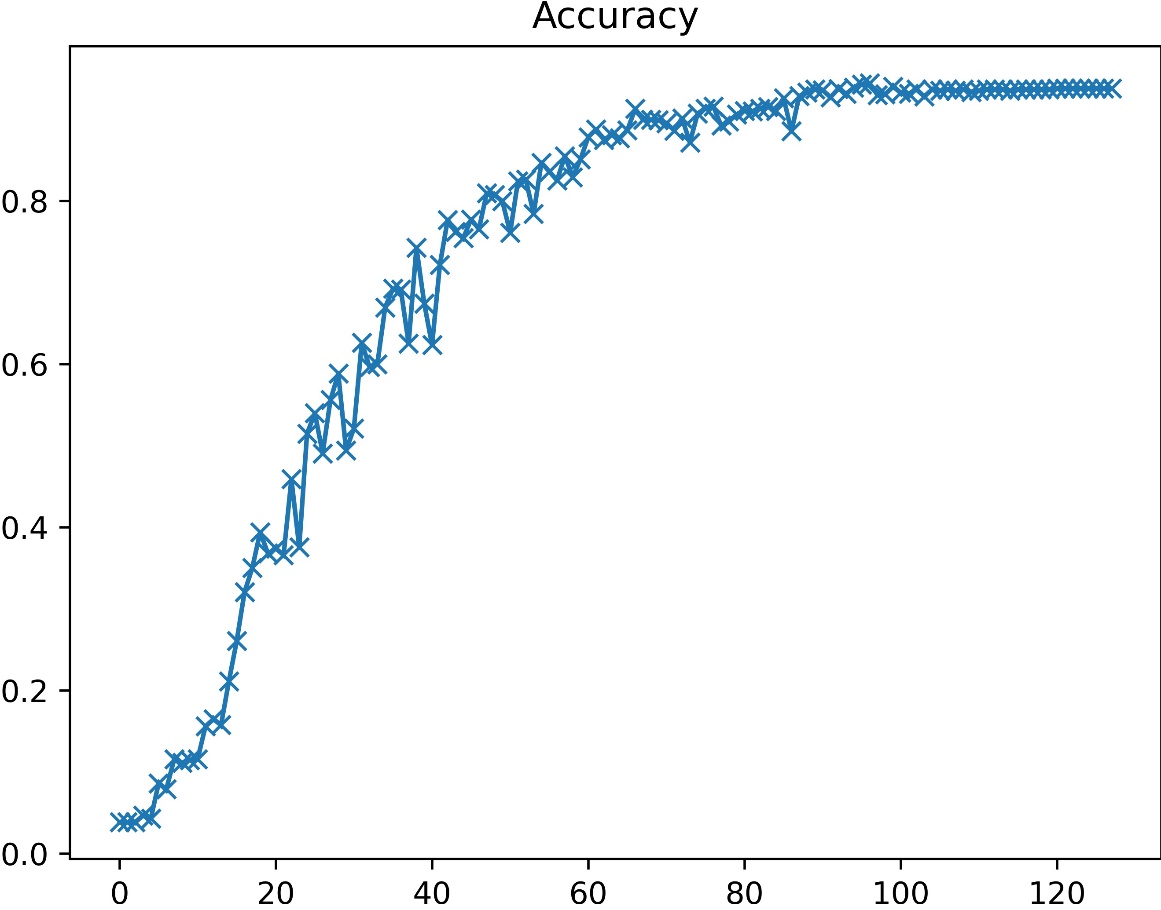
这是VGG16网络训练出来后在验证集上得到的混淆矩阵，正确率高达91.5%。可以看到VGG16对于10以后，即英文的识别相对中文来说没有那么理想，但是中文准确率相当高。

同样的混淆网络对于VGG618来说，正确率高达94.51%。与前两给网络不同的是，VGG618对英文识别效果非常好，中文没那么稳定。

### 4.1.2正确率(accuracy)与验证集的平均损失(mean loss)

由于在训练部分训练了128epoch，所以变化很明显，对于三个网络下面以VGG618为例，给出其随着epoch的增大（训练轮数的不断增加），正确率、验证集平均损失、学习率的变化图。

其中学习变化是因为使用了循环学习率来进行学习，这也是循环学习率的直观展示。



## 4.2理论分析与优化方向

我们可以看到三个网络对于中英文识别效果是有差异的，VGG16与DSP对中文效果优异，而VGG618对英文识别优异，三者相加然后综合起来看就可以达到一个非常不错的效果，在同学现场录的三个测试集（20个词）中分别达到了95%，100%，100%的正确率，是一个不小的提升。

事实上我有一个更好的想法，那就是再训练一个2分类的网络，识别中英文，因为中英文实际上发音是有很大区别的，尤其是孤立词，中文的音节较少，一般一个字一个音节，英文则会一个单词可能出现3-5个音节。所以先对其做一个中英文的识别，然后对于中文单词运用VGG16与DSP的综合模型，对于英文单词使用VGG618网络的模型，这样正确率可能会有进一步提高，下一步的模型优化可以朝这个方向做。

# A 附录

## A.1代码

见github\_id：8188zq (在验收结束后会上传)