**语音识别实践报告**

**09019203汪跃阳**

引言：本次报告分为四个部分，分别是模型搭建， 模型训练， 模型评估， 综合评估。 下面我将依次阐述这四个部分。

一、模型搭建

1.1 总览

从总体上看语音识别主要分一下几个模块， 首先我们的原始语言经过MFCC特征提取获取特征向量， 这些向量通过我们训练好的神经网络输出拼音， 然后我们的HMM模型根据神经网络生成的拼音生成汉字，最终实现了语音到文字的过程。

**原始语音**

**输入**

**特征**

HMM模型

语言

**拼音**

**汉字**

1.2 神经网络模型简述

3 Residual

Convolutional Layer

Linear Layer

5 Bidirectional GRU Layer

Linear Layer

Softmax

音调概率

解码函数

拼音

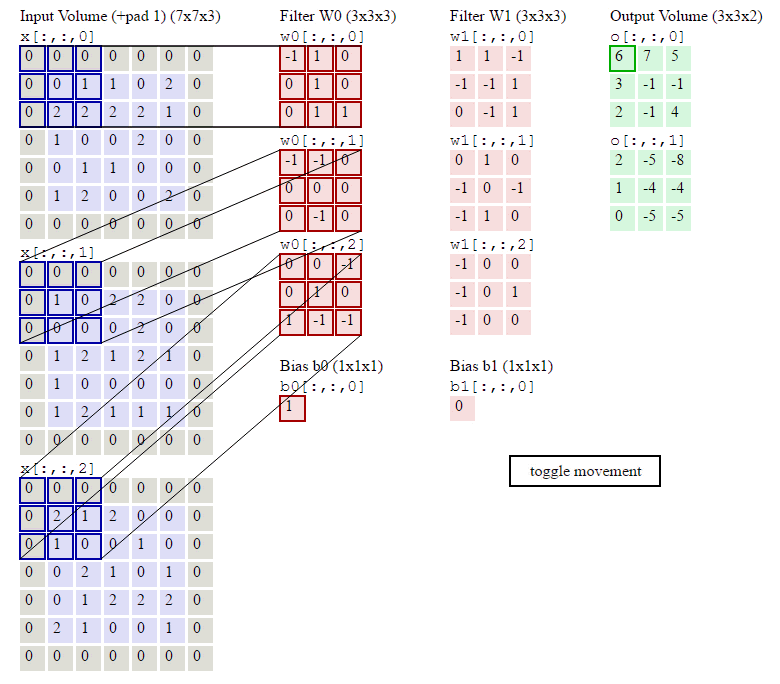
上图为我们搭建的神经网络模型，MFCC特征向量依次经过三层残差卷积网络进行特征提取，经过全连接层进行归一化， 经过循环神经网络双向GRU处理学习信号， 最后再经过全连接层和Softmax层输出预测结果， 最终我跟根据得到的影评概率我们调用解码函数生成拼音。

1.2.1 残差卷积层简述  
 为什么我们使用卷积网络而不使用全连接网络， 因为全连接网络：

• 参数数量太多 考虑一个输入1000\*1000像素的图片(一百万像素，现在已经不能算大图了)，输入层有1000\*1000=100万节点。假设第一个隐藏层有100个节点(这个数量并不多)，那么仅这一层就有(1000\*1000+1)\*100=1亿参数，这实在是太多了！我们看到图像只扩大一点，参数数量就会多很多，因此它的扩展性很差。

• 没有利用像素之间的位置信息 对于图像识别任务来说，每个像素和其周围像素的联系是比较紧密的，和离得很远的像素的联系可能就很小了。如果一个神经元和上一层所有神经元相连，那么就相当于对于一个像素来说，把图像的所有像素都等同看待，这不符合前面的假设。当我们完成每个连接权重的学习之后，最终可能会发现，有大量的权重，它们的值都是很小的(也就是这些连接其实无关紧要)。努力学习大量并不重要的权重，这样的学习必将是非常低效的。

• 网络层数限制 我们知道网络层数越多其表达能力越强，但是通过梯度下降方法训练深度全连接神经网络很困难，因为全连接神经网络的梯度很难传递超过3层。因此，我们不可能得到一个很深的全连接神经网络，也就限制了它的能力。



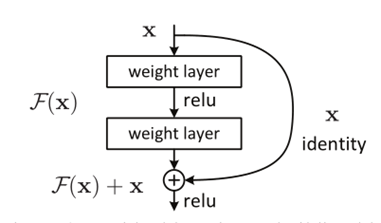
那么，卷积神经网络又是怎样解决这个问题的呢？主要有三个思路：

* 局部连接 这个是最容易想到的，每个神经元不再和上一层的所有神经元相连，而只和一小部分神经元相连。这样就减少了很多参数。
* 权值共享 一组连接可以共享同一个权重，而不是每个连接有一个不同的权重，这样又减少了很多参数。
* 下采样 可以使用Pooling来减少每层的样本数，进一步减少参数数量，同时还可以提升模型的鲁棒性。

对于语音识别任务来说，卷积神经网络通过尽可能保留重要的参数，去掉大量不重要的参数，来达到更好的学习效果。

尽管我们使用卷积网络，但非常深的神经网络是很难训练的，因为存在梯度弥散和梯度爆炸问题。什么叫梯度弥散呢？

假如从上一层传过来的梯度为1，经过一个block之后，得到的梯度已经变成了0.0001和0.01，也就是说，梯度流过一个blcok之后，就已经下降了几个量级，传到前一层的梯度将会变得很小！



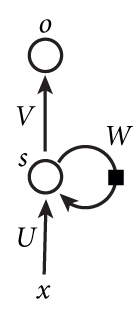
为了解决这个问题，有人提出了残差网络，残差网络的输出结果是普通网络的输出结果加上它本身， 这样，经过梯度传播后，现在传到前一层的梯度就多了一个x的梯度，假设为“1”！正是由于多了这条捷径，来自深层的梯度能直接畅通无阻地通过，去到上一层，使得浅层的网络层参数等到有效的训练！

残差网络有一下几个优点：

**1. 更容易优化；**

**2. 由于这些微小变化，残差网络在反向传播过程中梯度信息更容易传播。**

1.2.2 循环神经网络(RNN)简述

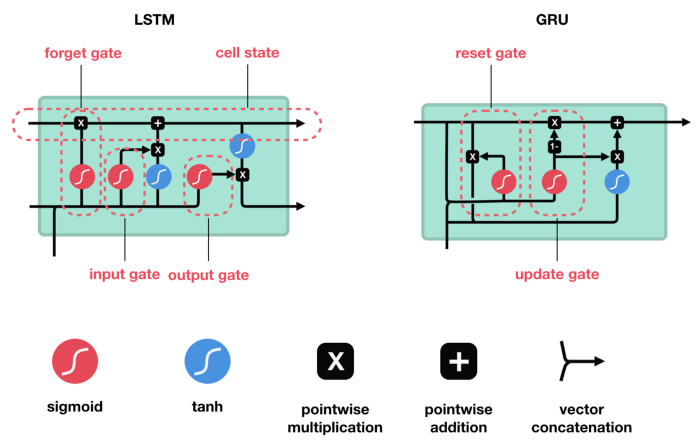


原始循环神经网络如上图， 当前输出的结果不仅由当前输入决定， 还由过去的权重W与输入相乘决定， 这一特性使RNN在处理上下文相关的任务， 如NLP任务时更加方便。 但是原始RNN存在梯度爆炸的问题， 当面对复杂的情况时， 效果不太理想。

1997年， Hochreiter和Schmidhuber两位科学家发明出长短时记忆网络，(LSTM)一举解决这个问题。

相比于原始RNN只有一个隐藏状态， 该隐藏状态只对短期的输入敏感， LSTM增加了一个保存长期的状态， 这也是长短时记忆网络名称的由来。

LSTM网络结构如下图所示：

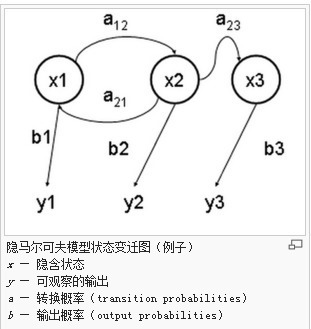


LSTM的主要结构是三个门， 分别是遗忘门(forget gate)、输入门(input gate)、输出门(output gate)。遗忘门决定了上一时刻的单元状态有多少保留到当前时刻；输入门决定了当前时刻网络的输入有多少保存到单元状态。LSTM用输出门来控制单元状态有多少输出到LSTM的当前输出值。

本次实验使用的GRU循环神经网络是LSTM的一种变体，GRU将输入门、遗忘门、输出门变为两个门：更新门（Update Gate）和重置门（Reset Gate）。 它对LSTM做了很多简化，同时却保持着和LSTM相同的效果。因此，GRU最近变得越来越流行。

1.3 HMM模型

HMM（隐马尔可夫模型）是用来描述隐含未知参数的统计模型，举一个经典的例子：一个南京的朋友每天根据天气{下雨，天晴}决定当天的活动{公园散步,购物,清理房间}中的一种，我每天只能在微博上看到她发的推“啊，我前天公园散步、昨天购物、今天清理房间了！”，那么我可以根据她发的微博推断南京这三天的天气。在这个例子里，显状态是活动，隐状态是天气。



hmm模型主要包含以下两个状态， 分别是观测序列（显状态）和隐状态， 分别对应我们的拼音和汉字， 那么我么如何根据隐状态推出显状态呢， 就要根据我们用样本训练的这三个概率，分别是start\_p:初始概率，trans\_p:转移概率，emit\_p: 发射概率。 在我们这次实验中对应概率如下：

obs:观测序列 -> 拼音

states:隐状态 -> 汉字

start\_p:初始概率（隐状态）-> 每个汉字的初始概率

trans\_p:转移概率（隐状态）->当前汉字转换到下一个汉字的概率

emit\_p: 发射概率 （隐状态表现为显状态的概率）-> 汉字转换到对应拼音的概率

有了三个概率后，输入拼音时， 我们使用Viterbi Algorithm，即一个基于图论的求概率最大路径的一种算法， 找出符合的汉字， 最终实现拼音转汉字。

在本次实验中， HMM模型算法主要借鉴github项目：Pinyin2Hanzi，

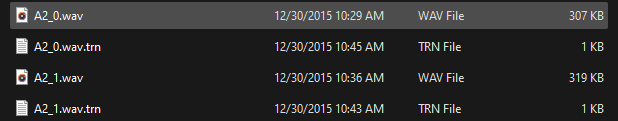
并修改了部分源码。

二、模型训练(基于pytorch)

2.1 数据集

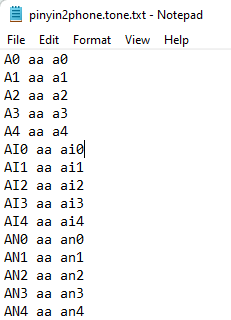
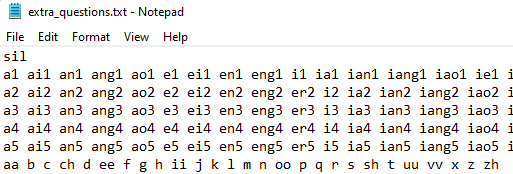
本次数据集使用使用清华大学**THCHS-30**数据集。

格式为音频+标签





同时该数据集为我们提供了220多个音调集合和音调转拼音集合，提供后续实验使用。

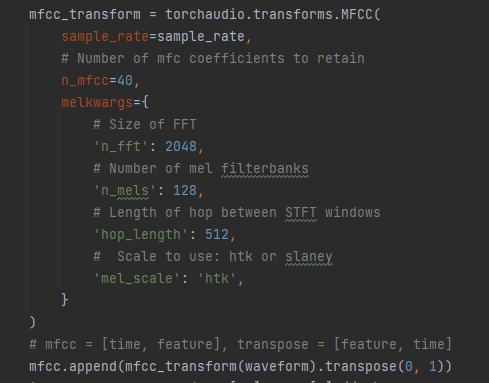
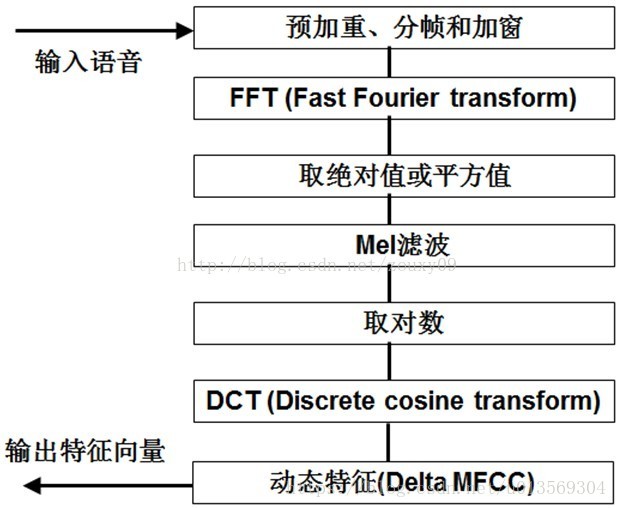


根据pytorch官方教程搭建数据集



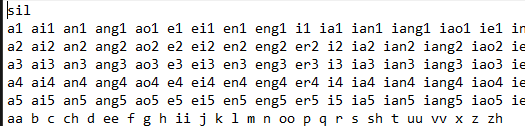
2.2 MFCC系数特征提取

MFCCs中文名为“ 梅尔倒频谱系数 ”（Mel Frequency Cepstral Coefficents）是一种在自动语音和说话人识别中广泛使用的特征。在本次实验中，我们使用torchaudio库中的MFCC提取函数提取特征。

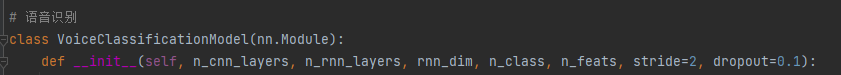


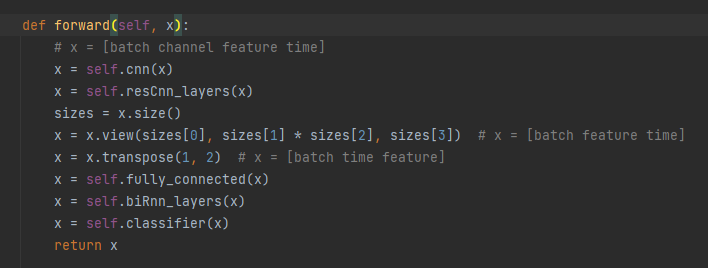
2.3 编码

将音调与数字编码



2.4 搭建神经网络





3 Residual

Convolutional Layer

Linear Layer

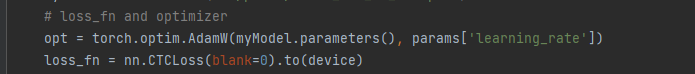
5 Bidirectional GRU Layer

Linear Layer

Softmax

音调概率

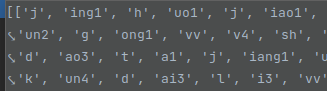
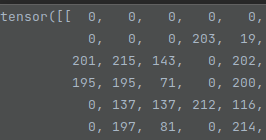
2.5 损失函数与优化器



CTC损失函数： CTC(Connectionist Temporal Classification)能连接时序分类损失，可以对没有对齐的数据进行自动对齐，主要用在没有事先对齐的序列化数据训练上。比如语音识别、ocr识别等等。

优化器： AdamW是Adam优化器的补充版，采用自适应梯度下降的方法，能帮助模型快速收敛，节省计算时间。

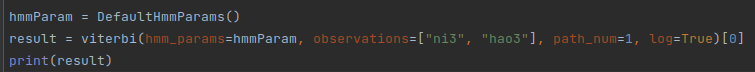
2.6 解码器



数字到拼音

2.7 拼音转汉字

如之前所说，我们拼音转汉字使用HMM模型， 我所使用的github项目中的HMM模型只支持不加音调的拼音转汉字， 但我神经网络输出的结果包含了音调， 删去音调这一信息， 一定会对识别精度造成影响。 所以我根据作者的项目修改了部分源码，使该HMM模型识别加音调的拼音， 同时我更改了训练数据集， 是数据集更贴合日常口语。最终结果如下。





三、模型评估

经过20轮的训练， 最终测试结果稳定在Average loss: 0.1209, Average Chinese CER: 0.475921, Average pinyin CER: 0.120287

四、综合评估

目前模型存在的主要问题有

**1. 拼音误差**

**2. 汉字转拼音误差**

**3. 适用性鲁棒性的问题**

在汇报时，经过和老师同学们的研讨， 我们为目前模型有一下几个改进方案。

1. 放弃HMM模型， 直接使用神经网络端到端的语音识别， 这样能减少汉字到拼音的误差。

2. 扩大训练数据集， 经过和同学们的讨论， 我的数据集过于单一， 鲁棒性不强，所以识别效果不好。

3. 我的模型参考与2014年的一篇论文， 要提高精度， 可以参考使用最近最新的模型。

09019203汪跃阳

2021.12.24