基于深度学习的中文语音识别模型设计与实现

杨焕峥[1,2]

（1.江苏省无线传感系统应用工程技术开发中心，江苏无锡 214153；
2.无锡商业职业技术学院，江苏无锡 214153）

【摘要】创建一个中文语音识别模型，采用近似深度全序列卷积神经网络（DFCNN）、门控循环单元（GRU）和连接时序分类（CTC）结构，将语音信号转换为二维语谱图像信号，将语音转换成中文文本。并在此模型的基础上，通过 Keras 编程搭建 HTTP 协议语音识别服务器，提供语音识别 API，实现语音识别功能。测试结果表明，该方法能获得平均 80%以上的语音识别准确率，取得了较好应用效果。

【关键词】深度学习；语音识别；应用程序接口；Keras 编程
【doi:10.3969/j.issn.2095-7661.2020.03.007】
【中图分类号】TN912.3　　【文献标识码】A　　【文章编号】2095-7661（2020）03-0024-04

Design and Implementation of Chinese Speech Recognition Model Based on Deep learning

YANG Huan-zheng[1,2]

（1.Jiangsu Research and Development Center of Application Technology for Wireless Sensing System, Wuxi, Jiangxi, China 214153; 2.Wuxi Institute of Commerce, Wuxi, Jiangsu, China 214153）

Abstract: A Chinese speech recognition model is created. The structure of approximate deep fully convolutional neural network (DFCNN), gated recurrent unit (GRU) and connectionist temporal classification (CTC) are adopted. The speech signal is converted into two-dimensional speech spectra, converting speech into Chinese text. Based on this model, the speech recognition server of HTTP protocol is built by keras programming. Speech recognition API is provided to realize speech recognition function. The test results show that the method can achieve over 80% in average speech recognition accuracy. Good application results have been obtained.

Keywords: deep learning; speech recognition; application program interface; Keras programming

目前，将人工智能与物联网相结合的 AIoT（AI 人工智能 +IoT 物联网）模式是一个新的发展方向，因为深入学习需要传感器采集数据，物联网系统要求人工智能能识别、检测异常并做出正确预测。而且，当 AI 与 IoT 一体化后，人工智能逐渐向应用智能发展。

深度学习在人工智能领域兴起，对语音识别技术产生了重要影响，深层的神经网络不断替换了之前的 GMM-HMM 等传统语音识别模型。项目建立的语音识别系统声学模型使用近似于深度全序列的卷积神经网络、门控循环单元[1]，语音信号特征提取不采用常规方法而直接采用频谱图作为模型输入，参考图像识别中的最佳网络配置模型深度卷积神经网络 VGG，可以看到较长的时间关联信息，与 RNN（Recurrent Newtral Network，循环神经网络）相比，具有更好的鲁棒性。在输出端，可以与连接时间序列分类（CTC）方案相结合，实现整个模型从输入端到输出端进行训练，将声音波形信号转译成标准汉语的拼音序列，在语言模型中，拼音序列通过隐马尔可夫模型（HMM）

[收稿日期] 2020-05-22
[作者简介] 杨焕峥(1980-)，男，江苏无锡人，无锡商业职业技术学院副教授，硕士，研究方向：嵌入式人工智能。
[基金项目] 2020 年江苏高校“青蓝工程”资助项目“优秀青年骨干教师培养对象”；2018 年无锡商业职业技术学院卓越师资队伍建设项目“校级教学名师培育对象”（项目编号：RS18MP03）；2018 年无锡商业职业技术学院卓越师资队伍建设项目“校级骨干教师培育对象”（项目编号：RS18GG38）。

转换为汉语文本。搭建 HTTP 等协议服务器，提供语音识别应用程序接口 API，STM32 MCU 嵌入式客户端通过 Internet 网络，调用 API 实现语音识别功能。服务器端使用 Python 语言及 Keras、TensorFlow 等库，客户端使用 C 语言进行编程。

1 语音特征提取

通过把普通的 wav 语音信号经分帧、加窗、傅立叶变换、取对数等手段处理，以形成二维语谱图像信号（语谱图、时频图）供神经网络使用。

1.1 获取语音进行数据变换

首先，在服务器上使用 Python 语言编写获取语音信号的函数，用于打开包含语音信号路径的文本文件，读取 wav 格式语音文件的列表，并返回存储列表的字典类型值。其次，编写获取语音符号的函数用于读取与指定数据集内包含的 wav 文件所映射的语音符号，并得到存储符号集的字典类型的值。再次，编写读取语音数据的函数读入音频文件，得到音频信号的收听时间与时域频谱矩阵，编写获取频率特征的函数对数据进行分帧化，帧长度为 25 毫秒，帧移动 10 毫秒，帧添加汉明窗口，并执行傅立叶变换，因为结果数据是对称的，所以可以取一半值使用，形成语谱图的过程如图 1 所示。

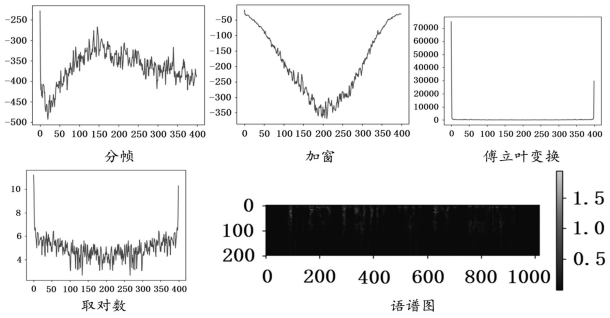

图 1　语音信号经变换得到语谱图过程

1.2 生成神经网络训练矩阵数据

编写获取数据的函数来读入数据并得到能够被用作神经网络模型训练的输入与输出值矩阵。首先，将具有 wav 特性的矩阵输入神经网络。其次，标定类别的矩阵作为神经网络输出值，提取出每个音频文件对应的拼音标签，打开之前生成的拼音词典，通过词典可以将读取到的拼音标签映射成对应的 id 序列。编写获取符号列表的函数加载拼音符号列表，编写符号转数字的函数将符号转为数字。如：准备辞掉一份工作，将拼音标签['zhun3','bei4','ci2','diao4','yi1','fen4','gong1','zuo4'] 转为 [11769 12972 2337 1091 72404 4736 15051 23163]。

2 声学模型搭建

搭建语音识别模型，采用近似 DFCNN+GRU+CTC 的结构，基于 Python 语言 Keras 框架，参考了 VGG 的深度卷积神经网络被用作训练模型[2]。

2.1 模型概述

首先，执行语音信号的时域傅立叶变换以获得语音信号的语谱图，被视为具有特定特征的图像。其次，语谱图直接用作声学模型的输入，输出单元对应于音节识别的最终结果。再次，时间与频率被视为图像的两个维度，通过结合卷积神经网络 CNN 更多的卷积层和池化层，对整个句子语音进行建模。从输入、模型结构和输出三个方面分析模型的优点：在输入端，常见的语音识别系统的特征提取方法会在傅立叶变换后使用滤波器组，造成语音信号高频区信息的明显丢失。此外，常见的语音特征提取会使用很大的帧移位来减少计算复杂度，这样会对语速较快的语音信号造成时域中的信息丢失[3]。模型以语谱图为输入解决了频域与时域的信息丢失问题。为了提高 CNN 的能力，模型结构参考了图像识别中的最佳网络配置 VGG[4]，为了能够表现语音信号的时间相关性，模型积累较多卷积层与池化层，结合 GRU 以看到一段较长的时间信息，具有较强的鲁棒性。在输出端，模型具有灵活性，可以容易地与 CTC 方法进行集成，从而实现对整个模型端到端的声学训练。

2.2 模型框架

定义近似 DFCNN+GRU+CTC 的模型，由输入层、大量卷积与池化层、全连接层、GRU 层、CTC 层、输出层等组成，如图 2 所示。输入层：200 维的特征值序列，一条语音数据的最大长度设为 1600（大约 16 s）；卷积与池化层：卷积核大小采用 3×3，池化窗口大小采用 2×2；全连接层：神经元数量为 self.MS_OUTPUT_SIZE，使用 softmax 作为激活函数；CTC 层：使用 loss 作为损失函数，实现连接性时序多输出。

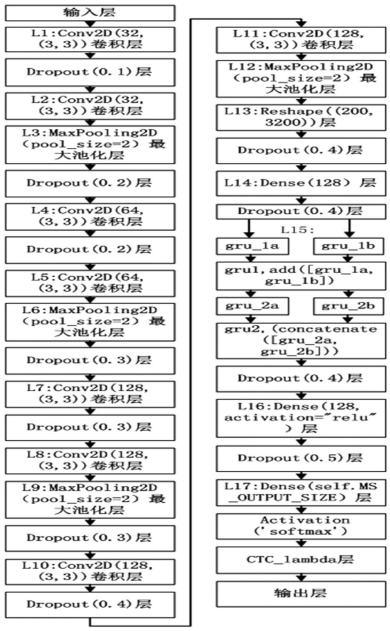

图 2　近似 DFCNN+GRU+CTC 的语音识别模型图

Convolution2D 层主要在二维输入上执行滑动窗口卷积运算，MaxPooling2D 层主要对空间域信号使用最大值池化，并且池化层用于得到最大值和局部平均值，Dropout 层用于预防当不断刷新参数时发生过度拟合的情况，随机中断连接某些数量的输入神经元，Reshape 层用于把输入 shape 转化为特定 shape，GRU 层进行记忆，Dense 层进行全连接。L1、L2：一个二维卷积层，包含 32 个卷积核，大小为 3×3，不使用偏置项，激活函数为线性整流函数“ReLU”，补 0 策略为“same”，即保留边界处的卷积结果，权值初始化方法为正态分布初始化方法“he_normal”。L3：一个二维最大池化层，池化窗口大小为 2，下采样因子为“None”，即默认值为 pool_size，补 0 策略为“valid”，即可能舍弃边上的某些元素。L4、L5：一个二维卷积层，包含 64 个卷积核，大小为 3×3，不使用偏置项，激活函数为线性整流函数“ReLU”，补 0 策略为“same”，即保留边界处的卷积结果，权值初始化方法为正态分布初始化方法“he_normal”。L6：一个二维最大池化层，池化窗口大小为 2，下采样因子为“None”，即默认值为 pool_size，补 0 策略为“valid”，即可能舍弃边上的某些元素。L7、L8：一个二维卷积层，包含 128 个卷积核，大小为 3×3，不使用偏置项，激活函数为线性整流函数“ReLU”，补 0 策略为“same”，即保留边界处的卷积结果，权值初始化方法为正态分布初始化方法“he_normal”。L9：一个二维最大池化层，池化窗口大小为 2，下采样因子为“None”，即默认值为 pool_size，补 0 策略为“valid”，即可能舍弃边上的某些元素。L10、L11：一个二维卷积层，包含 128 个卷积核，大小为 3×3，不使用偏置项，激活函数为线性整流函数“ReLU”，补 0 策略为“same”，即保留边界处的卷积结果，权值初始化方法为正态分布初始化方法“he_normal”。L12：一个二维最大池化层，池化窗口大小为 1，下采样因子为“None”，即默认值为 pool_size，补 0 策略为“valid”，即可能舍弃边上的某些元素。L13：一个 Reshape 层，将输入转换为 200×3200 的 feature_map。L14：一个全连接层，输出维度 128，不使用偏置项，激活函数为线性整流函数“ReLU”，权值初始化方法为正态分布初始化方法“he_normal”。L15：一个门控循环单元 gru1 层，将 gru_1a 和 gru_1b 相加。一个门控循环单元 gru2 层，将 gru_2a 和 gru_2b 连接。L16：一个全连接层，输出维度 128，不使用偏置项，激活函数为线性整流函数“ReLU”，权值初始化方法为正态分布初始化方法“he_normal”。L17：一个全连接层。利用 CTC 实现序列学习，构建在神经网络顶层的损失函数，Keras 目前不支持具有额外参数的损失函数，所以 CTC 损失在 lambda 层中实现。声学模型初始化，默认输出的拼音的表示大小是 1422，即 1421 个拼音加 1 个空白块，设置神经网络最终输出的每一个字符向量维度的大小及一次训练的 batch，标签字符串的最大长度 64，音频长度 1600，音频特性长度 200，利用函数启动模型训练。

2.3 模型特点

深层的 CNN 结构可以提高 CTC 语音识别系统的性能。然而，只使用深层 CNN 进行端到端建模的性能相对较差，因此将 CNN 和 GRU 的记忆相结合，并使用 3×3 小卷积核提高模型的性能。卷积神经网络层数的增加和滤波器数量的增多将明显对该模型的建模能力带来改变，故根据不同规模的语音训练数据库，使用不同的深层 CNN 模型配置来获得最佳的性能。语音识别系统声学模型的输出过程，存在许多接连的相同符号。因此，CTC 用于将连接的多个输出时序序列映射到一个输出，并且移除静音分离标记，得到最终的实际语音字母符号序列，在 CTC 层中，利用 loss 作为损耗函数。

借鉴图像识别中效果很好的网络配置 VGG，一方面加深网络层数，另一方面为了避免参数过多，把大型二维卷积核在计算链路中变为两个连续的 3×3 小卷积核，计算次数大大减少，计算复杂度降低。例如，把 5×5 卷积用两个连续的 3×3 卷积替换，计算次数从 25 次减少到 18 次。并且不是在每个卷积层后面跟上一个池化层，总数 4 个池化层，分布在不同的卷积层之下，而全连接层有 3 层。

3 语言模型搭建与语音识别测试

统计语言模型用于将拼音转换为最终文本以输出，本质被建模为隐含马尔可夫链，模型具有较高的精度。当传入的参数没有包含任何拼音时，返回为空。先取出一个字，即拼音列表中第一个字，依次从第一个字开始每次连续取两个字拼音，如果这个拼音在汉语拼音状态转移字典里的话，将第二个字的拼音加入，否则不加入，然后将现有的拼音序列进行解码，再重新从 i+1 开始作为第一个拼音。语音解码开始，如果这个拼音在汉语拼音字典里的话，获取拼音下属的字的列表，ls 包含了该拼音对应的所有的字。在第一个字中，设置 HMM 的初始状态值，将初始概率设置为 1.0，并将其添加到可能的句子列表中，否则，开始处理紧跟在第一个字后面的字。把现有的每一条短语取出来，尝试按照下一个音可能对应的全部的字进行组合，取出用于计算的最后两个字，判断它们是不是在状态转移表里，在当前概率上乘转移概率，公式化简后为第 n-1 和 n 个字出现的次数除以第 n-1 个字出现的次数，大于阈值之后保留，否则丢弃，最后对语言模型初始化。

语音数据集采用 ST-CMDS-20170001_1-OS 等，对模型加以大数据训练后，抽取若干语音数据进行测试，平均准确率超过 80%，效果较好，如表 1 所示。

表 1　语音识别测试表

序号	wav 语音文件名	语音原意	识别结果	准确率
1	20170001P00178I0104	向日葵永不落泪无所谓	上日葵永不落泪如所谓	80%
2	20170001P00178I0109	我回家给你们这群可怜娃带好吃的	我回家给你没这群可怜娃带好吃的	93.3%
3	20170001P00178I0113	你准备在你哥那里玩好久啊	你准备在你咯那的玩都久啊	75%
4	20170001P00178I0117	准备辞掉一份工作	准备辞掉一份工作	100%
5	20170001P00179A0001	俺家黑娃今天过生日呢	俺家黑娃惊天过生日	80%

4 语音识别 API 服务器与 STM32 MCU 客户端

使用 Python 语言编程实现基于 HTTP、TCP/IP 等协议的语音识别 API 互联网服务器，通过将声学模型与语言模型连接起来，使用服务器程序，提供 REST API 服务，通过 POST 与嵌入式 STM32 MCU 客户端进行 JSON 数据交换[5]。客户端包含有液晶屏显示、录音、语音识别 API 调用、通信等功能，STM32 MCU 通过 SPI 接口与高性能音频编解码 VS1053 模块相连，由 VS1053 模块所接麦克风对人的语音进行录音，保存到 SD 卡，并通过 WiFi 模块经互联网上传到语音识别 API 服务器，服务器将 wav 格式录音进行语音识别后，反馈的中文文本信息通过 WiFi 传输到 STM32 MCU 的 LCD 屏幕上显示。VS1053 语音模块支持 PCM 和 IMA-ADPCM 算法的 wav 音频录制，PCM 抽样声音数据并直接存储，IMA-ADPCM 采用 4:1 的数据压缩算法。当启用 STM32 MCU 时，先检测字库与 SD 卡，SD 卡文件根目录是否有录音文件夹，按下录音按键 VS1053 进入录音过程，松开按键会停止录音并保存该文件，可以在液晶屏上看到录音文件名和时间，STM32 MCU 由 WiFi 模块经互联网连接语音识别 API 服务器，上传录音文件，然后获取反馈的中文文本语音识别结果。

5 小结

系统算法模型在测试集上已经获得了约 80%的语音识别准确率，通过基于深度学习的中文语音识别服务器端与嵌入式 STM32 MCU 客户端的组合，取得了较好应用效果，进一步提高声学模型训练集的数据量，加深 DFCNN 层数以增加语音识别准确率。服务器和客户端的应用增加密钥机制，采用多个 GPU 训练模型，增大语音数据库标本，通过使用并行训练解决神经网络训练效率，使得声学模型更加完善。STM32 MCU 客户端通过互联网连接该语音识别模型的服务器，通过服务器 REST API 给予的标准通信端口，以 JSON 数据格式发送语音数据给服务器并接收识别结果，使得应用更加完善。

【参考文献】

[1]姚煜.基于双向长短时记忆—联结时序分类和加权有限状态转换器的端到端中文语音识别系统[J].计算机应用,2018(9):2495-2499.

[2]张建华,孔繁涛,吴建寨,等.基于改进 VGG 卷积神经网络的棉花病害识别模型[J].中国农业大学学报,2018(11):161-171.

[3]胡亚军.基于神经网络的统计参数语音合成方法研究[D].合肥:中国科学技术大学,2018.

[4]冯陈定.基于改进卷积神经网络与动态衰减学习率的环境声音识别算法[J].科学技术与工程,2019(1):177-182.

[5]杨焕峥,杨国华,徐玲.基于百度 AI 与 STM32 的人脸、语音与物体识别系统研究[J].湖南邮电职业技术学院学报,2018(4):28-30,36.