

基于卷积神经网络的语音识别人机交互系统设计

樊静

中国联合网络通信有限公司软件研究院

DOI:10.12238/deitar.v2i3.9178

[摘要] 研究针对传统语音识别系统准确率低、对环境条件依赖性强等问题,提出了基于卷积神经网络(CNN)的语音识别人机交互系统。系统利用CNN对语音信号进行特征提取与分类,采用多层卷积与池化结构,结合全连接层进行分类,在FPGA上映射CNN算法,旨在提高语音识别的准确性及人机交互的流畅性。实验结果表明,该系统具有较高的识别准确率与良好的稳定性,低延时、低功耗,可满足智能语音交互需求。

[关键词] 卷积神经网络; 语音识别; 人机交互; FPGA

中图分类号: TN912.34 文献标识码: A

Design of speech recognition human-machine interaction system based on convolutional neural network

Jing Fan

China United Network Communications Co., Ltd. Software Research Institute

[Abstract] A speech recognition human-machine interaction system based on Convolutional Neural Network (CNN) is proposed to address the issues of low accuracy and strong dependence on environmental conditions in traditional speech recognition systems. The system uses CNN for feature extraction and classification of speech signals, adopts a multi-layer convolutional and pooling structure, combined with fully connected layers for classification, and maps the CNN algorithm on FPGA, aiming to improve the accuracy of speech recognition and the smoothness of human-computer interaction. The experimental results show that the system has high recognition accuracy and good stability, low latency, low power consumption, and can meet the requirements of intelligent voice interaction.

[Key words] Convolutional neural network; Speech recognition; Human computer interaction; FPGA

随着科技的飞速发展,人机交互系统逐渐从传统的键盘、鼠标等输入方式,转向更为自然、便捷的语音交互方式。语音识别技术作为人机交互的关键环节,其性能直接关系到人机交互效率及用户体验。近年来,随着智能家居、智能车载、智能客服等的快速发展,用户对于智能语音交互的需求不断上升。传统的语音识别技术往往受限于识别准确率、环境噪声等因素,难以满足复杂场景下的应用需求。卷积神经网络(CNN)作为深度学习的代表性算法,具有强大的特征提取和分类能力。其独特的卷积层结构能够有效地提取语音信号中的时频特征,通过多层网络的逐层抽象与转换,实现对语音信号的自动识别、分类。因此,将CNN应用于语音识别领域,有望提升语音识别的准确性、鲁棒性,为用户提供更加便捷、高效的语音交互体验。

1 CNN的基本原理与优势

CNN是特殊的深度学习架构,其设计初衷是为了处理具有网格状拓扑结构的数据,如图像数据。通过引入卷积操作,CNN能够

高效地捕捉输入数据的局部特征,并在多层网络结构中逐步抽象出更高级别的特征表示,使得CNN在图像识别、自然语言处理等领域取得了显著成效。在语音识别领域,CNN通过卷积操作和池化操作,能够自动从原始语音信号中提取出有用的特征,无需人工干预,大大提高了特征提取的效率与准确性。其次,CNN能够处理语音信号的时空特性,通过卷积核对语音信号进行局部感知和参数共享,能够捕获语音信号中的局部特征和全局特征,从而更好地理解语音信号的本质。另外,由于语音信号往往受到噪声、口音、语速等多种因素的影响,传统的语音识别方法在复杂环境下往往表现不佳。而CNN通过训练大量的数据,可学习语音信号的内在规律及变化趋势,在复杂环境下也能保持较高的识别准确率。

2 基于CNN的语音识别模型设计

2.1 CNN模型设计

作为前向反馈神经网络,CNN主要包括输入层、卷积层、池

化层等，此次研究引入CNN LeNet5网络模型，以识别输入信号，其架构如图1所示。

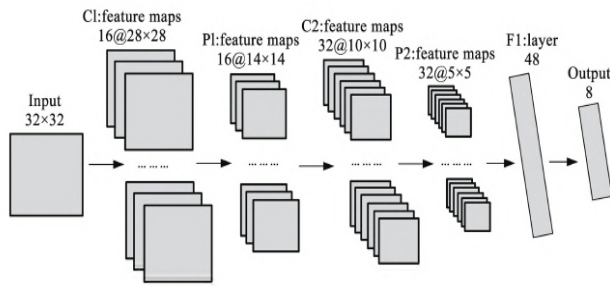


图1 语音识别系统网络结构示意图

在局部特征提取环节采用的是卷积运算，输入特征经过池化运算实现降维处理，将其输入全连接层，经Softmax函数可获得不同类别概率分类结果。函数公式如下：

$$P(i) = \frac{\exp(w_i^T x)}{\sum_{k=1}^K \exp(w_k^T x)}$$

其中，K表示的是类别个数， $w_i^T x$ 为卷积层输出，P(i)在0~1范围内取值。该模型激活函数后，上一层将会输出映射，并将其作为下一层的输入函数（非线性），常用函数类型包括ReLU、Tanh，前者具有计算简单、轻量化的优势，因此选择该激活函数，即

$$\text{ReLU}(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$

系统硬件的执行响应速度、资源消耗量很大程度上由模型参数量决定，本研究采用轻量化模型，能够在保证准确性基础上，最大程度上减少硬件资源的消耗，有利于硬件优化。网络模型参数如表1所示，共计51984个权重。对权重参数、特征图参数进行转化处理，表示为16bit定点数，然后再进行FPGA计算。该过程所消耗的存储资源约105K。

表1 模型网络参数表

网络层	输出尺寸	卷积核大小	权重参数尺寸	权重参数量
Input	32×32×1			
C1	28×28×16	5×5/1	5×5×16×1	400
P1	14×14×16	2×2/2		
C2	10×10×32	5×5/1	5×5×32×16	12800
P2	5×5×32	2×2/2		
F1	48×1×1	5×5	5×5×32×48	38400
Output	8×1×1	1×1	48×8	384

2.2 语音信号数据集准备

2.2.1 数据说明

语音信号数据集包含大量的语音样本，每个样本均为一段包含特定语音内容的音频文件。音频文件来自不同说话者、不同语音环境以及不同的语音内容。在选择数据集时，需要考虑以下因素：①数据集规模。足够大的数据集可以帮助模型学习到更多的语音特征和模式，从而提高识别准确率。在选择数据集时，应优先考虑规模较大的数据集。数据集应包含丰富的语音内容，涵盖不同的词汇、短语、句子，以便模型能够处理各种语音输入。数据集应具有多样性，包括不同的说话者、语音环境、语速、音调等，以增强模型的泛化能力。常用的语音信号数据集包括TIMIT、LibriSpeech、WSJ等。这些数据集都经过精心设计、标注，适合用于训练和评估语音识别模型。

2.2.2 数据预处理

数据输入语音识别系统前，应完成数据的预处理，提取有用的特征并消除噪声。使用librosa、scipy等库读取音频文件，将其转换为数字信号。将所有音频文件的采样率统一到相同的值，以便模型能够处理统一的输入数据。将连续的语音信号分割成较短的时间帧，提取局部特征。通常使用滑动窗口的方式进行分帧。从每个时间帧中提取有用的语音特征，如梅尔频率倒谱系数(MFCC)、线性预测系数(LPC)等。这些特征能够反映语音信号的频谱信息和时域信息，有助于模型识别语音内容。然后对提取的特征进行归一化处理，消除不同特征量纲差异，使模型更容易学习。对于长度不一的语音序列，可以通过填充或截断的方式将其调整为统一的长度，以适应模型的输入要求。经过上述预处理步骤后，语音信号数据集将被转换为适合用于模型训练的格式，为后续的训练和评估提供便利。

2.3 CNN模型的电路设计

HLS能够对C/C++代码进行转换，实现卷积与池化模块，经过逻辑综合得到网表，生成硬件电路。此次研究采用HLS，编写C++代码，将IP核导出。卷积操作的数学表达式如下：

$$[y_{i,j}] = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} w_{m,n} \cdot x_{i+m,j+n} + b$$
。其中， $(y_{i,j})$ 是输出特征图在位置 $((i,j))$ 的值， $(w_{m,n})$ 是卷积核在位置 $((m,n))$ 的权重， $(x_{i+m,j+n})$ 是输入特征图在位置 $((i+m,j+n))$ 的值， (b) 是偏置项， (M) 和 (N) 分别是卷积核的行数和列数。在硬件实现中，卷积层IP核需要能够并行处理多个卷积操作，加速计算过程。此外，还需要考虑数据的存储和访问方式，以优化内存带宽和降低功耗。

池化层用于对卷积层的输出进行下采样，以减少数据的空间尺寸和计算复杂度。常见的池化操作包括最大池化和平均池化。最大池化的函数为：

$$[p_{i,j}] = \max_{m=0}^{M-1} \max_{n=0}^{N-1} x_{i \cdot S + m, j \cdot S + n}$$

其中， $(p_{i,j})$ 是池化后特征图在位置 $((i,j))$ 的值， $(x_{i \cdot S + m, j \cdot S + n})$ 是输入特征图在池化窗口内的值， (S) 是池化步长， (M) 和 (N) 分别是池化窗口的行数和列数。在硬件实现中，池化层IP核需要能够高效地执行池化操作。通过设计专用

的比较器和选择器电路来实现。同时,需要考虑如何与卷积层IP核进行数据传输和同步,确保整个CNN模型的流畅运行。

3 基于CNN的语音识别人机交互系统设计与实现

3.1 人机交互界面设计

人机交互界面是用户与系统进行交互的桥梁,其架构见图2。在基于CNN的语音识别人机交互系统中,采用直观、简洁的界面设计风格。界面布局上,采用常见的窗口形式,将主要功能区域划分为输入区、显示区和控制区。输入区用于接收用户的语音输入,可以是一个麦克风图标或按钮,用户点击后开始录音并上传语音数据。显示区用于展示CNN语音识别模块的输出结果,即识别出的文本内容。控制区则包含一些常用的控制按钮和设置选项,如开始/停止录音、清空输入、调整识别参数等。在交互方式上,用户只需通过简单的点击和语音输入即可完成大部分操作,无需复杂的操作步骤或参数设置。同时,提供了辅助性的提示与反馈信息,以帮助用户更好地理解、使用系统。

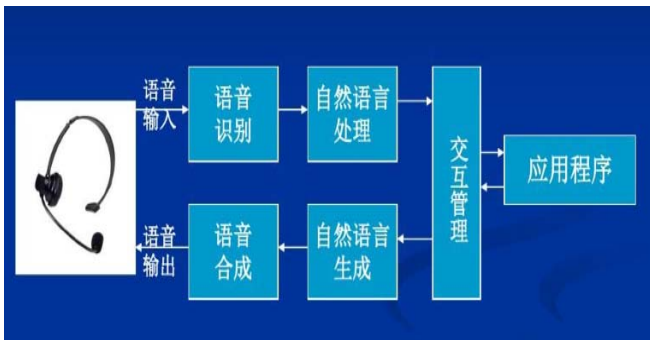


图2 语音人机交互架构图

3.2 语音输入与输出模块实现

语音输入模块主要负责接收用户的语音输入,并将其转换为计算机可以处理的数字信号。采用麦克风作为语音输入设备,通过录音软件或API接口获取原始的语音数据。为了保证语音信号的质量,需要对录音设备进行合理配置,设置合适的采样率和量化精度等。对原始语音数据进行预处理,如噪声抑制、分帧和加窗等,提取出语音信号的关键特征。语音输出模块采用文本转语音(TTS)技术将识别出的文本转换为语音信号。可以根据实际需求选择合适的TTS技术,并将其集成到系统中。通过调用TTS引擎或API接口,将识别出的文本转换为语音信号并播放给用户。

4 系统测试

采用了单元测试、集成测试、压力测试及性能测试,准备了丰富的测试数据集,包括不同场景、不同人物的语音样本,以确

保测试的全面性和准确性,测试内容及结果见表2:

表2 测试内容及结果分析

测试项目	测试方法	测试数据规模	预期结果	实际结果	偏差	结果分析
语音识别准确率	不同年龄、性别、地域的800段清晰语音样本	800段	准确率在95%以上	98%	2%	系统的识别准确率达到了95%以上,相较于传统语音识别方法有了显著提升。
响应速度	统计1000次语音输入的响应时间	1000次	平均响应时间在1s以内	0.5s	+0.5s	从接收到语音输入到输出识别结果的时间平均0.5秒,满足了实时交互的需求。
抗噪性能	在40dB、60dB、80dB噪声环境下分别测试300段语音	900段	在60dB噪声下准确率不低于85%	在60dB噪声下准确率为80%	-5%	模型对高强度噪声的处理能力有待加强,需改进降噪算法。
用户体验满意度	对200名用户进行问卷调查,1-10分评分	200份	平均满意度达到9分以上	9.5分	0.5分	多数用户对交互体验表示满意,认为系统易于使用、反应迅速且识别准确。

5 结束语

综上所述,基于CNN的语音识别人机交互系统具有较高的识别准确率、良好的实时性以及较好的交互体验,能够有效地实现人机交互,为用户提供便捷、高效的语音输入方式。未来,将进一步优化模型结构、提升性能,并探索更多应用场景,以满足不同用户的需求。

【参考文献】

- [1]彭凯贝,孙小明,陈皓炜,等.基于卷积神经网络的火车站语音情感识别方法[J].计算机仿真,2023,40(2):177-180,189.
- [2]李晶,邓伟,杨勇,等.基于多通道卷积神经网络的语音隐写分析方法[J].电子器件,2022,45(5):1105-1109.
- [3]乔栋,陈章进,邓良,等.基于改进语音处理的卷积神经网络中文语音情感识别方法[J].计算机工程,2022,48(2):281-290.
- [4]赵小芬,张开生.基于三层结构优化卷积神经网络的语音识别[J].石河子大学学报(自然科学版),2022,40(1):127-132.
- [5]杨帆,李祎男,乔涵.基于深度卷积神经网络的语音信号去噪关键技术研究[J].计算机与数字工程,2022,50(2):344-349.
- [6]刘虹,袁三男.基于多尺度残差深度卷积神经网络的语音识别[J].计算机应用与软件,2020,37(11):275-279.
- [7]张琨瑶,王华朋,牛瑾琳,等.基于深度卷积神经网络的语音降噪研究[J].刑事技术,2021,46(5):457-463.

作者简介:

樊静(1992—),女,汉族,安徽省安庆市人,硕士研究生,单位(中国联合网络通信有限公司软件研究院),职称:中级职称,研究方向:人工智能。