|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **分类号** |  | **密级** |  | |
|  |  |  |  |  | |
|  | | | | | |
| **重庆邮电大学研究生学位论文**  **开题报告** | | | | | |
|  | | | | | |
|  | **中文题目** | **面向车载嵌入式设备的** | | |  |
|  | **智能语音对话方法研究** | | |
| **英文题目** | **Intelligent Voice Dialogue Method** | | |
|  | **for Vehicle Embedded Device** | | |
| **学 号** | **S200303078** | | |
| **姓 名** | **黄子恒** | | |
| **学位层次** | **硕士研究生** | | |
| **学位类别** | **工学硕士** | | |
| **学科专业** | **控制科学与工程** | | |
| **研究方向** | **神经网络理论及应用** | | |
| **指导教师** | **李鹏华 教授** | | |
| **完成日期** | **2022年1月1日** | | |
|  | | | |

|  |
| --- |
| **一、选题依据** |
| **1.1 研究背景、目的和意义**  语音交互作为人机通信中最自然、直接的交互方式，具有天然的优势[1]。随着智能汽车的兴起，车载语音交互成为继通讯社交、智能家居之后的第三大应用场景。截止2020年9月，我国乘用车车载语音装配率为64.8%，预计2025年国内前装车载语音市场规模约32亿元，智能语音交互俨然成为汽车的标准化配置之一[2]。科大讯飞、亚马逊和谷歌等企业也相继推出了自主研发的车载语音对话系统，对车载语音市场的发展进程产生了显著影响[3]。自动语音识别（Automatic Speech Recognition, ASR）和自然语言理解（Natural Language Understanding, NLU）是智能语音交互的核心技术。在深度学习的驱动下，当前的ASR和NLU技术取得了重大进展，模型性能得到极大提升[4]。  然而，高精度的深度学习模型需要消耗庞大的计算资源才能快速运行。受限于车身结构和研发成本，当前主流的车载语音对话系统均采用“云—端”方式运行，即通过互联网将本地收集到的音频数据传输至配置有高性能GPU的云服务器进行处理，再将处理结果通过网络反馈到本地端，最后由汽车中控系统进行相应操作。事实上，此种运行方式在传输过程和云端处理过程中存在着一定的数据安全隐患，例如特斯拉间谍车、滴滴地图数据外泄等事件。为解决上述问题，实现深度学习下的ASR和NLU技术在车载平台上的高可靠性、强实时性应用，研发离线条件下的智能语音对话是有效的技术途径。  本课题面向计算资源有限的车载嵌入式设备，立足于语音识别和自然语言处理领域中已有的先进成果，重点研究针对ASR和NLU的技术创新应用方法，开发相应的轻量化神经网络模型，在满足性能要求的前提下降低其运行所需的计算资源，探索以离线方式运行的车载智能语音对话平台搭建，在本地实现数据安全、自然实时的语音对话。本课题提出的语音对话方法可以推广到智能家居、智慧教育和智慧医疗等其他智能语音交互领域，丰富相关领域的技术理论，实现语音交互的价值。特别地，除了具体的ASR和NLU技术外，本课题中涉及的一些关键共性技术，如模型轻量化、嵌入式移植等，亦可为研发其他依靠深度神经网络的设备提供成熟的案例支撑。  **1.2 国内外研究现状**  （1）自动语音识别技术  自动语音识别技术的发展过程可以分为三个阶段：第一阶段主要使用高斯混合模型-隐马尔可夫模型（Gaussian Mixture Model-Hidden Markov Model, GMM-HMM）进行建模，基于GMM-HMM的语音识别框架在上世纪得到广泛应用[5]；直至2009年，深度神经网络被应用于语音识别任务，并在TIMIT数据集的音素识别任务中取得重大进展和突破[6]。大量研究人员投入到深度学习研究中，并取得较大的成果[7]。至此，GMM-HMM框架被打破，进入到以深度神经网络-隐马尔可夫模型（Deep Neural Network- Hidden Markov Model, DNN-HMM）为主要方法的第二阶段；第三阶段是端到端（End-to-End, E2E）的时代，基于E2E的语音识别普遍采用深度学习的方法，它不需要提前进行语音帧对齐，而是直接采用带标签的语音进行训练。与经典方法相比，它更加简洁且具有较强的通用性，能够减少对专业语音、语言知识的依赖，大大降低了系统搭建难度。总体可分为两类：一类是基于联结时序分类（Connectionist Temporal Classification, CTC）[8]的E2E模型，另一类是基于注意力机制（Attention）的序列到序列（Sequence-to-Sequence, S2S）模型[9]。CTC方法不像经典方法需要对语音数据进行对齐操作，只需计算输出序列和真实序列的差距；而S2S方法在2017年被Google用于语音识别领域，并取得了非常好的效果。    （a）基于CTC的E2E模型 （b）基于Attention的S2S模型  图1.1 E2E模型和S2S模型结构示意图  基于CTC的E2E模型结构如图1.1（a）所示。在无先验性对齐情况下，该模型能够度量输入和输出序列的相似度，并且能刻画语音特征和字符序列的相关性。基于联结时序分类的语音识别由卷积神经网络（Convolutional Neural Network, CNN）/循环神经网络（Recurrent Neural Network, RNN）编码模块和CTC损失函数模块组成。2006年Alex利用空白字符对不同长度的序列进行对齐，首次提出CTC模型来解决MINIST手写数字识别和TIMIT语料库音素分类的问题[10]。2012年Alex对CTC语音识别模型进一步改进，将RNN作为转化模型[11]。2013年，Alex利用多层的长短时记忆（Long Short-Term Memory, LSTM）神经网络进行建模，进一步提升TIMIT语料库上的识别效果[12]。  基于Attention的S2S模型结构如图1.1（b）所示。该模型主要由编码器（Encoder）、注意力网络和解码器（Decoder）三个部分组成，在2013年由Alex首次提出，并被用于MINIST数据集的手写数字识别任务中[13]。2014年，Bahdanau进一步完善，提出基于Attention的编解码模型[14]。同年，Chorowski将注意力模型进一步应用于语音识别任务中，并在目标函数中加入约束，在TIMIT数据集上取得较好的效果[15]。至此，大量的研究人员开始投入到基于Attention的语音识别研究中，并提出多种改进模型。根据改进方法可分为三类：对Encoder的改进、对Attention的改进以及对语言模型的改进。对Encoder的改进方面，Bahar采用更复杂的二维长短时记忆网络（Two-Dimensional Long Short Term Memory, 2DLSTM）作为编码模型，从多维的角度增加语音时序信息的辨认能力[16]。近几年，研究者们热衷于研究基于深度自注意力网络（Transformer）的ASR系统，Transformer最先由Vaswani提出[17]，并被用于建立语言模型和机器翻译，由于Transformer采用全连接网络，因而相较于RNN等结构，其训练效率高，模型收敛效果好[18]；对Attention的改进方面，Merboldt提出局部注意力机制，对其添加约束，同时还考虑最大注意力得分，之后采用启发式搜索对模型进行训练，最终在SwitchBoard和LibriSpeech数据集上取得较好的效果[19]；对语言模型的改进方面，除了传统的语言模型，Zeyer采用LSTM作为语言模型，并采用字节对编码（Byte Pair Encoding, BPE）输出不同的识别基元，相比于传统语言模型，取得较好的效果[20]。  基于CTC的E2E模型存在解码过程复杂、速度慢等缺点，而基于Attention的S2S模型具有模型简单、联合训练、直接输出和无需强制数据对齐等优点。但是，该类模型存在参数量庞大、计算复杂、难以部署等缺点。例如，谷歌提出的Conformer模型参数量达到了1188亿[21]。这些因素为其在实际工程中的应用带来了巨大的挑战。因此，如何降低模型的参数量和计算复杂度、提高模型运行速度，在计算资源有限的车载平台上实现高效的语音识别是亟待解决的问题，也是学术界当前研究的热点。  （2）自然语言理解技术  自然语言理解通常包含意图检测（Intent Detection, ID）和槽位填充（Slot Filling, SF）两个子任务，前者属于文本分类（Text Classification）领域，而后者属于序列标注（Sequence Labeling）领域。自然语言理解技术的发展过程从建模方法上可以分为两个类别：基于独立建模的方法和基于联合建模的方法。  基于独立建模的方法如图1.2（a）所示。在意图检测方面，Xu利用CNN提取5-gram特征并应用最大池化层（Max Pooling）获取文字表示，取得了较好的效果[22]。Ravuri成功地将RNN和LSTM应用到意图检测中，表明序列特征有利于提升准确率[23]；在槽位填充方面，常用的方法包括条件随机场（Conditional Random Fields, CRF）、RNN和基于RNN的衍生模型。Yao采用基于RNN的语言模型来预测槽位标签，此外还研究了命名实体、句法特征和词类信息[24]。Mesnil研究了不同的用于NLU的RNN衍生模型，包括Elman RNN、Jordan RNN和双向Jordan RNN[25]。Mesnil利用Viterbi编码方式和循环CRF层消除了标签的预测偏差问题[26]。基于独立建模的方法虽然取得了不错的效果，但由于单独训练，单个模型的意图检测任务和槽位填充任务之间没有交互作用，存在共享信息泄漏，导致模型性能不足。    （a）独立建模 （b）隐式联合建模 （c）显式联合建模  图1.2 NLU的两种建模方式的结构示意图  基于联合建模的方法从两个子任务的交互方式可分为隐式交互和显式交互，分别如图1.2（b）和（c）所示。二者的区别在于：隐式联合建模仅采用共享编码器（Shared-encoder）来捕获特征，后续不再有任何交互；而显式联合建模在捕获特征后依旧通过交互模块进行交互，具有控制交互过程的特性。Zhang引入了共享RNN模型来捕获意图和插槽之间的相关性[27]。Hakkani-Tur提出了一种用于联合建模的共享RNN-LSTM架构[28]。隐式联合建模是一种直接整合共享信息的方法，但后续未有交互进行显式建模，导致模型可解释性不足，性能未达到预期。而越来越多的研究者提出通过显式交互建模的方法。Goo提出Slot-Gated模型，该模型允许槽位填充可以根据学习到的意图设置条件[29]。Li提出了一种新的Self-Attention模型，通过意图来引导槽位填充[30]。Qin提出Stack-Propagation模型，直接使用意图检测结果来指导槽位填充，并使用字符级（Token-level）信息来缓解二者间的误差传递[31]。E提出了一种新颖的SF-ID模型，为意图和槽位提供了双向关联机制[32]。Zhang引入动态路径胶囊网络（Capsule-NLU），将两个任务之间的分层和相关关系纳入其中[33]。Liu提出了一种新型的协同记忆网络（CM-Net）用于联合建模意图检测和槽位填充[34]。Zhang将Graph LSTM结构引入NLU进行了探索，取得了很好的性能[35]。Qin通过建立两个任务之间的双向连接，提出了一种考虑交叉影响的Co-interactive Transformer模型[36]。近年来，以BERT（Bidirectional Encoder Representation from Transformers）[37]为代表的预训练模型（Pre-trained Language Models, PLMs）在各种NLP任务中取得了振奋人心的结果。Chen使用BERT提取共享文本特征用于意图检测和槽位填充，取得了显著的性能提升[38]。  基于隐式联合建模的方法只是通过共享字符编码来隐式地考虑两个任务之间的相互连接，而基于显式联合建模的方法构建了交互模块，子任务间存在信息交互通道，使得模型性能有了进一步的提高。但是该类模型在交互模块中没有高效的交互信息融合手段，导致模型泛化能力不足。虽然可以在特征提取阶段引入预训练模型补充丰富的语义特征，但预训练模型的参数量往往非常惊人。以谷歌推出的基础模型Bert-base-uncased[37]为例，模型网络层数为12，隐含层维度为768，拥有12个注意力头，参数量达到1.1亿，如此庞大的模型难以部署于车载平台（嵌入式设备）上。因此，如何在不引入训练模型的前提下，构建意图和槽位的双向交互通道，高效融合所有的交互信息，深入挖掘跨任务的语义特征，提升模型的泛化性能，在计算资源有限的车载平台上实现高效的对话理解是亟待解决的问题，也是学术界当前研究的热点。  **1.3 当前存在的主要问题**  综上所述，将深度学习技术应用于车载智能语音对话领域取得了很好的效果，各种采用不同类型深度神经网络的模型使得预测结果的精度得到极大提升[4]，但针对自动语音识别和自然语言理解方法的研究还存在以下问题和难点：  **（1）****基于深度编—解码器的自动语音识别模型参数量庞大**  基于深度编—解码器的自动语音识别模型含有编码器和解码器两个结构。在这两个结构中都存在多头注意力模块（Multi-Head Attention），其本身由多层注意力网络堆叠而成，因此模型内部参数庞大、网络结构复杂；  **（2）基于显式联合建模的自然语言理解模型泛化能力不足**  基于显式联合建模的自然语言理解模型含有交互模块，用于捕捉两个子任务间的共享信息。但交互模块中没有高效的交互信息融合手段，导致模型对跨任务语义特征的利用率较低，因此模型的泛化能力不足；  **（3）“云—端”方式运行的车载智能语音对话存在数据安全隐患**  基于深度学习方法的智能语音对话需要庞大的计算资源作为支撑，但汽车受限于内部空间和研发成本无法搭载高性能处理器，所以当前车载语音对话普遍采用“云—端”方式，即在本地收集数据后传输至搭配有高性能GPU的云端服务器上进行处理。这种方式在传输过程和处理过程中存在着一定的数据安全隐患。 |
| **二、研究目标和主要研究内容** |
| **2.1 研究目标**  （1）针对基于深度编—解码器的自动语音识别模型参数量庞大的问题，探索基于残差分组线性变换的解码器结构。给出残差分组线性变换的具体形式，建立与注意力机制的拓扑关系，降低模型的参数量和计算复杂度；  （2）针对基于显式联合建模的自然语言理解模型泛化能力不足的问题，探索基于层级交互的全局—局部交互网络。给出全局—局部交互网络的具体结构，设计高效的层级信息融合算法，提高模型的泛化能力和预测精度；  （3）针对“云—端”方式运行的车载智能语音对话存在数据安全隐患的问题，探索面向车载嵌入式设备的本地语音对话系统集成与验证。针对嵌入式设备进行模型训练、移植和集成，提出面向实际应用场景的适配性优化方案，搭建离线运行的车载智能语音对话平台。  **2.2 主要研究内容**  **（1）****基于残差分组线性变换解码器的自动语音识别**  研究基于残差分组线性变换的解码器结构，首先引入残差分组线性变换方法，设计含有扩张和收缩操作的缩放单元，然后建立逐块缩放机制并将缩放单元放置于解码器自注意力模块前，最后耦合词嵌入层（Embedding Layer）和前馈网络（Feedforward Network, FFN），实现自动语音识别模型参数量和计算复杂度的降低；  **（2）基于全局—局部交互网络的自然语言理解**  研究基于层级交互的全局—局部交互网络，首先构造局部槽位交互层建立槽位自交互通道，其次构造全局意图—槽位交互层建立意图和槽位的双向交互通道，最后设计层级信息融合算法，高效聚合所有交互信息，实现自然语言理解模型泛化能力和预测精度的提升；  **（3）面向车载嵌入式设备的智能语音对话系统**  研究面向车载嵌入式设备的智能语音对话系统，首先根据车身结构选择嵌入式设备并配置相应的软件环境，其次面向企业和用户需求构建特定场景下的语音、文本数据集，接着训练适合特定场景的自动语音识别、自然语言理解等模型，依据车载嵌入式设备的特性进行模型移植与集成，最后根据实际使用情形提出适配性优化方案并搭建硬件平台，实现数据安全、自然实时的离线智能语音对话。 |

|  |
| --- |
| **三、拟解决关键问题及其研究方法** |
| **3.1 拟解决关键问题**  （1）自动语音识别模型中的解码器结构轻量化  研究自动语音识别模型中的解码器结构轻量化方法，其难点在于：为捕捉更丰富的语音特征，解码器包含多头注意力模块，但其本身由多层注意力网络堆叠而成，导致模型参数庞大、结构复杂。使用低维度的解码器，引入残差分层分组变换对解码器的表达能力进行补充，同时耦合词嵌入层和前馈网络模块，实现解码器结构的轻量化；  （2）自然语言理解模型中的全局—局部交互网络构建  研究自然语言理解模型中的全局—局部交互网络构建方法，其难点在于：为捕获两个子任务间的共享信息，模型中通常含有交互模块，但其中没有高效的交互信息融合手段，对跨任务语义特征的利用率较低，导致模型的泛化能力不足。构建局部槽位交互层和全局意图—槽位交互层，建立意图和槽位间的双向交互通道，设计层级信息融合算法，高效聚合所有交互信息，实现全局—局部交互网络的构建。  **3.2 采取的研究方法**  （1）基于残差分组线性变换解码器的自动语音识别  在基于Transformer的自动语音识别模型中，通常需要将解码器中隐含层神经元个数设置为较大值，然而这样不仅会显著增加模型的参数量，还会导致模型难以训练。在本课题中，构建引入残差分组线性变换的解码器，考虑在减少隐含层神经元个数的同时加深网络深度，进而在不影响模型性能的情况下大幅减少模型的参数量。  残差分组线性变换在分组线性变换的基础上，额外加入了残差连接、分组切分和混合器等相关操作和部件，如图3.2（a）所示。分组切分操作将输入和输出分成组，为所在层数，之后将分组的输入和输出使用混合器混合。以第一组为例，输入的第一组和层输出的第一组将被组合到一起作为第层第一组的输入，以此类推，直到产生组输入。该机制提高了输入特征的复用效率，由于使用稀疏连接，能大幅减少模型的参数量。    （a）残差分组线性变换 （b）“钻石”型缩放单元 （c）逐块缩放机制  图3.2 残差分组线性变换、“钻石”型缩放单元及逐块缩放机制示意图  基于残差分组线性变换，可以形成网络更深、包含扩张和收缩两个阶段的“钻石”型缩放单元，如图3.2（b）所示。从形式上来说，在扩张阶段，组数随着网络深度的加深而变多，总的神经元个数也会变多。而收缩阶段刚好与扩张阶段相反。  一般来说，该单元由5个配置参数决定：深度、宽度因子、输入维度、输出维度和最大组数。在扩张阶段，该单元将维度为输入映射到高维空间（限制最高维度），同时层数将会线性地增加到层。在收缩阶段，将维度为的向量逐渐变换到维度，收缩阶段将使用剩余的层，如下所示：    其中，为一个缩放单元中某一层的输出，为分组线性变换，的输入为输入特征序列或者中间层输出结果，函数首先将对进行切分，然后使用混合器将切分结果与输入特征序列的切分结果进行合并，该过程为一个残差分组线性变换。各层组数的划分依据为下式：    上述为一个块（Block）内部的特征扩张收缩过程，鉴于一些解码器（如Transformer中解码器）通常由若干个块堆叠而成，因此不同的块与块之间也存在深度的概念和逐块缩放机制，如图3.2（c）所示。为了充分提取特征，为各块设定不同深度和宽度因子：      其中，和为第块缩放单元的深度和宽度因子，表示总块数，与为超参数，表示设定的最小深度和最大深度。  将提出的“钻石”型缩放单元放置于解码器各自注意力模块之前，同时将注意力机制的维度变为原来的一半，再将前馈神经网络中间层的神经元个数调整为输入层的，这样便形成了低维度、窄而深的解码器结构，同时参数量大大减少。  （2）基于全局—局部交互网络的自然语言理解  在基于显式联合建模的自然语言理解模型中，通常引入预训练模型为模型提供丰富的语义特征。但预训练模型往往参数量巨大，不利于实际应用。在本课题中，构建基于层级信息融合的全局—局部交互网络，建立意图和槽位的双向交互通道，设计层级信息融合算法，高效聚合所有交互信息，进而提高模型的泛化能力。自然语言理解模型框架及全局—局部交互网络如图3.3所示。  利用双向GRU神经网络和自注意力机制获取字符向量，输入用于意图感知的专属双向GRU神经网络中，获取用于意图检测任务的隐层状态表示，得到位置字符的意图表示，经过意图解码器得到针对输入语句的意图结果并输入至全局—局部交互网络中；将与连接输入用于槽位感知的专属双向GRU神经网络中，获取用于槽位填充任务的隐层状态表示并输入至全局—局部交互网络中，其中。    （a）自然语言理解模型框架 （b）全局—局部交互网络  图3.3 自然语言理解模型框架及全局—局部交互网络结构示意图  如图3.3（b）所示，全局—局部交互网络由局部槽位感知交互层和全局意图—槽位交互层组成。在局部槽位感知交互层中，每个槽位都与其他槽位互相连接，使得可以实现跨槽位的依赖性建模，从而解决槽位不协调问题。每个槽位用一个节点表示，所有节点的第一层状态向量为，且对于节点只有相邻一个窗口大小的节点即与其连接，其中为超参数，表示滑动窗口的尺寸。从第一层逐层往上进行信息融合，第层的融合过程可以定义为如下式子：      其中表示互连槽位的节点集合，为可训练网络矩阵参数，表示激活函数LeakyRelu。叠加层之后，可得到局部槽位感知特征；在全局意图-槽位交互层中，将所有意图和槽位连接起来，实现并行输出。输入的槽位特征为，意图特征为，第一层表示意图和槽位状态节点的状态向量为    信息融合的过程可以表示为    其中表示相互连接的槽位和意图节点集。最终经过层的融合，采取softmax和argmax操作得到槽位填充的预测结果，如下式所示：      其中为可训练网络矩阵参数，为位置字符的槽位填充预测结果。  将设计的全局—局部交互网络放置于槽位填充任务的解码过程之前，建立意图和槽位的双向交互通道，使用层级信息融合算法聚合所有交互信息，提高模型的泛化能力和预测精度。  （3）面向车载嵌入式设备的智能语音对话系统  本课题提出的车载智能语音对话系统来源于本人所参与的企业研发项目。此项目旨在改变传统车载语音交互“云—端”的运行方式，基于车载嵌入式设备搭建硬件运行平台，实现离线运行的智能语音对话。  首先根据车内环境选择Nvidia Jetson TX2作为车载嵌入式设备，开发板如图3.4（a）所示。TX2是基于NVIDIA Pascal™架构的AI（Artificial Intelligence）单模块超级计算机。相较于上一代的TX1，TX2内存和eMMC（Embedded Multi Media Card）提高一倍，CUDA（Compute Unified Device Architecture）架构升级为Pascal，每瓦性能提高一倍，支持更大、更深、更复杂的深度神经网络。为配置TX2软件使用环境，根据官方教程利用SDK Manager软件对TX2进行刷机处理，JetPack版本为4.5.1，刷机操作过程如图3.4（b）所示。    （a）Nvidia Jetson TX2开发板 （b）SDK Manager  图3.4 Nvidia Jetson TX2开发板及刷机操作  然后面向企业和用户需求，构建特定场景下的语音、文本数据集。由于考虑到项目本身所涉及的控制需求较多，若通过线下采集数据的方式（例如利用录音笔录取志愿者的话语），无法将音频数据根据使用功能归类存放。为解决上述问题，决定采取线上采集的方式，使用手机端微信小程序（WeChat Mini Program）设计语音问卷，根据不同的意图来录取志愿者的话语，将音频文件传给后台分类存储形成语音数据集，然后对语音数据集进行人工标注得到相应的文本数据集，这样即可便捷地完成两部分数据集的采集工作。微信小程序使用界面如图3.5（a）所示，开发界面如图3.5（b）所示，录制音频的相关参数如表3.1所示。    （a）小程序使用界面 （b）小程序开发界面  图3.5 录音微信小程序界面  表3.1 录制音频相关参数   |  |  |  | | --- | --- | --- | | 参数 | 定义 | 数值 | | duration | 最大录音时长，超过自动停止（ms） | 600000 | | sampleRate | 采样率（Hz） | 16000 | | numberOfChannels | 录音通道数（个） | 1 | | frameSize | 指定帧大小（KB） | 50 | | format | 音频格式 | mp3 |     （a）驾驶室 （b）后备箱  图3.6 面向车载嵌入式设备的智能语音对话平台  最后根据收集到的特定场景下的语音、文本数据集，在PC平台上训练自动语音识别、自然语言理解等模型并完成算法应用质量验证，依据嵌入式设备特性进行模型移植与集成，定义相关输入、输出接口，根据实车使用环境提出适配性优化方案，搭建硬件离线运行平台，如图3.6所示。主要器件包括：驾驶室内的降噪麦克风、后备箱中的Nvidia Jetson TX2和无线路由器。TX2分别与麦克风、汽车系统应用软件通过USB、局域网进行连接。整个系统运行过程如下：驾驶员在驾驶室内发出语音指令，麦克风收集语音信号传输至TX2进行语义解析，然后将结果传输至汽车应用软件实现相应操作，同时通过车载喇叭播放回复语音。  **3.3 可能的创新**  （1）针对基于深度编—解码器的自动语音识别模型参数量庞大的问题，在解码器中使用基于残差分组线性变换的缩放单元耦合各层，实现自动语音识别模型参数量和计算复杂度的降低；  （2）针对基于显式联合建模的自然语言理解模型泛化能力不足的问题，搭建基于层级交互的全局—局部交互网络，设计层级信息融合算法，高效融合所有交互信息，实现自然语言理解模型泛化能力和预测精度的提升；  （3）针对“云—端”方式运行的车载智能语音对话存在数据安全隐患的问题，搭建基于车载嵌入式设备的本地运行平台，提出面向实际应用场景的适配性优化方案，实现数据安全、自然实时的离线智能语音对话。  **参考文献：**   1. Seaborn K, Miyake N P, Pennefather P, et al. Voice in Human–Agent Interaction: A Survey[J]. ACM Computing Surveys (CSUR), 2021, 54(4): 1-43. 2. 中泰证券研究所. Cerence：人工智能助力未来出行-车智能化海外公司系列报告[EB/OL]. (2021-04-25)[2021-12-25]. https://pdf.dfcfw.com/pdf/H3\_AP202104261487921362\_1.pdf?1619458440000.pdf. 3. 陈艳华. 基于智能交互的车载语音系统的设计与实现[D]. 北京交通大学, 2020. 4. Deriu J, Rodrigo A, Otegi A, et al. Survey on evaluation methods for dialogue systems[J]. Artificial Intelligence Review, 2021, 54(1): 755-810. 5. Rigoll G, Neukirchen C. A new approach to hybrid HMM/ANN speech recognition using mutual information neural networks[C]//NIPS. 1996: 772-778. 6. Mohamed A, Dahl G, Hinton G. Deep belief networks for phone recognition[C]//Nips workshop on deep learning for speech recognition and related applications. 2009, 1(9): 39. 7. Abdel-Hamid O, Mohamed A, Jiang H, et al. Convolutional neural networks for speech recognition[J]. IEEE/ACM Transactions on audio, speech, and language processing, 2014, 22(10): 1533-1545. 8. Graves A. Connectionist temporal classification[M]//Supervised Sequence Labelling with Recurrent Neural Networks. Springer, Berlin, Heidelberg, 2012: 61-93. 9. Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112. 10. Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd international conference on Machine learning. 2006: 369-376. 11. Graves A. Sequence transduction with recurrent neural networks[J]. arXiv preprint arXiv:1211.3711, 2012. 12. Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE international conference on acoustics, speech and signal processing. Ieee, 2013: 6645-6649. 13. Graves A. Generating sequences with recurrent neural networks[J]. arXiv preprint arXiv:1308.0850, 2013. 14. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014. 15. Chorowski J, Bahdanau D, Cho K, et al. End-to-end continuous speech recognition using attention-based recurrent NN: First results[J]. arXiv preprint arXiv:1412.1602, 2014. 16. Bahar P, Zeyer A, Schlüter R, et al. On using 2d sequence-to-sequence models for speech recognition[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 5671-5675. 17. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008. 18. Zhou S, Dong L, Xu S, et al. A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin chinese[C]//International Conference on Neural Information Processing. Springer, Cham, 2018: 210-220. 19. Merboldt A, Zeyer A, Schlüter R, et al. An Analysis of Local Monotonic Attention Variants[C]//INTERSPEECH. 2019: 1398-1402. 20. Zeyer A, Irie K, Schlüter R, et al. Improved training of end-to-end attention models for speech recognition[J]. arXiv preprint arXiv:1805.03294, 2018. 21. Gulati A, Qin J, Chiu C C, et al. Conformer: Convolution-augmented Transformer for Speech Recognition[J]. arXiv preprint arXiv:2005.08100, 2020. 22. Xu P, Sarikaya R. Convolutional neural network based triangular crf for joint intent detection and slot filling[C]//2013 ieee workshop on automatic speech recognition and understanding. IEEE, 2013: 78-83. 23. Ravuri S, Stolcke A. Recurrent neural network and LSTM models for lexical utterance classification[C]//Sixteenth Annual Conference of the International Speech Communication Association. 2015: 35-40 24. Yao K, Zweig G, Hwang M Y, et al. Recurrent neural networks for language understanding[C]//Interspeech. 2013: 2524-2528. 25. Mesnil G, He X, Deng L, et al. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding[C]//Interspeech. 2013: 3771-3775. 26. Mesnil G, Dauphin Y, Yao K, et al. Using recurrent neural networks for slot filling in spoken language understanding[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 23(3): 530-539. 27. Zhang X, Wang H. A joint model of intent determination and slot filling for spoken language understanding[C]//IJCAI. 2016, 16: 2993-2999. 28. Hakkani-Tür D, Tür G, Celikyilmaz A, et al. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm[C]//Interspeech. 2016: 715-719. 29. Goo C W, Gao G, Hsu Y K, et al. Slot-gated modeling for joint slot filling and intent prediction[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 753-757. 30. Li C, Li L, Qi J. A self-attentive model with gate mechanism for spoken language understanding[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 3824-3833. 31. Qin L, Che W, Li Y, et al. A Stack-Propagation Framework with Token-Level Intent Detection for Spoken Language Understanding[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 2078-2087. 32. Haihong E, Niu P, Chen Z, et al. A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 5467-5471. 33. Zhang C, Li Y, Du N, et al. Joint Slot Filling and Intent Detection via Capsule Neural Networks[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 5259-5267. 34. Liu Y, Meng F, Zhang J, et al. CM-Net: A Novel Collaborative Memory Network for Spoken Language Understanding[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 1051-1060. 35. Zhang L, Ma D, Zhang X, et al. Graph lstm with context-gated mechanism for spoken language understanding[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(05): 9539-9546. 36. Qin L, Liu T, Che W, et al. A co-interactive transformer for joint slot filling and intent detection[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 8193-8197. 37. Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4171-4186. 38. Chen Q, Zhuo Z, Wang W. Bert for joint intent classification and slot filling[J]. arXiv preprint arXiv:1902.10909, 2019. |

|  |
| --- |
| **四、进度安排和预期成果** |
| **4.1 预期成果**  发表SCI论文2篇。  **4.2 进度安排**   |  |  |  | | --- | --- | --- | | 时间段 | 工作内容 | 预期成果 | | 2022.01~2022.02 | 阅读相关文献，整理相关理论依据并与导师保持交流。 | 完成文献整理工作，撰写综述。 | | 2022.03~2022.04 | 收集、处理所需数据集。 | 完成数据集整理工作。 | | 2022.05~2022.08 | 设计ASR模型并在数据集上训练，进行模型评估与优化。 | 完成ASR模型训练与优化。 | | 2022.09~2022.12 | 设计NLU模型并在数据集上训练，进行模型评估与优化。 | 完成NLU模型训练与优化。 | | 2023.01~2023.03 | 搭建车载智能语音对话平台，撰写毕业论文。 | 完成平台搭建和论文撰写。 | | 2023.04~2023.05 | 导师和教研室审定，论文定稿，准备答辩。 |  |   **4.3 研究条件**  （1）项目支撑  ①李鹏华，教育部-中国移动科研基金（2018）项目：语音智能对话系统（MCM20180404）,2019.01-2020.12.  ②李鹏华，重庆市基础研究与前沿探索（重庆市自然科学基金）项目：基于深度学习的车载多轮对话系统（cstc2018jcyjAX0167），2018.01-2021.06  （2）硬件资源  实验环境为一台搭载Ubuntu操作系统的服务器，编程语言、框架与服务器配置信息如表4.1所示。  表4.1 实验环境及配置   |  |  | | --- | --- | | 实验环境 | 环境配置 | | 操作系统 | Ubuntu18.04 | | CPU | Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz × 2 | | 内存 | 256GB | | 编程语言 | Python3.7 | | GPU | NVIDIA Tesla P100 16GB × 4 |   （3）数据来源  AI-SHELL1：中文普通话语音数据集，共178小时。其中训练集包含340人、验证集包含40人、测试集包含20人。  MultiATIS++：中文自然语言理解文本数据集，共5871条语句，意图标签21个，槽位标签120个。其中训练集4478条、验证集500条、测试集893条。  （4）文献资料  重庆邮电大学图书馆，SCI、中国知网、IEEE等学术资源数据库。 |

备注：以上内容根据情况添页。

图示

描述已自动生成