基于堆叠式注意力网络的复杂话语领域分类方法

梁超杰, 黄沛杰 *, 丁健德, 朱建恺, 林丕源
华南农业大学, 数学与信息学院, 广东广州, 510642
ross@stu.scau.edu.cn, pjhuang@scau.edu.cn, bighead@stu.scau.edu.cn,
gabriel@stu.scau.edu.cn, pyuanlin@scau.edu.cn

摘要

话语领域分类（utterance domain classification, UDC）是口语语言理解（spoken language understanding, SLU）中语义分析的关键步骤。尽管带注意力机制的递归神经网络已经得到了广泛的应用，并将 UDC 的研究进展提高到了一个新的水平，但是对于复杂的话语，如长度较长的话语或带有逗号的复合句的话语，有效的 UDC 仍然是一个挑战。本文提出一种基于堆叠式注意力网络的话语领域分类方法 SAN-DC（stacked attention networks-DC）。该模型综合了对口语话语多层次的语言特征的捕捉，增强对复杂话语的理解。首先在模型底层采用语境化词向量（contextualized word embedding）得到良好的词汇特征表达，并在词法层采用长短期记忆网络 (long short-term memory) 将话语编码为上下文向量表示。接着在语法级别上使用自注意力机制（self-attention mechanism）来捕捉特定领域的词依赖，然后使用词注意力（word-attention）层提取语义信息。最后使用残差连接（residual connection）将低层语言信息传递到高层，更好地实现多层语言信息的融合。本文在中文话语领域分类基准语料 SMP-ECDT 上验证所提出的方法的有效性。通过与研究进展的文本分类模型对比，本文的方法取得了较高的话语领域分类正确率。尤其是对于较为复杂的用户话语，本文提出的方法较研究进展方法的性能提升更为显著。

关键词： 话语领域分类 ； 堆叠式注意力 ； 多层次语言信息 ； 复杂话语

Complex Utterance Domain Classification Using Stacked Attention Networks

Chaojie Liang, Peijie Huang*, Jiande Ding, Jiankai Zhu, Piyuan Lin
College of Mathematics and Informatics, South China Agricultural University, China
ross@stu.scau.edu.cn, pjhuang@scau.edu.cn, bighead@stu.scau.edu.cn,
gabriel@stu.scau.edu.cn, pyuanlin@scau.edu.cn

Abstract

Utterance domain classification (UDC) is a critical step of semantic analysis in spoken language understanding (SLU). Although attentive recurrent neural networks (RNN) have been widely used and have taken the state of the art of UDC to a new level, there remains challenge in effective DC for complex utterance, such as long utterances or utterances with commas. In this paper, we present SAN-DC, a DC method based on stacked attention networks, which integrates the capture of multi-level linguistic

* 通讯作者

features of spoken language to enhance the understanding of complex utterances. Particularly, we use contextualized word embedding and encode the utterance into vector representation and learn the lexical features by applying LSTM. Then we capture domain-specific word dependencies with self-attention mechanism at the syntax level, followed by a word-attention layer to extract semantic information. We finally employ the residual connection to join linguistic information from different levels. We conduct extensive experiments on the SMP-ECDT benchmark corpus. The results show that our model achieves a higher accuracy than the state-of-the-art baselines, especially on complex utterances.

Keywords: Utterance domain classification , Stacked attention , Multi-level linguistic information , Complex utterance

1 引言

近年来，口语智能助手在我们的日常生活中变得越来越重要，很多日常便携设备，如手机、智能手表、电脑都引入了智能助手应用程序 (俞凯等, 2015)。智能助手的关键组成部分是口语语言理解（spoken language understanding，SLU）模块，即机器通过理解用户特定的指令，有针对性地去执行任务 (Tür and Mori, 2011)。这种对“特定指令”进行理解的第一步是，将接收到的用户的话语分类到特定的领域中，以便进行进一步处理。此过程称为话语领域分类（utterance domain classification，UDC） (Tür et al., 2012; Xu and Sarikaya, 2014)。例如，用户话语“今天福州的天气怎么样?”应该首先分类到“天气”领域。

传统的多领域分类器使用简单的线性模型构建，它们使用一对多（one-versus-all）的方式进行分类 (Kim et al., 2018)，例如多项逻辑回归（MLR）或支持向量机（SVM）。这些模型通常使用词的 n-gram 特征以及基于静态词典匹配的特征。随着深度学习的发展，一些网络架构已经应用于话语分类，例如深度神经网络（deep neural networks，DNNs） (Tür et al., 2012)、卷积神经网络（convolutional neural networks，CNNs）(Xu and Sarikaya, 2013; Kim, 2014)、递归神经网络（recurrent neural networks，RNNs） (Xu and Sarikaya, 2014)、以及基于注意力机制（attention mechanism）的长期短期记忆网络（long short-term memory，LSTM） (Liu and Lane, 2016; Kim et al., 2018)。尽管已经证明了注意力机制能够选择性地关注句子特定部分来提高分类准确性，但这种方法不能完全捕捉语言的结构组成 (Cheng et al., 2016)。对于复杂的话语，如长度较长的话语或带有逗号的复合句的话语，有效的 UDC 仍然是一个挑战。较长的话语隐含着较多的噪声，而带有逗号的复合句的话语通常伴随着更复杂的表达方式。如图 1 所示是中文话语领域分类的基准语料 SMP-ECDT (Zhang et al., 2017; Zhao et al., 2019) 中的一个复杂话语例子在两种研究进展分类模型中的词注意力可视化。这两种模型都采用了 BERT 嵌入 (Devlin et al., 2019) 以及双向 LSTM (BiLSTM) (Cheng et al., 2016) 编码器，分别使用了软注意力（soft attention） (Liu and Lane, 2016; Yang et al., 2016) 和硬注意力（hard attention） (Shankar et al., 2018) 机制。从图 1 可以看到，两种不同的词注意力机制均未能对话语中心词“飞机票”给予足够的关注。

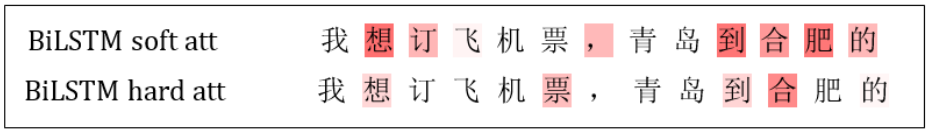


Figure 1: 一个复杂话语例子的词注意力可视化

本文面向复杂话语领域分类提出了一种新的堆叠式注意力网络模型 SAN-DC（stacked attention networks for domain classification），该模型综合了对口语话语多层次的语言特征的捕捉。我们的模型使用 BiLSTM 层将话语编码为上下文向量表示，并使用自注意机制 (Daniluk et al., 2017; Liu and Lapata, 2018) 捕获特定领域的词语依赖关系，然后使用词意机制提取语义信息。最后我们使用残差连接来连接不同语言层次的信息，解决了多层网络中特征传播和梯度消失（或爆炸）的固有困难。

上述架构和机制构成了一个理解复杂话语的有竞争力的模型，并在中文话语领域分类的基准语料 SMP-ECDT 上取得了优于研究进展方法的性能。本文的主要贡献如下：

- 本文提出了一种基于堆叠式注意力网络的话语领域分类模型 SAN-DC，实现词汇、词法、句法和语义多层语言信息的有效融合，增强复杂话语的语义理解，提高领域分类性能。
- 通过中文话语领域分类基准语料 SMP-ECDT 上的实验，验证了本文的模型取得了优于研究进展方法的效果。尤其对于复杂话语，本文的模型性能提升更为显著。

2 相关工作

本文的研究综合了对口语话语多层次的语言特征的捕捉，提高对复杂话语的理解性能。本节简要介绍相关技术方法，并阐述本文方案中融入这些技术方法的设计依据。

已经有很多话语或短文本分类的研究致力于提高分类器性能，早期的典型代表是 SVM 和最大熵等 (Haffner et al., 2003; Phan et al., 2008)。随后，深度学习在自然语言处理（Natural Language Processing，NLP）中受到关注，主流的应用包括深度信念网络（deep belief networks，DBN） (Sarikaya et al., 2011)、CNN (Xu and Sarikaya, 2013; Kim, 2014) 和 RNN (Xu and Sarikaya, 2014)，尤其是 RNN 中最常用的 LSTM (Cheng et al., 2016; Ravuri and Stolcke, 2016; Vu et al., 2016; 柯子烜等, 2018)。

近年来，注意力机制被引入到了 NLP 中，实验证明其善于在文本分类任务中抽取文本的含义，例如意图检测 (Liu and Lane, 2016)、领域分类 (Kim et al., 2018)、情感分类 (曾锋等, 2019) 和文档分类 (Yang et al., 2016) 等。尽管注意力机制在为关键词分配权重方面非常有效，但 Cheng 等人 (2016) 指出它无法完全捕捉语言的结构组成。

自注意力机制在 RNN 中的应用证明了它具有捕获词之间句法依赖的能力 (Liu and Lapata, 2018; Li et al., 2018; Lin et al., 2017)。然而，简单地在序列上应用自注意力机制对于学习依赖性是不够的。为了提高捕捉依赖的性能，Vaswani 等人 (2017) 把位置嵌入应用于机器翻译中，在编码和解码过程中引导自注意力机制提取重要特征。在本文的工作中，我们使用自注意力机制学习词之间的依赖关系，用于增强模型中话语内的词信息交互，有助于词注意力层更好的学习语义层次的特征。

我们的模型还依赖于残差连接。残差连接通过解决消失梯度问题来改善深度神经网络的学习过程 (He et al., 2016)。这种连接通过搭建跨越多层网络的直接路径，有助于不同层间信息的传输。最近，Zhang 等人 (2016)、Kim 等人 (2017)、Zilly 等 (2017)、以及 Wang 和 Tian (2016) 已经提出了几种用于序列预测的带有残差连接的 LSTM 架构。本文将残差连接应用于连接来自模型不同语言层次（词法特征层和语义特征层）的信息，使得 BiLSTM 中的参数通过注意力机制并行地更新，同时也通过 RNN 架构串行地更新。

最后是关于在模型中进行语言信息建模 (Hashimoto et al., 2017; Strubell et al., 2018) 的研究。Hashimoto 等人 (2017) 在模型的不同深度学习不同的任务，从模型的浅层到深层，利用 LSTM 和残差连接渐进地学习词法、句法和语义特征。他们的实验说明了模型中不同深度对各个任务的性能影响有所不同，更深层次可以处理更复杂的任务。同时本文也受到了 Yang 等人 (Yang et al., 2016) 的启发，他们通过词级别注意力和句级别注意力分层地计算文档语义。但本文没有使用多任务学习，也不是为了计算文档语义而使用层次性的注意力，而是利用自注意力先计算短语级别的注意力，然后计算句级别的注意力，自下而上地学习更高层次的语言特征，最后通过残差连接将低层语言信息传递到高层，更好地实现多层语言信息的融合，应用于领域分类任务，改善对复杂话语的有效理解。

3 模型

本文提出的 SAN-DC 模型的总体框架如图2所示。我们的模型旨在通过一个堆叠式注意力网络模型，综合多层次的语言特征（词汇、词法、句法和语义特征）的捕捉，来增强对复杂话语的有效理解。

首先，在底层采用语境化词向量（具体上，本文采用的是 BERT (Devlin et al., 2019)）得到良好的话语词汇特征表达，并通过一个双向 LSTM (Cheng et al., 2016) 层将话语编码为上下文向量表示。之后，我们在带有词法特征的 BiLSTM 层上使用自注意力机制 (Vaswani et al., 2017) 来提取词之间的依赖。然后在自注意力层上使用软注意力来提取序列的语义信息。最后，

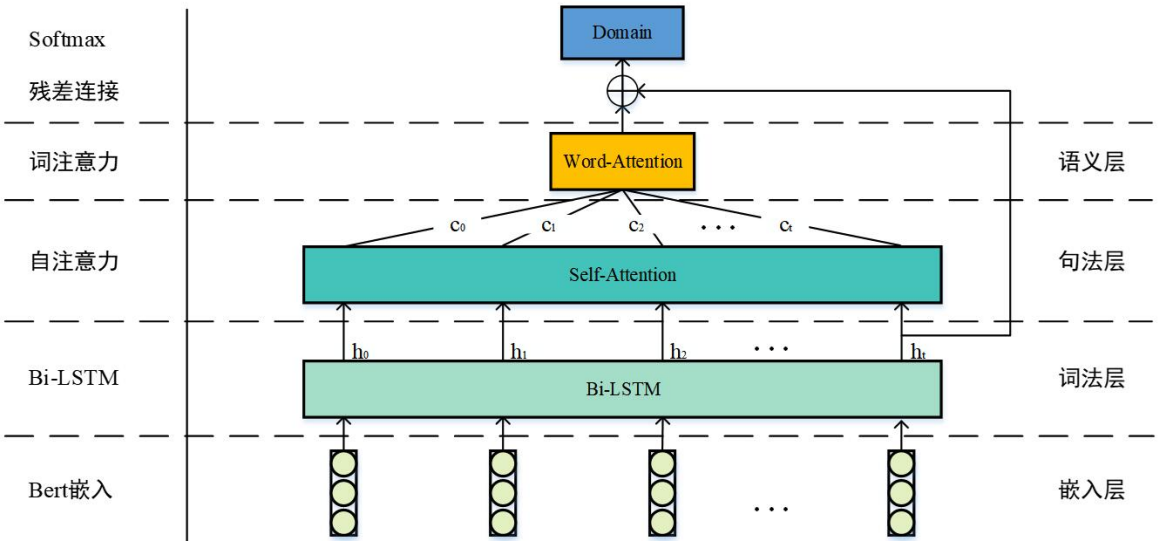


Figure 2: SAN-DC 模型框架

我们采用残差连接来合并 BiLSTM 最后的状态输出和软注意力层输出，将高层特征和低层特征融合在一起，我们将融合后的向量输入到带有 softmax 激活函数的全连接层以预测话语领域标签。

模型的总体架构受到 Hashimoto 等人 (2017) 的研究的启发，他们的研究证明了不同深度多个学习任务的有效性，并使用跳跃连接（skip-connection）来增强更深层次任务的性能，而我们的模型通过使用注意力机制和残差连接来提高话语领域分类性能。

3.1 BiLSTM

RNN 已经被广泛应用于 NLP 中。然而，想要解决需要学习长序列依赖性的问题，采用标准的 RNN 可能是困难的，因为损失函数的梯度随时序长度呈指数变化，并且通常会引发梯度消失/爆炸问题。LSTM (Hochreiter and Schmidhuber, 1997) 通过引入门控机制来缓解这个问题。本文中使用 LSTM 如下：

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{2}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{3}$$

$$\tilde{s_t} = tanh(W_s x_t + U_s h_{t-1} + b_s) \tag{4}$$

$$s_t = i_t \odot \tilde{s_t} + f_t \odot s_{t-1} \tag{5}$$

$$h_t = o_t \odot tanh(s_t) \tag{6}$$

其中 i，f 和 o 分别表示输入、遗忘和输出门。σ 代表 sigmoid 函数，$\odot$ 定义为两个向量的积。W，U 和 b 是可训练的参数。

考虑到充分利用某个时刻的历史信息和未来信息，我们在模型中使用双向 LSTM（BiLSTM）。令 $(x_1, x_2, ..., x_T)$ 表示输入序列，其中 x_t 是时刻 t 的向量。

$$\overrightarrow{h}_t = \varphi(x_t + \overrightarrow{h}_{t-1}) \tag{7}$$

$$\overleftarrow{h}_t = \varphi(x_t + \overleftarrow{h}_{t-1}) \tag{8}$$

$$h_t = [\overrightarrow{h}_t, \overleftarrow{h}_t] \tag{9}$$

其中 φ 定义为 LSTM 函数，$\overrightarrow{h}_t$ 是前向 LSTM 在时刻 t 的隐藏状态，而 $\overleftarrow{h}_t$ 是后向 LSTM 在时刻 t 的的隐藏状态，$\overrightarrow{h}_t$ 和 $\overleftarrow{h}_t$ 的拼接作为 BiLSTM 在时刻 t 的输出。

3.2 自注意力

我们使用自注意力机制来捕捉特定领域的词依赖。假设 $H \in \mathbb{R}^{(}d_h \times T)$ 是 3.1 小节中 BiLSTM 的输出序列，其包含 T 个状态向量 $[h_1, h_2, ..., h_T]$，其中 d_h 是输出状态的维度。根据 Vaswani 等人 (2017) 提供的方法，首先分别使用参数矩阵 W_Q、W_K 和 W_V 对 H 作线性变换，将其投影成 Q、K 和 V。当使单个注意力头（attention head）时，三个向量 Q、K 和 V 的形状与 H 的形状一样，带有放缩的点积注意力（scaled dot-product attention）机制计算注意力分数：

$$Q = W_Q H \tag{10}$$

$$K = W_K H \tag{11}$$

$$V = W_V H \tag{12}$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{13}$$

其中 d_k 是由等式 (11) 中的向量 K 的维数，$\sqrt{d_k}$ 是放缩比例因子。

对于随后的残差连接（3.4 小节），我们不对 H 作线性变换来得到 V。与前面带有放缩点积注意力机制中使用的方法相比，我们将 V 替换为 H。我们将等式 (13) 改写为以下公式：

$$Attention(Q, K, H) = softmax(\frac{QK^T}{\sqrt{d_H}})H \tag{14}$$

其中 Q 和 K 通过使用等式 (10) 和等式 (11) 来计算，而 H 是 BiLSTM 的输出序列。

3.3 词注意力

在大多数情况下，通过平均池化（mean pooling）或最大池化（max pooling）概括序列是切实可行的。然而，并非所有词在句子中都扮演着相同的角色，因此词在不同的上下文中应该具有不同的权重。Yang 等人 (2016) 利用词注意力（word attention）来计算句子中各个词的权重。词注意力层，我们考虑两种流行的选择，软注意力（soft attention） (Liu and Lane, 2016; Yang et al., 2016) 和硬注意力（hard attention） (Shankar et al., 2018) 机制。

假设给定自注意力层输出的序列 $[c_1, c_2, ..., c_T]$，c_i 代表某一时刻的向量表示，其中 $i \in [1, T]$。软注意力机制首先通过非线性变换把 c_i 转换为 s_i，以此作为 c_i 的隐藏状态表示，然后计算每个时刻的归一化注意力权重 a_i，用来表示每个时刻的隐藏状态 s_i 与上下文 s_ω 之间的相似性。最后，利用各个词的权重 a_i 来计算句子 c_i 中词的加权和来表示话语语义。特别地，

$$s_i = tanh(W_\omega c_i + b_\omega) \tag{15}$$

$$a_i = \frac{exp(s_i^\top s_\omega)}{\sum_i (s_i^\top s_\omega)} \tag{16}$$

$$u = \sum_i (a_i c_i) \tag{17}$$

其中 u 是用于概括话语中词的信息的话语向量。W、b 和 s 为可训练参数。

硬注意力机制则只关注句子中的一个词，可以描述为：

$$j = argmax_{i=1}^{T}(a_i) \tag{18}$$

$$u^{hard} = c_i \tag{19}$$

其中 j 为句子中注意力权重最大的词的索引，a_i 为等式（16）的注意力权重，u^{hard} 注意力权重最大的词向量。

3.4 残差连接

高级特征通常需要用更复杂的函数来映射和更深的神经网络来计算，有时这可能导致性能不稳定，尤其是在处理之前没出现过的话语时。因此，我们需要利用低级特征来使神经网络更加健壮。为了从模型的不同方面获得信息，并受到 Hashimoto 等人 (2017) 的启发，我们在模型中使用残差连接来加强更深层神经网络中的领域分类任务的性能。

使用加法操作来连接在模型中不同深度处的两个向量的残差连接 (He et al., 2016) 被广泛应用。我们使用残差连接来合并在具有相同形状（shape）的 BiLSTM 最终输出状态 h_t 的和软注意力的输出 u。特别地：

$$v_c = h_T + u \tag{20}$$

其中 v_c 是残差连接的输出。注意，v_c 的形状与 h_t 和 u 的形状（shape）相同。

由于词注意力层和自注意力层的注意力权重计算时使用了 softmax 激活函数，使得 BiLSTM 层获得的梯度比没使用上述两个注意力层时小得多，造成增加训练难度和参数更新难的问题，有必要使用残差连接来缓解这种问题。

3.5 话语领域分类

等式（20）中的向量 v_c 是与 BiLSTM 的最终输出状态和软注意的输出进行残差连接后得到的结果，我们将 v_c 输入到带有 softmax 激活函数的完全连接的层，以预测领域标签。给定数据集 D, 我们定义如下：

$$h_{do} = softmax(W_{do}v_c + b_{do}) \tag{21}$$

同样，我们使用领域标签的 NLL 作为其训练损失：

$$L(\theta) = -\sum_{(y^i,x^i)\in D} logP_\theta(y^i|x^i) \tag{22}$$

其中 y^i 是训练集中第 i 个样本的领域标签，x^i 是对应的话语序列向量，θ 为模型参数。

4 实验

4.1 数据集

为了评估我们提出的模型的性能，我们在中文口语领域分类基准语料 SMP-ECDT (Zhang et al., 2017; Zhao et al., 2019) 上进行了实验。该语料由哈尔滨工业大学社会计算与信息检索研究中心（哈工大 SCIR）和科大讯飞股份有限公司（iFLYTEK）提供。具体上采用的是 2018 年第二届中文人机对话技术评测任务一“中文话语领域分类”中的数据集。该数据集包含 3736 个训练样本和 4528 个测试样本，涵盖 31 个类别，包括闲聊类（chat）和垂类（30 个垂直领域）。在 SMP-ECDT 数据集中存在着一定比例的较长的口语话语（测试集中超过 10 个汉字的样本有 2143 个，接近测试集的 50%），以及带逗号的复合句的话语（测试集中带逗号的样本有 1290 个，接近测试集的 30%）。在这些相对复杂的话语中，尤其是那些领域特征信息不够明显的，其有效理解和领域分类存在较大困难。

4.2 实验设置

在预处理方面，我们使用 BERT (Devlin et al., 2019) 预训练模型获取字向量。在 BERT 嵌入模型，我们预训练 BERT-base 的倒数第二层的序列向量作为嵌入，训练期间冻结的嵌入层的权重，另外把集外（out-of-vocabulary，OOV）词用 <UNK> 标记代替嵌入。在 BiLSTM 层中，前向和后向的 LSTM 单元（units）数都为 64，它们拼接后维度总共是 128 维。此外，BiLSTM 层和残差连接层的输出的 Dropout (Srivastava et al., 2014) 概率都设置为 0.1。模型使用 Adam 优化器 (Kingma and Adam, 2015) 进行优化，其中 β_1=0.9，β_2=0.999，ϵ=10^{-8}，初始学习率为 0.001。模型训练轮数为 100。我们对训练数据进行 10 折交叉验证，然后选出交叉验证最好的超参数所对应的模型对测试集进行预测。所有结果是 5 次独立实验的平均值。

4.3 对比方法

我们将本文提出的 SAN-DC 与 BERT fine tune 以及 4 个基于 BERT 嵌入的研究进展文本分类方法进行对比：

- **BERT fine tune：**该方法直接将 BERT fine-tuning (Devlin et al., 2019) 应用到分类任务，在分类层增加了一个新的输出层。
- **BiLSTM：**该方法是文本分类的经典基线 (Vu et al., 2016)，非常适合于序列问题。
- **Multi-head att：**该方法采用多头注意力模型，它是 seq2seq 模型“Transformer”(Vaswani et al., 2017) 的主要组成部件，我们用它来进行领域分类任务。
- **BiLSTM hard att：**该方法使用具有硬注意力机制的 BiLSTM (Shankar et al., 2018) 进行领域分类任务。
- **BiLSTM soft att：**该方法使用具有软注意力机制的 BiLSTM (Yang et al., 2016) 进行领域分类任务。

值得注意的是，在采用中文 BERT 嵌入 (Devlin et al., 2019) 方案中，“词”都是指“字”。文本主要观察了两个变种，即分别在 BiLSTM hard att 和 BiLSTM soft att 基础上使用自注意机制捕获特定领域的词语依赖关系，使用残差连接来连接不同语言层次的信息，这两种变种我们分别命名为：SAN-DC-hard att 和 SAN-DC-soft att。

4.4 实验结果

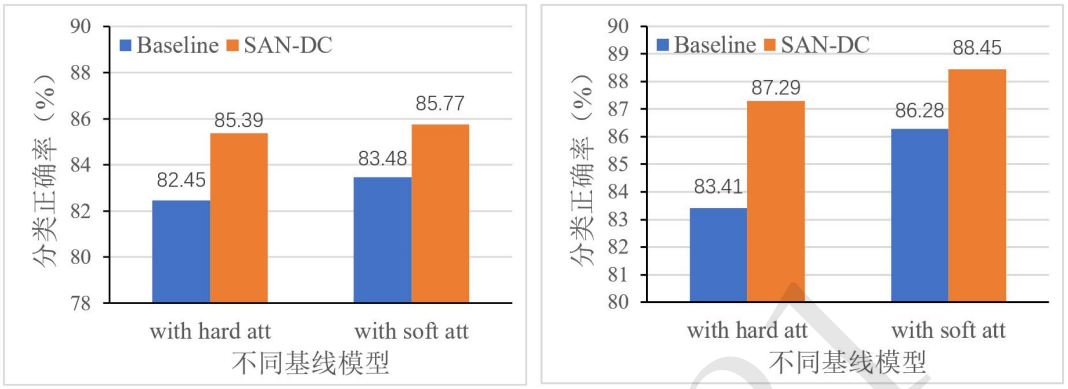
整个测试集上的性能对比。表1展示了 SAN-DC 与研究进展方法在整个测试集中进行预测的正确率对比。

方法		正确率（%）
BERT fine tune		80.60±0.28
BERT 嵌入	BiLSTM	82.07±0.41
	Multi-head att	82.06±0.65
	BiLSTM hard att	82.73±0.56
	BiLSTM soft att	82.99±0.24
	SAN-DC-hard att	84.49±0.16
	SAN-DC-soft att	**84.61**±0.26

Table 1: SAN-DC 与研究进展方法在测试集上的性能对比

从表1可以看到，在研究进展方法中，在 BERT 嵌入基础上加上编码器普遍都优于单纯 BERT fine tune 模型；词注意力机制有助于提高领域分类正确率，其中 BERT 嵌入的 BiLSTM soft att 在对比的研究进展方法中性能最好。本文的模型取得了优于所有研究进展方法的效果，其中 SAN-DC-soft att 领域分类正确率最高，达到 84.61%。对比基线模型中性能较优的硬注意力和软注意力模型，本文的模型的分类绝对正确率分别提高 1.76% 和 1.62%。

复杂话语上的性能对比。我们重点观察了堆叠式注意力在较复杂的话语上的性能。注意到长度较长或带有逗号的复合句话语是较复杂话语的典型代表，故在此类话语中进行量化分析，以体现堆叠式注意力在较复杂话语的场景下的性能。较长的句子隐含着较多的噪声，而带有逗号的复合句的话语通常伴随着更复杂的表达方式，需要更高层次的抽象的语义表示。词注意力根据每个词的表示独立地计算出注意力权重，并进行加权求和得到的向量作为话语的语义，忽略了词之间的依赖关系，不足以表示话语的语义。故在使用词注意力之前，使用自注意力捕捉词之间的依赖关系，并与该词相关的短语作为该词的向量表示（等式 (14)），词注意力层能够基于各个短语计算出话语的语义的向量表示。图3所示是 SAN-DC 与研究进展基线（图中的 Baseline 代表采用 BERT 嵌入的 BiLSTM hard/soft att）在两类复杂话语上的性能对比。



(a) 较长的话语（字数 >10）　　(b) 带逗号的话语

Figure 3: 本文的方法与研究进展基线在复杂话语的性能对比

从图3可以看到，相比于总体性能的提升（如表1所示，对于 hard att 和 soft att 基线，分别提升 1.76% 和 1.62% 的正确率），不论在较长话语的情况，还是在带逗号情况，本文的方法在词注意力基础上增加自注意力机制，都能获得更为显著的性能提升。在 hard att 基线，较长话语和带逗号话语部分的分类正确率分别提升了 2.94% 和 3.88%。在 soft att 基线，较长话语和带逗号话语部分的分类正确率分别则提升了 2.29% 和 2.17%。

值得注意的是，从图3中，我们看到，各模型在复杂话语的领域分类正确率比在表1所示总体正确率还高，但这并不意味着复杂话语容易分类。实际上，复杂话语中，有一部分是领域特征信息比较充分，并且口语化不强烈的，这部分话语的领域分类正确率较高。而对于领域特征信息不够充分，或者存在较大的跨领域混淆，这类话语则会因为句子表述上的结构或者信息复杂性而导致分类正确率较低。我们的模型主要改善的是后面这类话语的情况。

4.5 进一步的分析

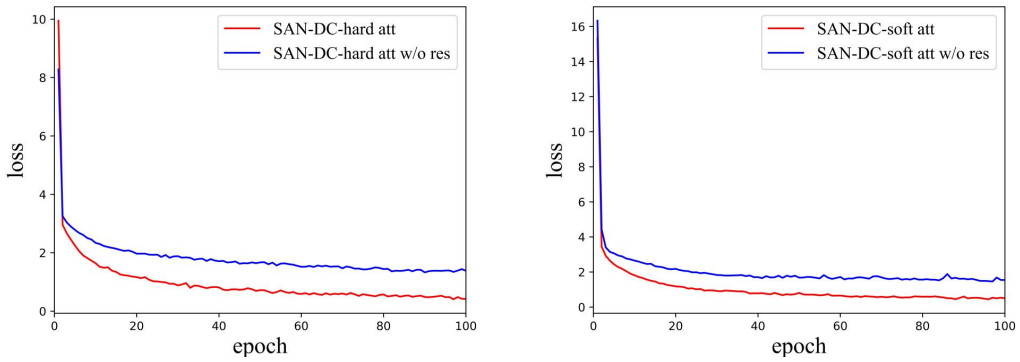
残差连接的影响。为了验证了残差连接可以优化深层神经网络的训练性能，我们记录了训练过程中 loss 的变化，结果如图4所示。从图4中我们可以看到，无论是在 SAN-DC-hard att 还是 SAN-DC-soft att 中，残差连接都起到了增强性能的效果，在相同的训练轮数下，模型拟合的效果变得更好了，具体体现在它能够使得模型的收敛速度更快、损失函数更低。

可视化分析。我们进一步采用可视化分析，通过一个带逗号并且长度较长的复杂话语例子，观察本文的 SAN-DC 模型在词注意力层和自注意力层的表现。

（1）词注意力层可视化

我们对词注意力层使用热力图进行可视化，以更好地展现基线模型和本文提出的 SAN-DC 模型在词注意力层的差异，例子如图5所示。

在图5中，注意力权重高的词颜色更深。通过对软/硬注意力进行可视化，结果显示了本文提出的结构可以提升词注意力层性能。在软注意力中，我们计算每一个词的权重，使用这些权重来表示模型对于词的注意程度。而在硬注意力中，根据注意力权重，使用了蒙特卡洛采样来选取一个词作为句子向量，在理论上，硬注意力会比软注意力注意得更加集中，从热力图中的可以看出来，硬注意力注意到的词比软注意力更加的集中，即表现为硬注意力热力图的颜色块比软注意力的颜色块少。在图5的例子中，前半句带有订飞机票的意图，后半句是对飞机票的补充说明，



(a) SAN-DC-hard att (b) SAN-DC-soft att

Figure 4: 带和没带残差连接的 SAN-DC 训练性能对比

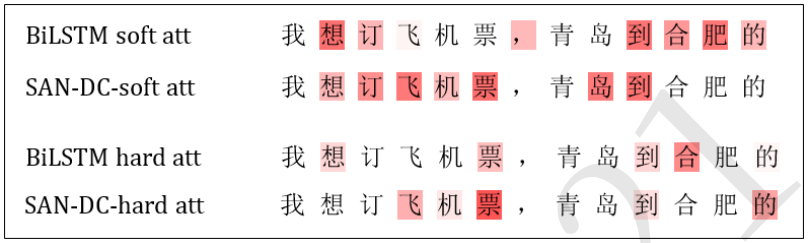


Figure 5: SAN-DC 与两种基线的词注意力层可视化例子对比

显然“飞机票”是话语的中心词，在话语中应该有更高的权重。从词注意力层本文的方案与基线方案的对比，可以看到，无论是硬注意力基线还是软注意力基线，本文的 SAN-DC 使用自注意力机制来捕捉特定领域的词依赖，都有效地提高了领域关键词在词注意力层的注意力权重。

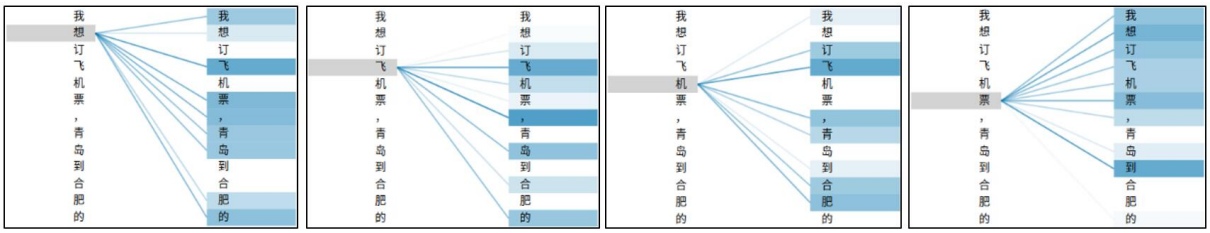
（2）自注意力层可视化

为了进一步说明自注意力层对词注意力层的影响，我们使用 BertViz (Vig, 2019) 对自注意力层进行了可视化。我们选取了图5中一些词注意力权重变化较大，并且可能影响分类正确性的词（“想”、“飞”、“机”、“票”等）进行了自注意力层的可视化，在一定程度上可以更直观地说明自注意力层的作用，结果如图6所示。

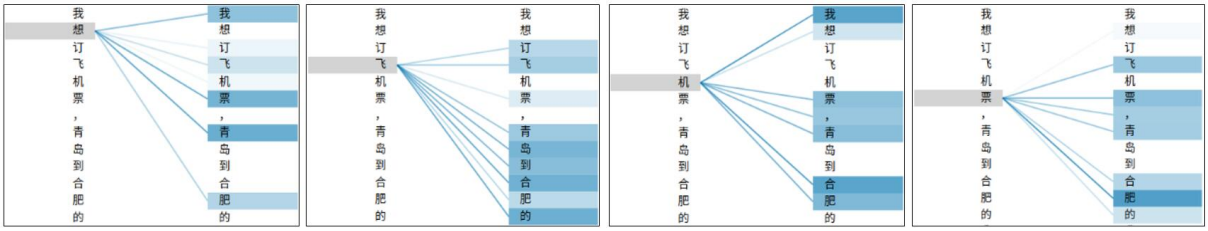
在图6中，展示了每个词有关联的词，以及相关强度。可以看到“飞机票”的每个字都有较多的关联依赖，这使得它们在词注意力层能够得到更大的注意力权重。而“想”字得到的依赖关联相比之下并不是很多，尤其是在采用硬注意力机制的方案里。并且，在词注意力层词的注意力权重是相对的，在“飞机票”的每个字权重提高之后，“想”字的权重就相对变小了。在自注意力和词注意力机制的共同作用下，领域关键词的权重的上升，领域性模糊的字的权重相对变小，有助于提高话语领域分类正确的概率。

5 结束语

本文面向人机对话中复杂话语的有效理解和领域分类，提出了一种堆叠式注意力网络模型 SAN-DC。为了更好地实现多层次语言信息的融合，本文的模型在底层采用语境化词向量获取更好的词汇特征表达，并在词法层采用 BiLSTM 将话语编码为上下文向量表示。然后我们使用自注意力机制在句法层面来捕捉话语中的词依赖，并借助软注意力机制来提取话语的语义特征，最后利用残差连接融合高层特征和低层特征，加快收敛速度并使得模型更加健壮。我们在中文话语领域分类基准语料 SMP-ECDT 进行实验，实验结果证明了在领域分类中，在软和硬注意力两种基线模型上，SAN-DC 模型都取得了优于研究进展方法的效果。尤其对于较长的话语和带有逗号的话语等为代表的较复杂的话语，SAN-DC 性能提升更为显著。此外，本文还进一步通过可视化观察了词注意力层以及自注意力机制对词注意力的影响。



(a) SAN-DC-hard att



(b) SAN-DC-soft att

Figure 6: 基于两种基线的 SAN-DC 模型的自注意力层可视化例子

致谢

本文受到广东省自然科学基金 (2021A1515011864)、国家自然科学基金 (71472068)、广东省普通高校特色创新项目 (2020KTSCX016) 的资助。

参考文献

柯子烜, 黄沛杰, 曾真. 2018. 基于优化“未定义”类话语检测的话语领域分类. 中文信息学报, 32(4):105-113.

俞凯，陈露，陈博等. 2015. 任务型人机对话系统中的认知技术——概念、进展及其未来. 计算机学报, 38(12):2333-2348.

曾锋, 曾碧卿, 韩旭丽, 等. 2019. 基于双层注意力循环神经网络的方面级情感分析. 中文信息学报, 33(6):108-115.

J. P. Cheng, L. Dong, and M. Lapata. 2016. Long short-term memory-networks for machine reading. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pp. 551–561.

M. Daniluk, T. Rockt¨aschel, J. Welbl, et al. 2017. Frustratingly Short Attention Spans in Neural Language Modeling. *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*.

J. Devlin, M. Chang, K. Lee, et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pp. 4171–4186.

P. Haffner, G. Tur, and J. H. Wright. 2003. Optimizing SVMs for complex call classification. *Proceedings of the 28th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, pp. 632-635.

K. Hashimoto, C. M. Xiong, Y. Tsuruoka, et al. 2017. A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pp. 1923-1933.

K. M. He, X. Y. Zhang, S. Q. Ren, et al. 2016. Deep residual learning for image recognition. *Proceedings the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pp. 770–778.

S. Hochreiter and J. Schmidhuber. 1997. Long Short-Term Memory. *Neural computation*, 9(8):1735–1780.

Y. Kim. 2014. Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 1746–1751.

J. Kim, M. El-Khamy, J. Lee. 2017. Residual LSTM: Design of a Deep Recurrent Architecture for Distant Speech Recognition. *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, pp. 1591-1595.

Y. Kim, D. Kim, A. Kumar. 2018. Efficient large-scale neural domain classification with personalized attention. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 2214–2224.

D. Kingma and J. Ba. Adam. 2015. A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015).*

C. L. Li, L. Li, and J. Qi. 2018. A self-attentive model with gate mechanism for spoken language understanding. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pp. 3824-3833.

Z. H. Lin, M. W. Feng, C. Nogueira dos Santos, et al. 2017. A Structured Self-Attentive Sentence Embedding. *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017).*

B. Liu and T. Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *Proceedings of the 17th Annual Conference of the Inter-national Speech Communication Association (INTERSPEECH 2016)*, pp. 685-689.

Y. Liu and M. Lapata. 2018. Learning Structured Text Representations. *Transactions of the Association for Computational Linguistics*, 6: 63–75.

X .H. Phan, L. M. Nguyen, S. Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from largescale data collections. *Proceedings of the 17th International Conference on World Wide Web (WWW 2008)*, pp. 91—100.

S. Ravuri and A. Stolcke. 2016. A comparative study of recurrent neural network models for lexical domain classification. *Proceedings of the 41th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2016)*, pp. 6075-6079.

R. Sarikaya, G. E. Hinton, and B. Ramabhadran. 2011. Deep belief nets for natural language call-routing. *Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, pp. 5680–5683.

S. Shankar, S. Garg, and S. Sarawagi. 2018. Surprisingly easy hard-attention for sequence to sequence learning. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pp. 640–645.

N. Srivastava, G. E. Hinton, A. Krizhevsky, et al. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 2014*, 15: 1929–1958.

E. Strubell, P. Verga, D. Andor, et al. 2018. Linguistically-Informed Self-Attention for Semantic Role Labeling. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pp. 5027-5038.

G. Tür and R. Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*, John Wiley and Sons, Inc.

G. Tür, L. Deng, D. Hakkani-Tür, et al. 2012. Towards deeper understanding: Deep convex networks for semantic utterance classification. *Proceedings of the 37th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'12)*, pp. 5045–5048.

A. Vaswani, N. Shazeer, N. Parmar, et al. 2017. Attention is all you need. *Proceedings of the 41th Annual Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 6000–6010.

J. Vig. 2019. Visualizing Attention in Transformer-Based Language Representation Models. *Computing Research Repository, arXiv:1904.02679. 2019*, Version 2.

N. T. Vu, P. Gupta, H. Adel, et al. 2016. Bi-directional recurrent neural network with ranking loss for spoken language understanding. *Proceedings of the 41th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, pp. 6060–6064.

Y. R. Wang and F. Tian. 2016. Recurrent residual learning for sequence classification. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pp. 938–943.

P. Y. Xu and R. Sarikaya. 2013. Convolutional neural network based triangular CRF for joint intent detection and slot filling. *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2013)*, pp. 78–83.

P. Y. Xu and R. Sarikaya. 2014. Contextual domain classification in spoken language understanding systems using recurrent neural network. *Proceedings of the 39th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, pp. 136–140.

Z. C. Yang, D. Y. Yang, C. Dyer, et al. 2016. Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*, pp. 1480-1489.

W. N. Zhang, Z. G. Chen, W. X. Che, et al. 2017. The first evaluation of Chinese human-computer dialogue technology. *Computing Research Repository, arXiv:1709.10217*, Version 1.

Y. Zhang, G. G. Chen, D. Yu, et al. 2016. Highway long short-term memory RNNs for distant speech recognition. *Proceedings of the 41th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, pp. 5755–5759.

Z. Y. Zhao, W. N. Zhang, W. X. Che, et al. 2019. An Evaluation of Chinese Human-Computer Dialogue Technology. *Data Intelligence*, 1(2):187-200.

J. G. Zilly, R. K. Srivastava, J. Koutník, et al. 2017. Recurrent highway networks. *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pp. 4189–4198.