基于多头注意力和 BiLSTM 改进 DAM 模型的中文问答匹配方法

秦汉忠	于重重*	姜伟杰	赵霞
北京工商大学	北京工商大学	北京工商大学	北京工商大学
793973698@qq.com	chongzhy@vip.sina.com	359206371@qq.com	zhaox@btbu.edu.cn

摘要

针对目前检索式多轮对话深度注意力机制模型 DAM（Deep Attention Matching Network）候选回复细节不匹配和语义混淆的问题，本文提出基于多头注意力和双向长短时记忆网络（BiLSTM）改进 DAM 模型的中文问答匹配方法，该方法采用多头注意力机制，使模型有能力建模较长的多轮对话，更好的处理目标回复与上下文的匹配关系。此外，本文在特征融合过程中采用 BiLSTM 模型，通过捕获多轮对话中的序列依赖关系，进一步提升选择目标候选回复的准确率。本文在豆瓣和电商两个开放数据集上进行实验，实验性能均优于 DAM 基线模型，R_{10}@1 指标在含有词向量增强的情况下提升了 1.5%。

关键词：检索式多轮对话；DAM；多头注意力；BiLSTM

Chinese question answering method based on multi-head attention and BiLSTM improved DAM model

Hanzhong Qin	Chongchong Yu*	Weijie Jiang	Xia Zhao
Beijing Technology and Business University	Beijing Technology and Business University	Beijing Technology and Business University	Beijing Technology and Business University
793973698@qq.com	chongzhy@vip.sina.com	359206371@qq.com	zhaox@btbu.edu.cn

Abstract

Aiming at the current problem of Deep Attention Matching Network(DAM) can not effectively match response details, and will cause semantic confusion, a Chinese question answering method based on multi-head attention and Bi-directional Long Short-Term Memory (BiLSTM) improved DAM model was proposed. This method can model longer multiple rounds of dialogue and handle the matching relationship between the response selection and the context. In addition, this paper uses the BiLSTM Network in the feature fusion process to improve the accuracy of multi-turn response selection tasks by capturing the time-dependent relation. In this paper, we test the improved DAM on two public multi-turn response selection datasets, the Douban Conversion Corpus and the E-commerce Dialogue Corpus. Experimental results show our model outperforms the baseline model by 1.5% in Recall-10-at-1 with the word vector enhancement.

Keywords: multi-turn response selection ,Deep Attention Matching ,multi-head attention ,Bi-directional Long Short-Term Memory

1 引言

人机对话系统是一个复杂的研究方向，构建人机对话系统的方法之一是检索式方法。检索式方法首先需提取输入对话的特征，随后在候选回复库匹配多个目标候选回复，按照某种指标进行排序，输出得分最高的回复。将之前的对话输出作为历史对话，即形成多轮对话的形式。

近年来，随着深度学习的发展，关于人机对话的研究重点逐渐由基于模板、规则的传统方法转变为基于端到端的深度学习模型方法。（Wu Y et al.,2016）提出序列匹配网络（Sequence Matching Network，SMN）模型，模型可分为“表示—匹配—融合”三个部分，整体上基于 CNN 和 RNN 实现以语义融合为中心的多轮对话回复选择。（Zhang Z et al.,2018）提出深度表达融合（Deep Utterance Aggregation，DUA）模型，针对 SMN 模型将历史对话直接拼接为上下文存在噪声和冗余的问题，采用注意力机制挖掘关键信息并忽略冗余信息，最终获得对话表达和候选响应的匹配得分。（Zhou X et al.,2018）提出深度注意力匹配（Deep Attention Matching，DAM）模型。在 SMN 模型的基础上，省去 CNN 和 RNN 等结构，仅依靠注意力机制完成多轮对话的回复选择，使模型参数大幅度减少，大幅提升了训练速度。而这类方法的局限性在于候选集中被选定的回复仅适用于本轮对话，与上下文并不能形成良好的匹配，或在匹配模型中没有学习到真正的语义关系，对多轮对话的内容产生了混淆，难以选择正确的候选回复。

在“表示—匹配—融合”这一框架下，同时优化三个部分是现阶段的研究难点。在前人研究的基础上，本文对比 DAM 模型，通过引入多头注意力机制，使模型更适合处理含有细微变化的数据，能让选定的目标候选回复与上下文形成良好的匹配关系。此外，本文在特征融合过程中采用 BiLSTM 模型，通过捕获多轮对话中的序列依赖关系，帮助模型建立每轮对话与前一轮对话、候选回复之间的匹配信息，使匹配模型学习到真正的语义关系，进一步提高选择目标候选回复的准确率，基于此建立基于多头注意力和 BiLSTM 改进的 DAM 模型 Ex-DAM，在豆瓣、电商这两个多轮中文对话数据集上进行研究。

论文的结构安排如下：第 1 节介绍了“检索式人机多轮对话”的概念及特点，概述了近几年深度学习模型方法；第 2 节介绍了深度注意力机制模型 DAM 的整体结构；第 3 节介绍基于多头注意力和 BiLSTM 改进的 DAM 模型 Ex-DAM，主要包括多头注意力模块、语义表示网络和双通道 BiLSTM 特征融合网络；第 4 节介绍实验数据、实验内容、实验结果及分析，验证 Ex-DAM 模型的有效性；最后在第 5 节进行总结。

2 DAM 模型

（Zhou X et al.,2018）提出的 DAM 模型的整体结构如图 1 所示，可以分为输入、表示、匹配、聚合四个部分。模型的输入是多轮对话和候选回复，输出是每个候选回复的得分。

第十九届中国计算语言学大会论文集，第313页-第323页，海南，中国，2020年10月30日至11月1日。

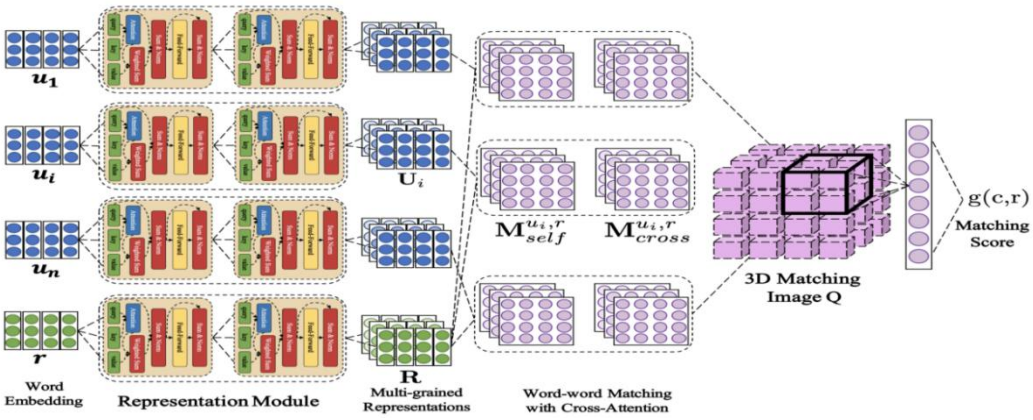

图 1. DAM 模型的整体结构

DAM 模型的注意力模块含有查询向量 Q、键向量 K 和值向量 V 三个输入。模块首先利用以下公式计算输入的缩放点积注意力：

$$\mathrm{V}_{\mathrm{att}} = Attention(Q,K,V) \tag{1}$$

之后模块将 V_{att} 和 Q 直接相加，产生的和包含二者的联合语义信息。为防止梯度消失或梯度爆炸，对 V_{att} 和 Q 相加的结果应用层归一化，结果记为 V_{ln}。接着将 V_{ln} 传入一个基于 ReLU 函数激活的双层前馈网络 FFN，进一步处理融合信息。FFN 的输出与输入进行一次残差连接，产生的结果再次应用层归一化，此时的 O 是整个注意力模块计算过程的最终输出，将其表示为：

$$O = AttentionModule(Q,K,V) \tag{2}$$

DAM 模型的语义表示网络由多个相同的注意力模块首尾相连，形成堆叠的网络结构。网络中每个注意力模块的三个输入相同，计算自注意力，公式表示为：

$$U_{\mathrm{i}}^{l+1} = AttentionModule(U_i^l,U_i^l,U_i^l) \tag{3}$$

$$R^{l+1} = AttentionModule(R^l,R^l,R^l) \tag{4}$$

其中 l 的范围从 0 到 $L-1$，L 表示模块堆叠的数量，$U_i^0 = u_i$ 以及 $R^0 = r$ 是原始输入。

DAM 模型中特征匹配的输入是语义表示网络的输出 U_i 和 R。针对不同粒度 l，网络将产生两种匹配矩阵，一种是自匹配矩阵 $M_{self}^{u_i,r,l}$，另一种是互匹配矩阵 $M_{cross}^{u_i,r,l}$。$M_{self}^{u_i,r,l}$ 是 U_i 和 R 的点积，矩阵中含有 U_i 和 R 元素间的语义依赖关系。$M_{cross}^{u_i,r,l}$ 的产生基于对注意力模块输入的修改，通过令 U_i 和 R 中对应元素互相关注，构造出新的语义表示 $\widetilde{U}_i^l$ 和 $\widetilde{R}^l$，用于捕捉跨越对话表达与候选回复之间的交叉关联特征。二者的计算公式为：

$$\widetilde{U}_{\mathrm{i}}^l = AttentionModule(U_i^l,R^l,R^l) \tag{5}$$

$$\widetilde{R}^1 = AttentionModule(R^l,U_i^l,U_i^l) \tag{6}$$

通过计算 $\widetilde{U}_i^l$ 和 $\widetilde{R}^l$ 的点积，得到 $M_{cross}^{u_i,r,l}$，DAM 模型中的特征融合网络将多个粒度的 $M_{self}^{u_i,r,l}$ 和 $M_{cross}^{u_i,r,l}$ 作为输入，在 i 和 l 两个维度拼接两个矩阵，得到矩阵 $P_{i,l}$ 如下：

$$P_{\mathrm{i,l}} = M_{self}^{u_i,r,l} \oplus M_{cross}^{u_i,r,l} \tag{7}$$

在 DAM 模型中，$P_{i,l}$ 被称为像素点，由 $P_{i,l}$ 组合形成的高维矩阵 P 被称为图像，这样的命名是为了方便使用 CNN。P 中的图像深度对应于多轮对话的轮次，图像宽度对应于每轮对话和候选回复在句子层级的匹配信息，图像高度对应于每轮对话和候选回复在单词层级的匹配信息。由于 P 含有三个维度的特征，DAM 模型采用 3D 卷积进行特征提取。经过两次 3D 卷积和最大池化，P 最终变成一维特征 f，再经过一个线性分类器即可获得匹配分数 $g(c,r)$。

3 基于多头注意力机制和 BiLSTM 网络的 Ex-DAM 模型

3.1 基于多头注意力和 BiLSTM 的 Ex-DAM 模型

为使 DAM 模型更适合处理含有细微变化的数据，进一步提高选择目标候选回复的准确率，本文利用多头注意力表示网络和双通道特征融合网络，结合 DAM 模型中的交互匹配网络，基于此构成一个新的端到端检索式多轮对话系统模型，将该模型命名为基于多头注意力和 BiLSTM 的 Ex-DAM 模型。模型的整体结构如图 2 所示。

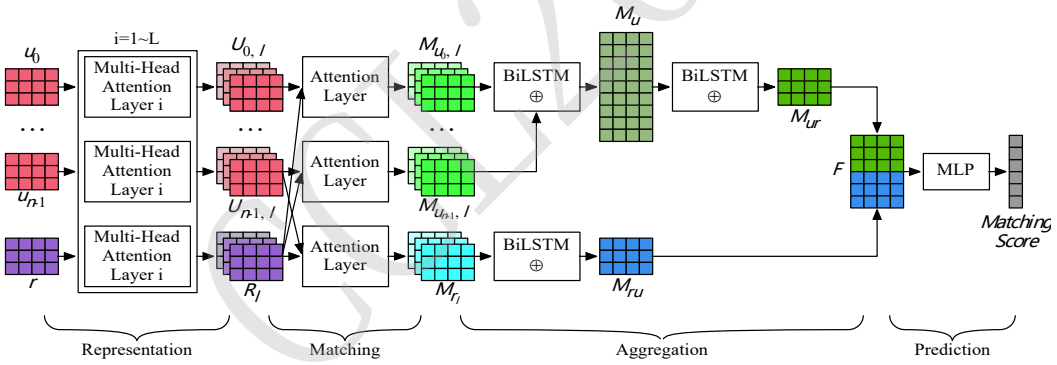

图 2. Ex-DAM 模型整体结构

模型的输入是词向量形式的多轮对话和候选回复，首先经过 L 个多头注意力层以获取它们的多粒度表示。在这些表示向量中，每一轮对话都和候选回复进行一次普通的注意力计算，得到多个主匹配矩阵。此外，候选回复再额外地与最后一轮对话计算一次注意力，以获得次匹配矩阵。随后，主、次匹配矩阵分别作为两个通道的输入进行特征融合。在这个过程中，所有的匹配矩阵经过 BiLSTM 和拼接操作依次进行序列特征提取和维度统一。最后，把两个通道的输出向量首尾拼接，经过多层感知器就能获得每个候选回复与多轮对话之间的匹配分数。

3.2 Ex-DAM 模型中的多头注意力模块和语义表示网络

普通的注意力机制在文本序列中可以很好地提取词向量角度的关键信息，但几乎无法识别对词向量进行统一修改的操作。多头注意力机制正好可以解决此类问题，在计算时首先输入多次映射，每个映射使用不同参数进行相同计算，最后将各个输出合并在一起，因此比缩放点积注意力更适合处理含有细微变化的数据，本文使用多头注意力模块结构如图 3 所示。

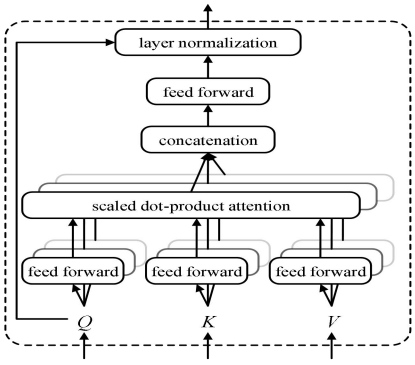

图 3. 多头注意力模块

多头注意力模块含有查询向量 Q、键向量 K 和值向量 V 三个输入，整体计算公式为：

$$MHAModule(Q,K,V) = LayerNorm(Q + MultiHead(Q,K,V)) \quad (8)$$

本文在模块中应用了残差连接，将输入 Q 与前馈层的输出恒等叠加，不会引入额外的参数，也不会增加模型的计算复杂度，在叠加过程中可强化输入中的重点内容，提升训练效果。多头注意力的头部能在不同子空间处理同一序列，从而获得更丰富的语义表示信息。由多头注意力模块组成 Ex-DAM 模型的语义表示网络，结构如图 2 中的 Representation 模块所示。

输入 $[u_i]_{i=1}^{i=T}$ 和 $[r_i]_{i=1}^{2\le i\le 10}$ 分别是每轮对话和候选回复的词向量，这些词向量已经过增强处理，同时含有意图信息和语义信息。词向量进入多头注意力模块计算语义表示的公式为：

$$U_{\mathrm{i}}^{l+1} = MHAModule(U_i^l, U_i^l, U_i^l) \quad (9)$$

$$R^{l+1} = MHAModule(R^l, R^l, R^l) \quad (10)$$

以 $U_i^0 = u_i$ 和 $R^0 = r$ 为多头注意力模块的原始输入，经 $L-1$ 个堆叠的相同模块可逐层获得多粒度的语义表示，即 $U_i = [U_i^0, ..., U_i^L]$ 和 $R = [R_i^0, ..., R_i^L]$。由于后续处理的计算量较大，为避免内存溢出需控制输入数量，本文针对 U_i 和 R 中的元素选择制定了以下三个处理策略：

第一个策略是当 L 数值较小时，保留所有 U_i 和 R，即使用所有粒度的语义表示作为特征匹配网络的输入，记作 Ex-DAM$_{\mathrm{L}}$；第二个策略是当 L 数值较大时，保留 U_i 和 R 中的后 m 个元素，即仅使用深层粒度的语义表示作为特征匹配网络的输入，记作 Ex-DAM$_{\mathrm{L\text{-}m}}$；第三个策略是当 L 数值较大时，保留 U_i 和 R 中的第一个元素和后 m 个元素，将原始输入同时作为语义表示网络和特征匹配网络的输入，而后者的输入还包含原始输入的多粒度语义表示，记作 Ex-DAM$_{\mathrm{L\text{-}0\text{-}m}}$。

3.3 Ex-DAM 模型中的双通道 BiLSTM 特征融合网络

本文提出的 Ex-DAM 模型的特征匹配网络仿照 DAM 模型，以语义表示网络的输出 U_i 和 R 作为输入，利用 DAM 模型中的注意力模块构造自匹配矩阵 $M_{self}^{u_i,r,l}$ 和互匹配矩阵 $M_{cross}^{u_i,r,l}$，如图 2

中的 Matching 模块，由于本文在该网络中未做改动，这里不做赘述。在 DAM 模型的特征融合网络中，3D 卷积作为注意力机制的一种辅助策略，在进一步提取特征的同时缩小数据维度。由于本文已将缩放点积注意力替换为多头注意力，因此本文在 Ex-DAM 模型中也将 3D 卷积进行了替换，Ex-DAM 模型的特征融合网络如图 2 中 Aggregation 模块所示。

在 Ex-DAM 模型的特征融合网络中，文本首先对 $M_{self}^{u_i,r,l}$ 和 $M_{cross}^{u_i,r,l}$ 进行加权求和，提取匹配矩阵中的重要匹配特征，其中 w_l 是共享权重，作用是增强模型的泛化性并减少计算开销。计算得到的 M_{self}^{i} 含有每轮对话与前一轮对话之间的匹配依赖信息，而 M_{cross}^{i} 中含有每轮对话与候选回复之间的匹配依赖信息。

接下来，本文利用两个不同的 BiLSTM 网络分别处理 M_{self}^{i} 和 M_{cross}^{i}，提取其中细微片段之间的匹配关系。与原始 DAM 模型中使用的 3D 卷积不同，Ex-DAM 模型放弃从“轮次—对话—候选回复”的角度入手，转而考虑以轮次为单一主线，分别对“对话—对话”以及“对话—候选回复”进行独立计算。同时，由于将在实验中引入基于意图识别的词向量增强，以累加形式嵌入到对话中的意图嵌入向量也较容易引起 BiLSTM 的关注。将代表不同轮次的隐藏状态矩阵首尾拼接，便可得到两种不同的融合特征矩阵——M_{self}^{agr} 和 M_{cross}^{agr}。特征融合网络的末尾部分与之前的相关研究保持一致，采用全连接神经网络处理 M_{self}^{agr} 和 M_{cross}^{agr}，将其中蕴含的融合特征表示为一组匹配分数，并使用 softmax 函数进行归一化处理，形成匹配概率。

4 数据集和实验设置

4.1 数据集与数据预处理

本文使用的中文多轮对话数据集是（Wu Y et al.,2016）提供的豆瓣对话数据集和（Zhang Z et al.,2018）提供的电商对话数据集。这两个数据集已由提供者进行了中文分词处理，每个数据集含有六个文件，其中 responses.txt 用于索引特定候选回复，word2vec.txt 用于预训练词向量，vocab.txt 用于索引特定单词，test.txt 作为测试集，train.txt 作为训练集，valid.txt 作为验证集。本文针对两个数据集中的数据组成进行了统计，统计结果如表 1 所示。

表 1. 数据集统计结果

统计指标	豆瓣对话数据集			电商对话数据集		
	训练集	验证集	测试集	训练集	验证集	测试集
多轮对话总数	495389	25000	1000	386478	5003	1000
候选回复总数	1000000	50000	10000	1000000	10006	10000
多轮对话轮次数最小值	3	3	3	1	1	1
多轮对话轮次数最大值	98	91	45	119	111	49
一轮对话单词数最小值	1	1	1	1	1	1
一轮对话单词数最大值	10624	5617	862	207	94	100

数据预处理的第一部分是数据清洗，利用 response.txt 中的候选回复表，可索引 i^T 和 i^F 得到

2~10 个候选回复，记作 $R=[r_i]_{i=1}^{2\le i\le 10}$。统计 R 中每个候选回复 r_i 的单词数 n_i^r，若满足

$$n_i^r \le W_{\max} \tag{11}$$

表示 r_i 中的单词数符合常规，其中 $W_{\max}$ 是本文自行设置的单词处理最大值，本文设为 100。本文利用 C 中的<EOS>将 C 分割成多个单轮对话，同时根据<EOS>的个数快速统计 C 的轮次数，得到 $C=[c_i]_{i=1}^{i=t}$，其中 t 是轮次数，若满足

$$t \le T_{\max} \tag{12}$$

表示 C 的轮次数符合常规，其中 $T_{\max}$ 是本文自行设置的轮次处理最大值。此外，还需针对满足条件的 C 统计其中 c_i 的单词数 n_i^c，若满足

$$W_{\min} \le n_i^c \le W_{\max} \tag{13}$$

表示 n_i^c 中的单词数符合常规，其中 $[W_{\min},W_{\max}]$ 是单词处理阈值区间，$W_{\max}$ 与本文对 n_i^r 的限制相同，$W_{\min}$ 是特别针对 c_i 设置的单词处理最小值，本文设为 2。

随后，本文将清洗后得到的数据进行数据规范。针对候选回复 $R=[r_i]_{i=1}^{2\le i\le 10}$，本文在 $[W_{\min},W_{\max}]$ 区间内设置一个表达长度值 W，通过处理所有 r_i，使其满足

$$n_i^r = W \tag{14}$$

若 $n_i^r < W$，将若干特殊标识符<PAD>增加至 r_i 尾部，使其长度达标。<PAD>本身不具备语义，因此也不会影响到 r_i 的语义。若 $n_i^r > W$，删除超出部分的单词。

经过上述处理，每个多轮对话都由 T 轮对话组成，每轮对话和每个候选回复都由 W 个单词组成。借助 vocab.txt 中单词与词序之间的对应关系表，本文将数据中的所有单词（包括<EOS>）转换为数字，实现文本数据的向量化过程。

4.2 基于意图识别的词向量增强

得到上述规范数据，其中将单词转化成的数字同时与 word2vec.txt 文件中的预训练词向量一一对应，根据这种对应关系可将每个数字转换成 200 维的词向量，其中<PAD>的词向量是 200 维的零向量。r_i 和 c_i 均是维度为 $(W,200)$ 的向量，C 的维度是 $(T,W,200)$。

受位置编码的启发，本文引入意图嵌入向量以处理意图识别结果。针对候选回复集合 $R=[r_i]_{i=1}^{2\le i\le 10}$，取其正确候选回复集合 R^T，对含有某种意图的 C 和 R^T 同时采用如下的编码方式：

$$ID(a)=\frac{1}{10}\sin(1/10000^{1-\frac{d}{dm}}) \tag{15}$$

$$ID(b)=\frac{1}{10}\sin(1/10000^{\frac{d}{d_m}}) \tag{16}$$

本文将 200 维的意图嵌入向量直接与相应的词向量相加，由<PAD>构成的单词和对话保持不变，得到的结果即为词向量增强的结果。这样设计的好处是，对于相同意图的 C 和 R^T，二者的意图嵌入向量完全相同，增强了二者的相关程度。此外，预训练词向量的数值范围是 0~1，而意图嵌入向量的数值范围是 0~0.1，二者直接相加的数据变化对预训练词向量影响很小。

4.3 实验设置与结果分析

4.3.1 评价指标及实验设置

本文采用检索式多轮对话系统常用的几种评价指标，用来衡量模型的性能。假设多轮对话数据集 C 由 N 个集合 c 组成，每个集合 c 包含正确回复 t 个、错误回复 f 个。在整个数据集上计算各项评价指标的平均值，得到平均精度均值（Mean Average Precision，MAP）、倒数排序均值（Mean Reciprocal Rank，MRR）、首位准确率（Precision-at-1，P@1）和计算召回率（Recall-n-at-k，$R_n@k$）：

$$MAP=\frac{1}{N}\sum_{c\in C}AP(c) \tag{17}$$

$$MRR=\frac{1}{N}\sum_{c\in C}RR(c) \tag{18}$$

$$P@1=\frac{1}{N}\sum_{c\in C}P@1(c) \tag{19}$$

$$R_n@k=\frac{1}{N}\sum_{c\in C}R_n@k(c) \tag{20}$$

其中 AP 为平均精度（Average Precision），RR 为倒数排序指数（Reciprocal Rank）。

本次实验中使用的数据均已经过数据预处理，训练过程相关配置使用 Adam 优化器调节模型参数，DAM 模型和 Ex-DAM 模型超参数的取值均如表 2 所示。

表 2.DAM(Ex-DAM)模型超参数表

超参数	参数含义	参数值
W	表达长度值	50
T	对话轮次值	9
epoch	数据集迭代次数	3
layer	堆叠多头注意力模块数	5
batch_size	单次训练样本数	128
learning_rate	初始学习率	0.001
decay_step	学习率衰减步长	500
decay_rate	学习率衰减率	0.9

4.3.2 DAM 模型实验结果

本文先在两个数据集的训练集上训练 DAM 模型，然后在测试集上评估模型的性能，实验结果如表 3 所示。

表 3. DAM 模型实验结果

实验内容	豆瓣对话数据集				电商对话数据集		
	MAP	MRR	P@1	R_{10}@1	R_{10}@1	R_{10}@2	R_{10}@5
DAM模型	0.550	0.601	0.427	0.254	0.406	0.547	0.810
+词向量增强	0.539	0.583	0.409	0.238	0.392	0.530	0.798
+数据预处理	**0.556**	**0.606**	**0.434**	**0.259**	**0.414**	**0.557**	**0.819**
+数据预处理+词向量增强	0.548	0.587	0.425	0.248	0.396	0.539	0.807

由表 3 可以看出，数据预处理有助于模型性能的提升，对原始数据集进行数据预处理，在豆瓣对话数据集上的各项评价指标获得了 0.5%~0.7%的提升，在电商对话数据集上的各项评价指标获得了 0.8%~1%的提升。还可以看出，将基于意图识别的词向量增强直接应用于 DAM 模型产生了不理想的效果，在两个数据集上的各项评价指标均下降了 1%以上，即使经过数据预处理，模型效果有了略微提升，但也始终低于基线水平。此结果的产生原因主要是 DAM 模型完全由自注意力机制构造而成，其中的计算过程依赖于词向量，本文在词向量层面进行的任何改动都将逐层干扰注意力机制的计算，从而导致 DAM 模型性能急剧下降。

4.3.3 Ex-DAM 模型实验结果

为了验证本文提出的 Ex-DAM 模型是否有效，将经过数据预处理的两种实验模型作为基线模型，从是否进行词向量增强的角度进行了独立实验。其中 Ex-DAM$_5$ 表示堆叠注意力模块数设置为 5，保留所有U_i和 R，即使用所有粒度的语义表示作为特征匹配网络的输入；Ex-DAM$_{5\text{-}4}$ 表示模块数设置为 5，保留U_i和 R 中的后 4 个元素；Ex-DAM$_{5\text{-}0\text{-}4}$ 表示模块数设置为 5，保留U_i和 R 中的第 1 个元素和后 4 个元素，实验结果分别如表 4 和表 5 所示：

表 4. 不含词向量增强的 Ex-DAM 模型实验结果

实验内容	豆瓣对话数据集				电商对话数据集		
	MAP	MRR	P@1	R_{10}@1	R_{10}@1	R_{10}@2	R_{10}@5
DAM模型	0.556	0.606	0.434	0.259	0.414	0.557	0.819
Ex-DAM$_5$	**0.562**	**0.610**	**0.441**	**0.264**	**0.423**	**0.563**	**0.822**
Ex-DAM$_{5\text{-}4}$	0.557	0.605	0.437	0.260	0.419	0.561	0.819
Ex-DAM$_{5\text{-}0\text{-}4}$	0.559	0.607	0.438	0.259	0.420	0.561	0.820

表 5. 含有词向量增强的 Ex-DAM 模型实验结果

实验内容	豆瓣对话数据集				电商对话数据集		
	MAP	MRR	P@1	R_{10}@1	R_{10}@1	R_{10}@2	R_{10}@5
DAM模型	0.548	0.587	0.425	0.248	0.396	0.539	0.807
Ex-DAM$_5$	0.568	0.607	0.442	0.265	0.425	0.564	0.830
Ex-DAM$_{5\text{-}4}$	0.565	0.605	0.440	0.262	0.421	0.561	0.828
Ex-DAM$_{5\text{-}0\text{-}4}$	**0.570**	**0.615**	**0.448**	**0.270**	**0.427**	**0.566**	**0.831**

由表 4 和表 5 看出，无论是否对数据集进行词向量增强，Ex-DAM 模型的实际表现都优于基线模型。即使词向量增强曾在之前的实验中导致基线模型的性能不升反降，却帮助 Ex-DAM 模型达到了最佳性能，说明多头注意力机制与 BiLSTM 的共同作用要优于普通自注意力机制。

在表 4 中，Ex-DAM$_5$ 模型性能优于其余模型，该模型与其余模型的不同之处在于语义表示网络的输出含有 5 种粒度的语义表示。若将最底层粒度语义表示去除或以原始输入替换最底层语义表示，都将损失一部分模型性能。然而在表 5 中，以原始输入替换最底层语义表示的 Ex-DAM$_{5\text{-}0\text{-}4}$ 模型却比 Ex-DAM$_5$ 模型性能更优。这是由于对原始输入进行了词向量增强，导致原始输入含有额外的意图特征，而语义表示网络中每个粒度的语义表示都源于原始输入，相当于在计算过程中不断强化这种意图特征，促使模型重点对意图特征建模。

本文进行的实验均将堆叠注意力模块数设置为 5，为探究 Ex-DAM$_{L\text{-}0\text{-}m}$ 模型中 L 和 m 的取值对模型性能的影响，本文使用经数据预处理和词向量增强的电商对话数据集进行了额外的实验，将评价指标 $R_{10}@1$ 结果绘制成折线图 4 所示。

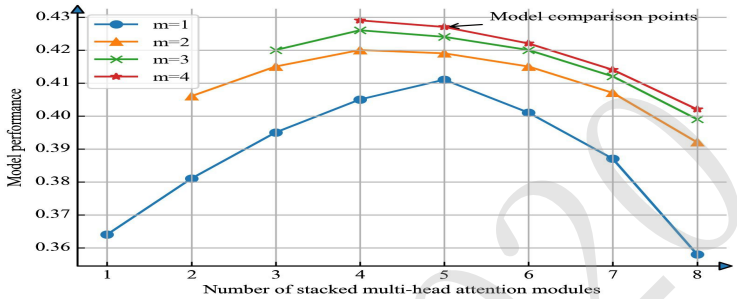

图 4. 不同参数搭配对 Ex-DAM$_{L\text{-}0\text{-}m}$ 模型的影响

本文在实验中始终保持语义表示网络的输出粒度不超过 5，这是由于模型占用的显存所致，若超出此值必须修改超参数，而计算方面的消耗将呈指数级增长，很难与之前的实验做对比。图 4 标记的模型对比点是上一实验中的 Ex-DAM$_{5\text{-}0\text{-}4}$ 模型，由图可知，多粒度语义表示确实能在一定程度上提升模型性能，当模型将 4 个堆叠多头注意力模块的输出与原始输入共同作为语义表示网络的输出时，通常能取得最高性能。若堆叠多头注意力模块数超过 5，无论如何选择 m 的值，模型都将逐渐出现过拟合现象，这是由于随着堆叠多头注意力模块数的增加，被多头注意力机制重点关注的信息会从前一层不断累加到下一层，导致这些信息在深层计算过程中基本保持不变，严重影响模型训练。

5 结论及后续工作

本文提出了基于多头注意力和 BiLSTM 改进的 DAM 模型 Ex-DAM，该模型用于处理中文多轮对话问答匹配问题。本文将 DAM 模型作为基线模型，利用多头注意力机制在多个不同子空间内计算特征，从而有能力建模较长的多轮对话。Ex-DAM 模型的卷积核使用 BiLSTM 来捕获序列上的依赖关系。实验证明，Ex-DAM 模型性能在电商和豆瓣的数据集上均优于基线模型。

在未来的研究中，将尝试加入命名实体识别、情感分析等多种辅助手段，使得 Ex-DAM 模型可以在文本片段中尽可能提取更多的特征。

致谢

*通信作者（chongzhy@vip.sina.com），本文承国家教育部人文社会科学研究规划基金资助项目（16YJAZH072）、国家社会科学基金重大项目（14ZDB156）、食品安全大数据技术北京市重点实验室资助。

参考文献

Ba J L, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.

Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval[M]. New York: ACM press, 1999.

陈晨, 朱晴晴, 严睿, 等. 基于深度学习的开放领域对话系统研究综述[J]. 计算机学报, 2019, 42(7): 1439-1466.

Glorot X, Bordes A, Bengio Y, et al. Deep Sparse Rectifier Neural Networks[C]. international conference on artificial intelligence and statistics, 2011: 315-323.

He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]. computer vision and pattern recognition, 2016: 770-778.

Kingma D P, Ba J. Adam: A Method for Stochastic Optimization[J]. Computer ence, 2014

Tran D, Bourdev L, Fergus R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]. international conference on computer vision, 2015: 4489-4497.

Voorhees E M. The TREC-8 question answering track report[C]//Trec. 1999, 99: 77-82.

Wu Y, Wu W, Xing C, et al. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots[J]. arXiv preprint arXiv:1612.01627, 2016.

Zhang Z, Li J, Zhu P, et al. Modeling multi-turn conversation with deep utterance aggregation[J]. arXiv preprint arXiv:1806.09102, 2018.

Zhou X, Li L, Dong D, et al. Multi-turn response selection for chatbots with deep attention matching network[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1118-1127.

左彬靖. 基于 word2vec 和自注意力机制的文本分类研究[D]. 广东工业大学, 2019.