计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(9):18−24 [doi: 10.15888/j.cnki.csa.007057]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向聊天机器人的多注意力记忆网络①

任建龙[1,2], 杨　立[1,2], 孔维一[1,2], 左　春[1,2,3]

[1](中国科学院 软件研究所 精准计算联合实验室, 北京 100190)
[2](中国科学院大学, 北京 100049)
[3](中科软科技股份有限公司, 北京 100190)
通讯作者: 任建龙, E-mail: ren_jian_long@163.com

摘　要: 如何对多轮的对话历史进行建模和推理是构建一个智能聊天机器人的主要挑战之一. 基于循环或门控的记忆网络已经被证明是进行对话建模的有效方式. 然而, 这种方式有两个缺点, 一是使用复杂的循环结构, 导致计算效率较低; 二是使用代价较大的强监督信息或先验信息, 不利于扩展和迁移应用到新的领域. 针对上述问题, 本文提出了一种端到端的多注意力记忆网络. 首先, 该网络采取结合词向量和位置编码的方式对文本输入进行表示; 其次, 使用并行的多层注意力在不同子空间捕获对话交互中的关键信息来更好地建模对话历史; 最后, 通过捷径连接的方式叠加多注意力层管理信息流, 实现对建模结果的多次推理. 在 bAbI-dialog 数据集上的实验表明, 该网络可以有效地对多轮对话进行建模和推理, 而且具有较好的时间性能.

关键词: 聊天机器人; 多轮对话; 多注意力; 捷径连接

引用格式: 任建龙,杨立,孔维一,左春.面向聊天机器人的多注意力记忆网络.计算机系统应用,2019,28(9):18–24. http://www.c-s-a.org.cn/1003-3254/7057.html

Memory Network with Multi-Head Attention for Chatbot

REN Jian-Long[1,2], YANG Li[1,2], KONG Wei-Yi[1,2], ZUO Chun[1,2,3]

[1](Laboratory of Precise Computing, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](SinoSoft Co. Ltd., Beijing 100190, China)

Abstract: Modeling and reasoning about the multi-turn dialogue history is a main challenge for building an intelligent chatbot. Memory Networks with recurrent or gated architectures have been demonstrated promising for conversation modeling. However, it still suffers from two drawbacks, one is relatively low computational efficiency for its complex architectures, the other is costly strong supervision information or fixed priori knowledge, which hinders its extension and application to new domains. This paper proposes an end-to-end memory network with multi-head attention. Firstly, the model adopts a method using word embedding combined with position encoding to represent text input; Secondly, it uses multi-head attention to capture important information in different subspaces of conversational interactions. Finally, multi-layered attention is stacked via shortcut connections to achieve repeatedly reasoning over the modeling result. Experiments on the bAbI-dialog datasets show that the network can effectively model and reason for multi-turn dialogue and has a better time performance.

Key words: chatbot; multi-turn dialogue; multi-head attention; shortcut connections

① 基金项目: 中国科学院 A 类战略性先导科技专项 (XDA20080200); 国家重点研发计划 (2018YFB1005002)
Foundation item: Strategy Priority Research Program of Chinese Academy of Sciences (XDA20080200); National Key Research and Development Program of China (2018YFB1005002)
收稿时间: 2019-02-27; 修改时间: 2019-03-22; 采用时间: 2019-03-29; csa 在线出版时间: 2019-09-05

近来, 人机对话由于其潜在的商业价值受到广泛关注[1–5]. 智能聊天机器人也就是对话系统在工业界得到了广泛的应用, 如苹果的 Siri, 谷歌的 Google Assistant, 阿里巴巴的阿里精灵, 小米的小爱, 百度的小度等. 根据用户的使用意图, 聊天机器人可以分为两种, 1) 目标导向的聊天机器人, 它的目的是帮助用户完成诸如订票、订餐、预定等特定任务; 2) 非目标驱动的聊天机器人, 旨在通过与人交互提供有意义的回复和娱乐功能, 也被称为开放域的聊天机器人或闲聊机器人. 传统的聊天机器人采用流水线的设计, 如图 1, 包括 1) 语音识别; 2) 自然语言理解; 3) 用户状态追踪; 4) 策略学习; 5) 自然语言生成; 6) 文本转化为语音共六个部分. 这种流水线的设计存在如下两个缺点: (1) 错误定位问题, 研发人员从最终客户得到关于系统质量的反馈, 而误差定位和分析需要大量繁重的工作, 因为上游的误差会传播到下游任务; (2) 过程的内部依赖, 给系统的自适应和调整带来了困难, 例如当其中一个模块使用新数据进行训练, 那么下游模块也需要重新训练, 而这个过程还需要大量的人工工作. 当前深度学习领域的成功大大促进了多轮对话系统的发展, 其中端到端的深度神经网络成为研究的重点. 与传统流水线方式相比, 端到端的方式不需要手工设计规则, 并且不存在错误定位和对齐问题[2]. 然而构建一个智能的对话机器人仍然面临着如何进行多轮交互式建模的问题, 需要机器人理解来自不同对话轮次中的历史对话信息和避免产生无意义的、冗余和重复的回复.

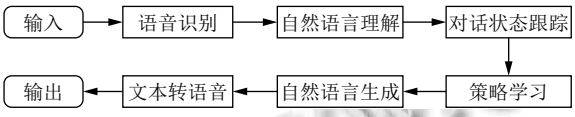

图 1 传统流水线式对话组件

循环神经网络 (Recurrent Neural Networks, RNNs), 具体包括长短期记忆 (Long Short-Term Memory, LSTM)[6]和门控循环神经网络 (Gated Recurrent Neural Network, GRU)[7], 作为强的基准方式被广泛应用在自然语言处理领域 (Natural Language Processing, NLP) 中, 如机器翻译、对话建模和序列标注[8–10]. 最近, 记忆网络被提出来用于自然语言理解中的语言建模[11–14]. 文献[13,14]使用 LSTM 或 GRU 结合注意力机制, 在聊天机器人的多轮对话中可以有效地进行建模. 然而现有的方法存在以下两个方面的不足:

(1) 计算效率较低

大多数方法使用 RNNs 类网络模型, 这种网络结构在建模时需要进行序列对齐操作, 执行计算时是串行的, 无法并行执行和有效利用当前高效的、加速计算的硬件资源, 而这在大规模数据集上是十分重要和有意义的[9];

(2) 依赖强监督信息和先验知识

一些方法使用强监督信息, 这些监督信息包括手工设计的特征函数, 代价较大, 同时不利于迁移到新的领域[13]. 还有的方法使用一些固定的先验知识[14], 同样不利于模型的扩展和迁移, 在特定领域构建这类模型需要领域专家制定大量的规则.

本文对深度神经网络在聊天机器人领域内的应用展开研究, 提出一种端到端的基于多注意力机制的记忆网络, 试图解决以上问题. 该网络具有相对简单的并行结构, 首先通过多注意力机制在不同子空间捕获对话轮次中的重要信息对对话历史进行建模, 其次使用捷径连接叠加多注意力层来控制信息流通和对话历史记忆的获取, 对历史对话信息建模结果进行多次推理.

本文的主要贡献如下:

(1) 使用多注意力机制对对话历史和上下文进行建模, 该机制允许模型学习去对齐关注相关重要的信息和获取输入和输出之前的全局依赖. 与当前 LSTM 或 GRU 使用的注意力机制不同的是, 多注意力机制具有相对简单的并行结构;

(2) 使用捷径连接叠加多注意力层, 形成多层的推理结构, 对记忆单元进行迭代推理, 记忆单元就是多注意力机制建模的最新结果. 该结构通过控制信息在前一层和当前层之间的流动来控制对记忆单元的动态读取.

本文剩余内容的组织方式如下, 第 1 节介绍背景知识, 第 2 节对本文提出的网络结构进行了详细阐述, 第 3 节描述了实验过程并对实验结果进行分析, 第 4 节介绍了相关工作, 第 5 节总结了本文工作.

1 背景知识

一个对话任务中一般包含历史对话c和当前用户输入u, 产生回复r, 可用如下公式表示:

$$r = f(c,u) \tag{1}$$

记忆网络[11–15]整体上提供一个可读写的记忆模块, 从而实现对大量长期的记忆进行建模和推理, 网络结构如图 2 所示, 包括输入、记忆、输出和响应 4 个基本组件.

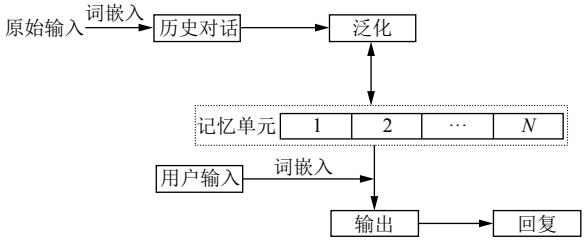

图 2 记忆网络的一般结构

给定输入x, 记忆网络各个模块的执行流程和作用如下:

(1) Input: 将原始输入x进行编码表示为一个低维紧凑的向量$R(x)$;

(2) Memory Generalization: 给定新的输入时更新记忆单元;

$$m_{i+1} = f(m_i, R(x_i)) \tag{2}$$

(3) Output: 使用注意力机制对当前输入和记忆单元进行读取和推理;

$$o = attention(m, R(x)) \tag{3}$$

(4) Response: 根据当前输出产生最终的回复.

$$r = decode(o) \tag{4}$$

记忆网络及其相关变体整体上都遵循这个结构, 在具体的一些表示方法、记忆更新以及读取方式上略有不同. 现有的大多数基于注意力机制的方法使用循环结构 LSTM 或 GRU, 这些循环结构相对复杂从而导致计算效率比较低. 有些方法使用强监督信息或先验知识, 不利于扩展和迁移应用.

2 多注意力记忆网络

针对现有记忆网络中存在的计算效率低和对额外信息输入的依赖问题, 本文提出一种端到端的多注意力记忆网络. 该网络结构如图 3 所示, 包括文本表示、多注意力层、捷径连接的堆叠结构, 以下对这三个部分分别展开阐述.

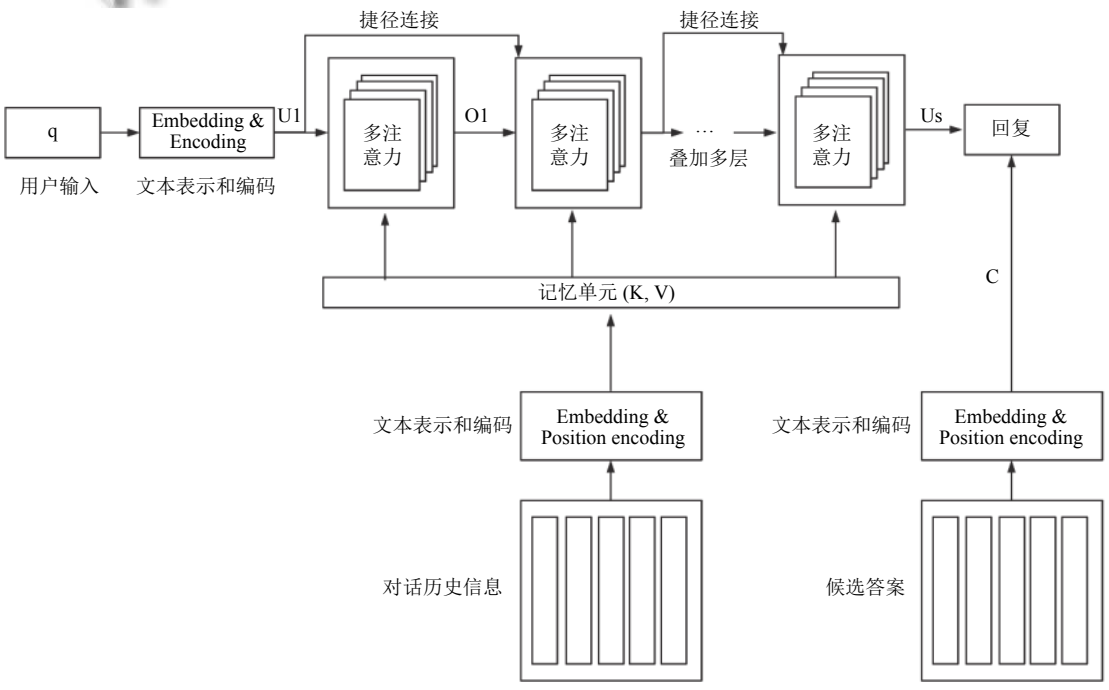

图 3 本文提出的多注意力记忆网络

2.1 文本表示和编码

深度神经网络具有自主的学习表征的能力, 因此也被称为表示学习. 词向量是神经网络训练语言模型的产物, 它是一个紧凑和低维的分布式向量表示, 维度经常在 20 至 500 之间取值, 相比传统的 one-hot 表示方式, 它包含了可学习的语义信息. 本文对于输入的文本采用直接嵌入的词向量, 在神经网络训练和学习的过程中固定词向量表示矩阵. 具体地, 词向量表示矩阵被初始化为E, 如公式 (5) 所示, 对于所有的输入包括用户输入、机器响应以及候选答案都使用这个矩阵, 目的是更充分地训练词向量矩阵, 词嵌入相当于一个查表操作.

$$Emb(word_i) = E(i) \tag{5}$$

因为基于多注意力的网络结构不能利用序列的位置信息, 因此本文在词向量的基础上采用位置编码 (Position Encoding, PE) 来获取单词的全局和局部信息. 如式 (6) 和式 (7) 所示, 其中pos表示单词的位置信息, i表示词向量的维度信息, d_{model}与词向量的维度一样, 使二者可以相加, 在本文中取值为 128.

$$PE(pos,2i) = sin(pos/10000^{2i/d_{model}}) \tag{6}$$

$$PE(pos,2i+1) = cos(pos/10000^{2i+1/d_{model}}) \tag{7}$$

最终文本的表示为词向量嵌入和位置编码的和, 即:

$$R(word_i) = Emb(word_i) + PE(word_i) \tag{8}$$

2.2 多注意力机制

在进行对话历史和上下文的建模之前, 先利用上一节所述的文本表示方法将输入转化为向量表示. 记忆单元的 key 表示为K, value 表示为V, 用户输入表示为Q. 这 3 个输入的维度是一样的, 与式 (6) 和式 (7) 中的d_{model}维度相同.

如图 4 所示, 左侧为单层点乘注意力, 对用户输入Q和记忆单元的 key-K 进行点积运算, 然后利用$softmax$函数计算注意力的值 (注意力的值分布在 0 至 1 之间), 如式 (9) 所示, 与基础的点乘注意力不同的是, 这里还对点乘的结果除以K的维度d_K也即 d_{model}.

$$Attention(Q_i,K_i,V_i) = \frac{e^{\left(Q_iK_i^T/\sqrt{d_k}\right)V_i}}{\sum_j e^{\left(Q_jK_j^T/\sqrt{d_k}\right)V_j}} \tag{9}$$

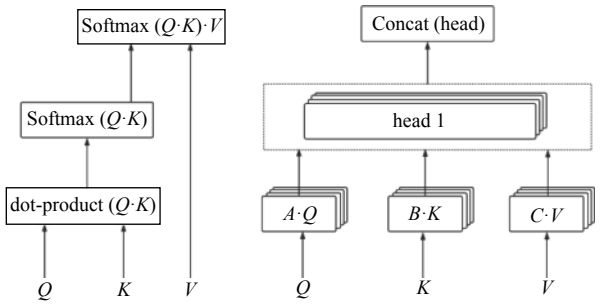

图 4 单层点乘注意力 (左), 并行多注意力 (右)

如式 (10) 所示, 多注意力机制[9]就是并行地执行单层点乘注意力, 这个并行执行过程是先将维度空间 (例如词向量维度) 划分为h个部分, 然后在每个部分分别计算单层的点乘注意力. 这样做的目的是通过在不同子空间进行注意力机制的计算, 使模型在不同子空间获得关于不同信息成分不同程度的关注, 进而获得全局或局部的依赖信息, 让模型去关注对话不同轮次中重要的信息. 此外, 对于三个输入部分乘以一个线性矩阵做线性映射的转换, 进一步加强神经网络的表征能力.

$$head_h = Attention(AQ,BK,CV) \tag{10}$$

最后, 这些并行的不同空间经过不同关注的信息值被按照顺序拼接, 然后再乘以一个线性矩阵做线性映射作为输出.

$$O = Concat(head_1,\cdots,head_H)W^O \tag{11}$$

2.3 捷径连接的多层推理结构

在记忆网络的相关研究中, 已经证明了一种称为 multi-hops 的重复多层结构对于记忆和用户输入的推理是有效的. 在本文的工作中, 通过叠加多注意力层实现这种重复推理的方法. 同时, 本文集成了捷径连接 (shortcut connections) 机制. 这种机制允许网络自主学习它应该从前一层流到下一层的信息量, 进而实现对于信息的灵活读写和获取.

捷径连接的机制首先应用于计算机视觉的图像识别任务[16], 在残差网络 (residual network) 中使用的方式是恒等映射 (identity mapping), 如式 (12) 所示, 其中$g(\cdot)$表示激活函数, x表示输入.

$$y = g(x) + x \tag{12}$$

恒等映射是一种固定的连接前一层与当前层的机制, 即它无法灵活自主得决定信息的流通量. 在文献[17]的 highway network 中采取可学习的门控连接来控制信息的流通量, 如式 (13) 所示, 其中⊗表示逐元素对应的乘法操作, $T(x)$是一个可学习的非线性转换函数, 本文中采取双曲正切$tanh$函数.

$$y = g(x)\otimes T(x) + x\otimes(1-T(x)) \tag{13}$$

如式 (14) 所示, 本文通过门控的捷径连接方式来堆叠多注意力层. 其中S表示堆叠的层数, o是式 (11) 中的输出, 表示上一多注意力层的输出. 这种门控机制是与输入相关的非线性函数, 在神经网络训练的过程中, 它可以自主调整到一个合适的参数空间, 进而控制信息的流通.

$$U_{S+1} = O\otimes\frac{1}{1+e^{-U_S}} + U_S\otimes\left(1-\frac{1}{1+e^{-U_S}}\right) \tag{14}$$

最终的输出选取最后堆叠层S的输出和候选答案做相关性计算, 选择概率最大的回复作为机器人的最终响应, 如式 (15) 所示, 其中U_1表示用户输入, U_S

表示堆叠层最后的输出, C表示候选答案的向量表示.

$$a = max\frac{e^{W_u(U_S+U_1)C^{\mathrm{T}}}}{\sum_n W_U(U_S+U_1)C_n^{\mathrm{T}}} \tag{15}$$

3 实验

为了验证多注意力记忆网络的有效性, 本文选取 bAbI 对话数据集[15]进行实验验证, 该数据集为订餐领域的多轮对话实例, 由 11 个任务组成, 涵盖回复准确性, 对话完成率和可扩展性等多方面要求. 实验环境为: Intel (R) Xeon (R) CPU E5-2640 @ 2.40 GHz, 32 GB 内存, NVIDIA Quadro P2000 显卡, Centos 7. 3 操作系统.

3.1 数据集

bAbI 对话数据集是从典型的面向任务的多轮对话系统中收集或生成的. bAbI 对话数据集包括 6 个任务: 1) T1: 发出 API 调用, 2) T2: 更新 API 调用, 3) T3: 显示选项, 4) T4: 提供额外信息, 5) T5: 进行完整对话 (上述 4 个任务的集成), 6) T6: 对话状态跟踪挑战语料库 (DSTC 2). 所有 6 个任务都属于餐厅预订领域. 每项任务, 有 1000 个对话用于训练, 1000 个对话用于验证, 1000 个对话用于测试. 对于任务 1-5, 提供了包含对话的第二测试集 (具有后缀-OOV. txt), 所述对话包括训练和开发集中不存在的实体. 任务 6 由真实的人机对话生成, 其数据来自第二对话状态跟踪挑战[18]. DSTC2 数据集最初是为对话状态跟踪而设计的 (每一轮都标有状态: 用户意图+插槽). 因此, 为了评估端到端模型的性能, 文献[15]将该数据集转换为与上述 5 个任务格式相同的第 6 个任务. 任务 6 与前 5 个任务具有类似的统计信息, 由于其包含部分语音识别错误并且存在噪声信息, 因此具有更高的难度.

3.2 实验结果

bAbI 对话数据集主要指标是每轮响应准确率, 即每次对话机器人回复正确的轮次占所有对话轮次的比例. 本文选取监督词向量 (LTR-SEM)[15]、循环端到端记忆网络 (MEMR)[13]作为与多注意力记忆网络 (MEMMA) 比较的标准. 如图 5 左侧所示为每轮响应准确率的比较, LTR-SEM 除 T1 外准确率均低于 MEMR 和 MEMMA, 在 T1, T2, T2-OOV, T3, T3-OOV 这五个任务上 MEMR 和 MEMMA 的表现相近, 其中 T1 和 T2 两个对话任务两个模型都获得 100% 的准确率, 但是在 T1-OOV, T4, T4-OOV, T5, T5-OOV 这 5 个任务上, 本文提出的模型表现优于 MEMR, 在所有 10 个任务上的平均准确率提高了 1.65%. 在每轮对话时可以添加匹配特征, 包括对话轮次的时间信息和对话发起者 (机器人或用户), 添加匹配特征的方法后, MEMR 和 MEMMA 关于每轮响应准确率的比较结果如图 5 右侧所示, 在 T1, T3, T4, T4-OOV 这 4 个任务上二者的表现相近, 其中 T1, T3, T4-OOV 都取得了 100% 的准确率, 而 MEMMA 在 T2, T2-OOV, T3-OOV, T5, T5-OOV 这 5 个任务上的准确率都高于 MEMR, 在所有 10 个任务上的平均准确率, 前者比后者高 1.46%. 第 6 个任务来自真实对话场景的数据, 该数据中包含语音识别的错误和噪音, 同时又因为其从填槽对话中转换为端到端的类型, 任务整体难度更大. 在该任务上的实验结果如表 1 所示, 在每轮对话回复准确率的比较上, MEMMA 比 LTR-SEM 和 MEMR 分别提高 18.5% 和 1.4%; 在添加匹配特征后的比较中, MEMMA 相比 MEMR 在每轮回复准确率上高 2.6%. 实验结果表明, 在多轮对话的建模中, 本文提出的 MEMMA 比 MEMR 更加有效.

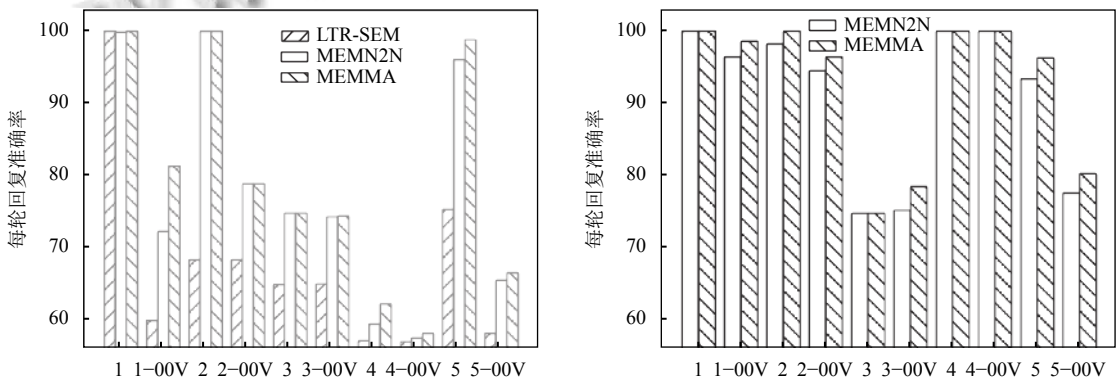

图 5 左侧为每轮回复准确率比较, 右侧为添加匹配特征的的每轮回复准确率比较

多注意力记忆网络具有相对简单的并行结构, 为了验证其有助于提高计算效率, 针对循环端到端记忆网络[13]和本文提出的网络, 我们选取二者在相同实验环境下的训练时间进行比较. 实验结果如表 2 所示, 取 200 次迭代训练的时间 (重复 10 次取平均值) 作为比较基准, 在全部 11 个任务中, 本文提出的网络所需的训练时间相比 MEMR 均有所减少. 实验结果表明, 本文提出的多注意力记忆网络大约提升了 50% 的计算效率, 这是因为对于给定一个长度为n的序列, 循环结构的注意力执行序列操作的时间复杂度是O(n), 而并行的多注意力机制执行序列操作只需要常数级的时间复杂度O(1).

表 1 任务 6 即 DSTC2 数据上的每轮回复准确率比较

	LTR-SEM	MEMR	MEMMA	添加匹配特征	
				MEMR	MEMMA
T6-DSTC2	22.6	41.1	**42.5**	41.0	**43.6**

表 2 模型训练时间的比较 (单位: s)

	T1	T1-OOV	T2	T2-OOV	T3	T3-OOV	T4	T4-OOV	T5	T5-OOV	T6
MEMR	300	306	502	503	660	662	173	173	1158	1156	760
MEMMA	**200**	**204**	**334**	**342**	**439**	**438**	**120**	**118**	**760**	**763**	**512**

4 相关工作

在大多数 NLP 任务中, 例如机器翻译, 文本分类和对话建模[8,9,19], 现有工作已经证明循环或门控网络和自注意力机制的结合可以有效地进行语言建模. 这种序列对齐的循环结构使得其难以解决计算效率、梯度消失和长期依赖性的问题. 最近, 一系列语言建模和语言理解专注于记忆类网络. 记忆类网络可以通过学习读写内容来推理长期记忆模块, 并且已被广泛应用, 如语言建模[11], 问答[13,14]和多轮对话[15]. 相关的工作最早提出的是非端到端的记忆网络, 这种结构依赖额外的强监督信息输入, 所以并不是端到端的[20]. 后来的工作改进了记忆网络的组件, 例如, 读, 写和记忆模块. 一些研究使用动态记忆机制来调节记忆的相互作用, 但它具有代价较高的的手工特征功能和较复杂的循环或门控结构[13,14]. 本文提出基于多注意力机制[9]的记忆网络, 整体结构与记忆类网络类似, 包含可供读写的记忆单元, 但与现有记忆网络的不同之处在于: 1) 使用并行的多注意力机制代替循环注意力机制, 结构较简单可以提高计算效率, 同时可以更有效地捕获全局依赖关系; 2) 未依赖额外的监督信息或先验知识, 可以进行端到端的训练.

在神经网络优化的背景下, 梯度消失问题已经被研究了很长时间[21], 文献[16]提出了一种用于图像识别的残差连接 (residual connections), 已被证明可有效地克服消失梯度问题. 类似的工作是 Highway Network 采用可微分的门控机制[17]. 本文引入捷径连接机制, 将多个多注意力层与捷径连接相结合构建叠加的多层重复结构.

5 结论

本文提出一种多注意力记忆网络来对聊天机器人进行多轮对话的建模和推理. 具体而言, 首先采用关注不同对话轮次重要信息的多注意力机制对会话语境和历史记忆进行建模, 该机制具有相对简单和并行的结构. 然后提出了具有捷径连接的叠加多层推理结构. 与现有方法相比, 本文提出的网络不依赖于代价较大的附加监督信息或先验知识, 提供了一种更为简洁的端到端方式. 通过在 bAbI 数据集所有 11 个任务上的实验表明, 多注意力记忆网络可以有效地建模和推理多轮对话交互, 性能优于循环端到端记忆网络.

参考文献

1 商雄伟, 张志祥. 限定领域智能导学系统问题生成及对话管理技术. 计算机系统应用, 2015, 24(11): 242–246. [doi: 10.3969/j.issn.1003-3254.2015.11.041]

2 Chen HS, Liu XR, Yin DW, *et al*. A survey on dialogue systems: Recent advances and new frontiers. ACM Sigkdd Explorations Newsletter, 2017, 19(2): 25–35. [doi: 10.1145/3166054]

3 Huang YF, Li ZC, Zhang ZS, *et al*. Moon IME: Neural-based Chinese pinyin aided input method with customizable association. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations. Melbourne, Australia. 2018. 140–145.

4 Zhang ZS, Li JT, Zhu PF, *et al*. Modeling multi-turn conversation with deep utterance aggregation. Proceedings of the 27th International Conference on Computational Linguistics. NM, USA. 2018. 3740–3752.

5 Qiu MH, Li FL, Wang SY, *et al*. AliMe Chat: A sequence to

sequence and rerank based chatbot engine. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada. 2017. 498–503.

6 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

7 Chung J, Gulcehre C, Cho K H, *et al*. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv: 1412.3555, 2014.

8 Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada. 2014. 3104–3112.

9 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Advances in Neural Information Processing Systems. Long Beach, CA, USA. 2017. 6000–6010.

10 夏天赐, 孙媛. 基于联合模型的藏文实体关系抽取方法研究. 中文信息学报, 2018, 32(12): 76–83. [doi: 10.3969/j.issn.1003-0077.2018.12.010]

11 Sukhbaatar S, Weston J, Fergus R, *et al*. End-to-end memory networks. Proceedings of the 29th Annual Conference on Neural Information Processing Systems. New York, NY, USA. 2015. 2440–2448.

12 Miller A H, Fisch A, Dodge J, *et al*. Key-Value Memory networks for directly reading documents. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Austin, TX, USA. 2016. 1400–1409.

13 Kumar A, Irsoy O, Ondruska P, *et al*. Ask me anything: Dynamic memory networks for natural language processing. Proceedings of the 33rd International Conference on Machine Learning. New York, NY, USA. 2016. 1378–1387.

14 Xiong CM, Merity S, Socher R. Dynamic memory networks for visual and textual question answering. Proceedings of the 33rd International Conference on Machine Learning. New York, NY, USA. 2016. 2397–2406.

15 Bordes A, Boureau YL, Weston J. Learning end-to-end goal-oriented dialog. arXiv: 1605.07683, 2016.

16 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778.

17 Srivastava RK, Greff K, Schmidhuber J. Highway networks. arXiv: 1505.00387, 2015.

18 Henderson M, Thomson B, Williams JD. The second dialog state tracking challenge. Proceedings of the SIGDIAL Conference. Philadelphia, PA, USA. 2014. 263–272.

19 Weston J, Chopra S, Bordes A. Memory networks. Proceedings of International Conference on Learning Representations. New York, NY, USA. 2015.

20 Devlin J, Chang M W, Lee K, *et al*. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv: 1810.04805, 2018.

21 Srivastava RK, Greff K, Schmidhuber J. Training very deep networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada. 2015. 2377–2385.