软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(11):2905−2924 [doi: 10.13328/j.cnki.jos.005334]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

记忆神经网络的研究与发展*

梁天新, 杨小平, 王 良, 张永俊, 朱艳丽, 许 翠

(中国人民大学 信息学院,北京 100872)

通讯作者: 王良, E-mail: wangliang@ruc.edu.cn

摘 要: 首先,根据记忆神经网络训练形式的不同,介绍了强监督模型和弱监督模型的结构特征和各自应用场景以及处理方式,总结了两类主要模型的优缺点;随后,对两类模型的发展和应用(包括模型创新和应用创新)进行了简要综述,总结了各类新模型在处理自然语言过程中所起的关键作用;最后梳理了记忆神经网络处理自然语言所面临的复杂性挑战,并预测了记忆神经网络未来的发展方向.

关键词: 自然语言处理;人工智能;记忆神经网络;递归神经网络;机器学习;问答

中图法分类号: TP181

中文引用格式: 梁天新,杨小平,王良,张永俊,朱艳丽,许翠.记忆神经网络的研究与发展.软件学报,2017,28(11):2905−2924. http://www.jos.org.cn/1000-9825/5334.htm

英文引用格式: Liang TX, Yang XP, Wang L, Zhang YJ, Zhu YL, Xu C. Review on research and development of memory neural networks. Ruan Jian Xue Bao/Journal of Software, 2017,28(11):2905−2924 (in Chinese). http://www.jos.org.cn/1000-9825/5334.htm

Review on Research and Development of Memory Neural Networks

LIANG Tian-Xin, YANG Xiao-Ping, WANG Liang, ZHANG Yong-Jun, ZHU Yan-Li, XU Cui

(School of Information, Renmin Universityof China, Beijing 100872, China)

Abstract: Firstly, in this paper, the key features of memory neural networks in the strongly supervised model and the weakly supervised model and introduced. Then the corresponding application scenarios and processing methods, as well as the advantages and disadvantages of the two models are summarized. Next, a brief survey on the development and application of the two models (including the innovation on the model and the innovation in application) is provided, and the key roles of individual innovative models in the natural language processing are summarized. Finally, the complex challenges of memory neural networks in the natural language processing and the future development of memory neural networks are also discussed.

Key words: natural language processing; artificial intelligence; memory neural network; recurrent neural network; machine learning; question answering

1 记忆神经网络的提出与发展

2006 年,多伦多大学的 Hinton 等人提出了深度置信网络(deep belief networks,简称 DBN)[1],自此,深度学习(deep learning,简称 DL)的热潮开始兴起,DL 成为机器学习领域中最有发展潜力的研究方向,特别是伴随着几个神经网络模型的兴起,在图像识别[2]、游戏[3]、语言生成[4]、语音识别[5−7]等领域大放异彩.DL 成为推动人工智能(artificial intelligence,简称 AI)进步的重要手段.近年来,随着 DL 的出现,深层架构的各类神经网络模型对各类不同行业产生了积极的影响,比如 1989 年由 leCun 等人提出的卷积神经网络(convolutional neural network,简称

* 基金项目: 国家自然科学基金(61772537)
Foundation item: National Natural Science Foundation of China (61772537)
本文由复杂环境下的机器学习研究专刊特约编辑张长水教授推荐.
收稿时间: 2017-01-09; 修改时间: 2017-04-11; 采用时间: 2017-06-16

CNN)[8],Hinton 等人应用 CNN 在 2012 年的 ImagesNet[2]比赛中拔得头筹.随着递归神经网络(recurrent neural network,简称 RNN)[9,10]的提出,为神经网络带来了一定的短时记忆能力.1997 年,长短记忆神经网络(long short-term memory,简称 LSTM)[11]的出现,更是让神经网络的记忆能力有了大幅度的提升.Mikolov 等人将 LSTM 应用在语言建模上[10−13]、语法识别[14]、机器翻译[15],甚至在一些游戏上有了非常优异的表现[3].

长久以来,递归神经网络的记忆能力不断地被开发挖掘,同时,新的记忆模型也不断涌现.2014 年就有两个很有影响力的模型被提出来,一个是 Google DeepMind 的神经图灵机(neural turing machine,简称 NTM)[16],另一个是 Facebook 人工智能实验室的记忆神经网络(memory neural network,简称 MemNN)[17].这类模型将外部记忆方式引入到神经网络之中,解决了神经网络长程记忆的问题,开创记忆模型的新领域.

计算机的记忆机制历史久远,从计算机被发明开始,存储就是计算机的核心部件.神经网络最开始提出的模型不依赖于硬件的方式存储信息,而是将通过训练得到的特征提取能力和记忆能力保存在各种层之中.最常用的方式是标准递归神经网络模型,如果用这种模型进行自然语言的信息挖掘,比如对话生成[18]、信息检索[19]、语言建模[10−13]等,则是依靠数据集对递归神经网络模型进行训练,最终预测单词流的输出.RNN 模型的隐藏状态本身就是一种记忆,Yoshua 等人总结了 RNN 由于存在梯度消失问题[20]而无法从长距离中进行依赖学习,而 LSTM 则通过门控制机制对此做了改善,允许记忆被显式地更新和删除.然而隐藏节点和权重所能记住的数据往往十分有限,此模型往往无法记忆足够多的回答信息,在问答类问题中,对需要两种以上支持信息才能回答的问题效果欠佳.这是因为 LSTM 中记忆仅仅是被压缩成为数据向量,且它只能记住一些有一定内在规律和特征的信息,对于完全不相关信息并没有足够的记忆能力.

LSTM 擅长于语言建模,它也能被作为问答模型使用[21−23].这是因为某些类型的问答实际上更像是语言建模,而非对问答中支撑因子的理解,一旦涉及到含义问答或推理的时候,LSTM 就显得很无力.特别是针对一些需要长程记忆的信息,如小说和长篇评论等,没有足够的能力进行理解、推断或者回答问题.正因为以上原因,机器学习领域长久以来一直缺少一种能够对长程记忆进行读写,并结合长期记忆进行推理和演绎的模型.

20 世纪 90 年代,为了解决神经网络长程记忆的问题,Das 等人曾经引入过用“栈”的方式读取长程记忆[24],Tsungnam 等人也曾经利用 RNN 解决长程依赖的问题[20].实验结果表明,改进的 RNN 可以比传统的递归神经网络保留 2 倍~3 倍长的信息,在非线性问题识别上取得了一定的进步,但以上尝试皆没有彻底解决如何对长程记忆进行反复读取的问题.

2014 年,Facebook 人工智能实验室的 Weston 等人提出的 MemNN[17]在解决上述问题时有良好的表现.MemNN 不像 RNN 或 LSTM 那样舍弃训练信息,它保存全部训练信息在记忆模块中.因此,MemNN 在 bAbI 问答数据集[25]上得到了远超其他模型的表现.MemNN 的核心思想是:将有推论能力的机器学习方法与可读写的记忆模块结合在一起,在训练模型推理能力的同时,构建长效的动态知识库.

2015 年,Facebook 人工智能实验室的 Sukhbaatar 又在原实验的基础上提出了新的实验模型,即弱监督模型(weakly supervised mode)[26],原来的模型被称为强监督模型(strongly supervised mode)[17].相对于原来的强监督模型,弱监督模型是一种端到端的训练模式,减少了中间训练步骤.虽然目前在 bAbI 数据集的某些子集上表现稍逊于强监督模型,但是端到端的训练方式节省了大量数据标注,也更加容易应用于其他类型的数据集上,比如情感分析[27]、对话训练[28];同时,弱监督模型引入的软关注机制(soft attention mechanism)不仅简化了模型,而且提高了训练的准确率.

强监督模型在自然语言处理的应用中不断拓展,应用在更多的数据集上,比如:Facebook 人工智能实验室在新的文献[29]中改进了模型并尝试新的数据集;Bordes 等人开始利用知识库进行问答训练[30],并尝试了知识库下的迁移学习;Chandar 等人则以新的存储方式来适应大数据下的应用[31].

弱监督模型是在强监督模型上的一种模型创新,应用领域继续扩大,随之产生了许多改进模型,如 Dodge 等人尝试将模型应用在了多种文本领域[28],Felix Hill 等人尝试将硬关注机制(hard attention)加入到监督模型中[32],Miller 等人尝试了对信息源的索引和非索引部分用不同嵌入式矩阵进行映射[33],哈尔滨工业大学的唐都钰博士等人则将弱监督模型简化后应用在情感分析领域[27],Chen 等人在弱监督模型中加入 RNN 应用在口语指

令的识别上[34].综上所述,记忆神经网络近两年的发展概况如图 1 所示.

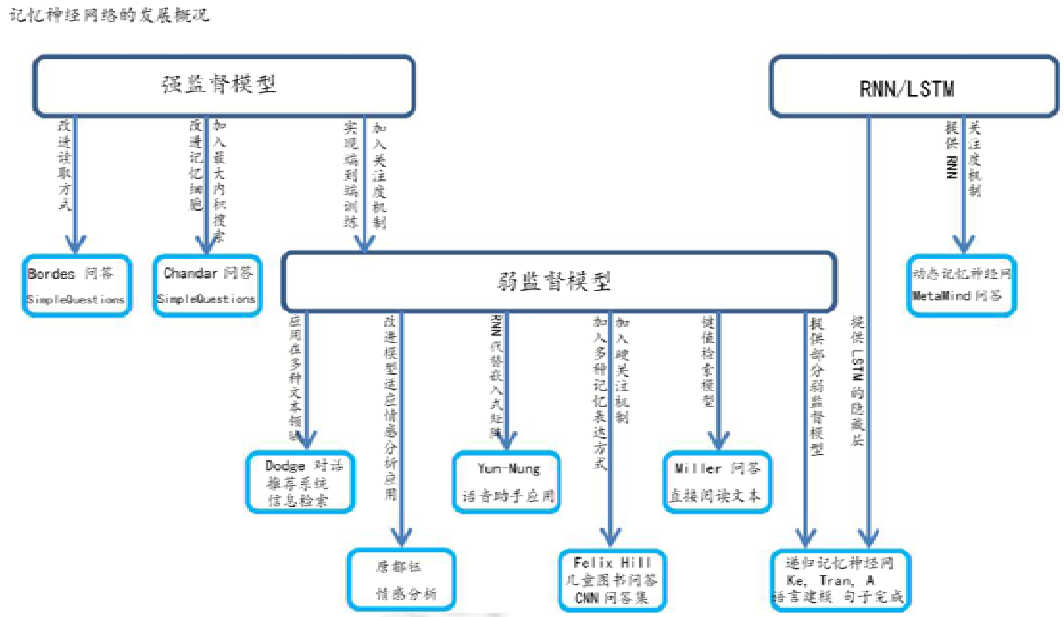


Fig.1 Memory networks development overview

图 1 记忆神经网络发展概况

2 记忆神经网络原理

MemNN 利用精确记忆的机制来解决机器阅读理解问题.首先,向神经网络输入一段文本和一个问题,它们被拆分成句子序列;然后转化为一种知识表示形式,再存储到数组之中;然后,依靠神经网络的方式反复读取,训练输出答案文本,直到神经网络最终收敛,存储在数组中的信息可以被更新或扩展.

强监督模型和弱监督模型都遵循着两个基本步骤:第 1 步,通过存储的记忆和问题,使用机器学习的方式找到包含支撑因子的句子;第 2 步,通过前步的支撑因子和问题,使用机器学习的方式找到最终答案,如图 2 所示.

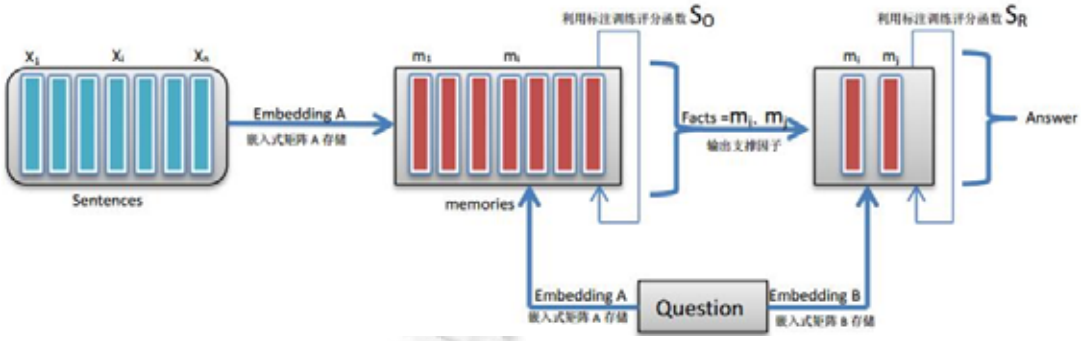


Fig.2 Strongly supervised mode training

图 2 强监督训练方式

记忆神经网络的实现过程是一个描述、预测、再加以引导的过程.Facebook 人工智能实验室将自然语言中的一部分问题描述成 20 类典型问题[25,29],每类问题子集中的单个问题都被描述为 3 部分.

(1) 故事(story/sentences):陈述段落,包含多个句子,每个句子有索引序号.

(2) 问题(question):一个问句包含一定的指向性信息,不使用代词.

(3) 答案(answer):答案以及与在故事中与答案相关的句子索引序号.

文献[17]以这种抽象式的表示方法生动地描述了问答模型的结构.将这样的数据集合输入到记忆神经网络的模型进行训练,最后达到预测答案的目标.然而,这样的预测过程需要利用数据集合内在的抽象关系来对模型进行引导训练,才能使其预测答案.

模型通过输入构建记忆细胞,应用训练数据对模型中的两个评分函数进行训练,得到两个评分函数:第 1 个函数用于判断文本中哪个句子与问题有相关性;第 2 个函数用于判断候选答案与支撑因子和问题的相关性.实际应用时,通过对这两个评分函数的评价,最终可以得到推理后的答案.

2.1 递归神经网络的局限性

得益于 RNN 的记忆能力,LSTM 在很多数据集上都有很好的表现.LSTM 将表示知识压缩成为密集的向量,通过隐藏层和权重来实现记忆功能,但记忆能力还是有限,并不能精确记录数据集原始信息.Karpathy 等人通过实验证实 LSTM 是通过概率的方式记录高纬向量[35],而非精确记忆.LSTM 完全不依赖于已经有的知识,直接通过模型的拟合能力得到答案的概率分布.但是,当优化策略需要长时间跨度规划时,LSTM 系统与人的表现相差甚远.例如,Hausknecht 等人利用 LSTM 玩《太空侵略者》游戏时,模型的长程记忆能力就比较差[3].不同于 LSTM 通过隐含层保存记忆信息,MemNN 是通过记忆模块将信息完整地保存下来,它与 LSTM 相比具有如下优点.

(1) 在自然语言处理上,特别在问答类处理上,能够适应多种问答数据集,并优于 LSTM[30].

(2) 完整保留数据集信息,在提取文本特征的准确度上显著强于 LSTM.

(3) MemNN 轻易地实现迁移学习,在存储空间足够大的情况下能够实现不间断的学习.

(4) MemNN 的开发为神经网络的记忆可读写性的设计开辟了道路.

2.2 记忆神经网络的基础模型

记忆神经网络的基础模型有强监督学习和弱监督学习两种,两者的主要区别在于训练方式上:强监督模型的训练模块要多于弱监督模型,模型整体复杂度也高于后者;弱监督模型是强监督模型加入关注度机制的改进模型,具有处处可微的特性,对数据集标注的要求低于强监督模型.

2.2.1 强监督学习

强监督模型有两个主要训练模块,并且都需要训练,第 1 模块的训练的输出要作为第 2 模块的输入,需要两次端到端的训练,如图 3 所示,强监督模型共包含 4 个训练模块和 1 个存储器.

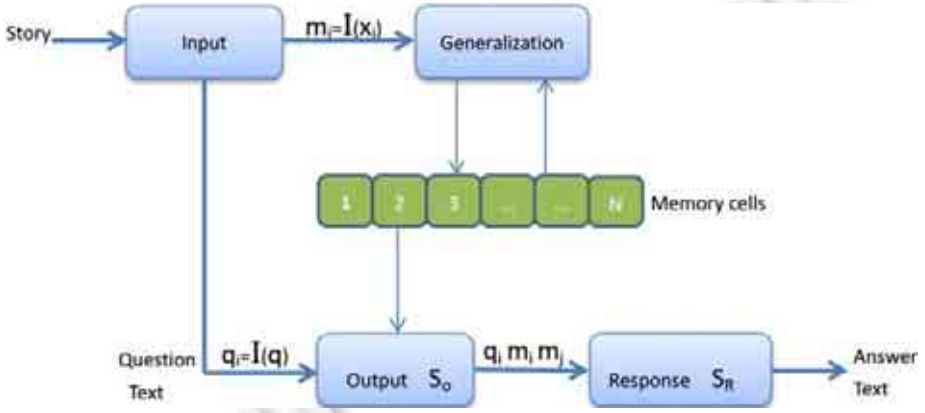


Fig.3 Strongly supervised model

图 3 强监督模型

文献[17]的强监督模型结构中,*I* 模块接受输入文本后将其转化为语言模型,一般使用词袋模型,也可以是词

向量模型.词向量模型所表达的内部含义要远比词袋丰富,但是词向量模型会随着训练文件的增加发生变化,导致存储向量过期,影响输出结果.如果存储向量能够保持不变,则词向量模型在读取速度上有优势.读取文本之后,将文本用词袋向量的形式逐个保存在记忆细胞中,简单的记忆神经网络中 G 模块仅仅将输入的信息逐个存储到空置的记忆细胞中,后续的扩展模型中将会有各类存储和检索算法应用其中,数据流程如图 3 所示.

图 3 的 4 个模块都是训练模块,性能都是通过训练或设计后获得,其中,

- Input,I 模块:将故事、问题、答案这个 3 个主要的输入转化成为特征表示,通常表示为词袋模型向量或词向量,为 G 模块存储做准备,x_i 为文本句子,m_i 为转化后向量.
- Generalization,G 模块:根据输入确定是否更新当前存储器状态,先将 I 模块中得到的向量 m_i 存入到存储器数组中,Bordes 将这些存储数组被称为记忆细胞[30].G 模块决定了长期记忆的存储方式,存储方式直接影响对记忆细胞的读取和写入效率.
- Output,O 模块:O 模块同时读取代表问题的向量 $q_i=I(q)$和记忆细胞中的向量,利用评分函数 S_O 找到并输出包含支撑因子的记忆细胞向量 m_i,m_j.
- Response,R 模块:R 模块称为反应模块,R 模块将 O 模块的输出(m_i,m_j)作为输入,同时读入代表问题的向量 $I(q)$,通过评分函数 S_R 计算得到最高分单词,输出答案 Answer Text.

强监督的 bAbI 训练集样例见表 1.其中,1~6 是故事文本;7 的第 1 句话是问句,问号后是答案,3 和 6 是与答案相关的句子.

Table 1 An example of bAbI

表 **1** bAbI 数据集样例

分类	训练集	与答案相关性
故事	1. Mary moved to the bathroom.	×
	2. Sandra journeyed to the bedroom.	×
	3. Mary got the football there.	√
	4. John went to the kitchen.	×
	5. Mary went back to the kitchen.	×
	6. Mary went back to the garden.	√
问题	7. Where is the football? garden 3 6	–

O 模块与 R 模块是强监督模型的推理核心,训练和推理工作都在这两个模块中完成,这两个模块的核心是评分函数 S_O,S_R,并且有相同的函数形式,如公式(1)所示

$$S(x,y)=\Phi_x(x)^T U^T U \Phi_y(y) \tag{1}$$

其中,$\Phi_x(x)$和$\Phi_y(y)$分别代表输入文本的向量表示(通常是词袋模型),维度都是 $n\times 1$;U 是一个 $n\times D$ 维度的权重矩阵,n 是向量维数且是一个超参数,D 是嵌入式矩阵维度,通常选择 $D=3|W|$,W 代表词袋向量的长度,它是由 S_O,S_R 函数通过矩阵运算后得到实数值.

用词袋模型时,两个向量的内积可以表示相关度,比如$\Phi_x(x)^T\Phi_y(y)$越大,代表相关性越高,但这是一种线性关系.在实际应用中,很多相关语句之间的关系是非线性或间接的,因此需要通过机器学习的方式实现这种非线性关联或者间接的关联.文献[17]中引入权重矩阵 U_O 来解决这个问题,首先将记忆细胞中的全文本$\{m_1,m_2,\ldots,m_n\}$和问题 $I(q)$作为输入,以含有支撑因子的记忆细胞 m_i,m_j 为目标输出,采用随机梯度下降法(stochastic gradient descent,简称 SGD)反复对矩阵 U_O 进行训练,得到评分函数 S_O,如公式(1)所示.$S_O(x,y)$最终以实数表示得分,得分的范围可以判断相关程度的高低,还设置一个参数 K 来确定求几个相关的记忆细胞.

当设置 $K=1$ 时,即只求相关度最高的 m_i,函数如公式(2)所示.

$$o_1 = O_1(x,m) = \arg\max_{i=1,2,\ldots,N} S_O(x,m_i) \tag{2}$$

当 $K=2$ 时,求与 $I(q)$和上式 m_{o1} 这两项都最相关的 m_{o2},函数如公式(3)所示.

$$o_2 = O_2(x,m) = \arg\max_{i=1,2,\ldots,N} S_O([x,m_{o1}],m_i) \tag{3}$$

最终的输出列表为$[q_i,m_{o1},m_{o2}]$,这个列表将作为 R 模块的输入.

R模块的参数矩阵U_R使用与U_O类似的训练方式,以O模块的输出作为输入,以最终答案Answer作为输出,进行端到端的训练,其函数形式如公式(4)所示.

$$r=\text{argmax}_{w\in W}S_R([x,m_{o1},m_{o2}],w) \tag{4}$$

在文献[17]中,Weston 等人对两个模块的训练采用边缘排序误差和随机梯度下降法,通过逐步缩小损失函数的方式训练U_O和U_R矩阵.一个bAbI数据子集至少要反复迭代训练2 500次.

在模型的运用中,不可避免地要涉及到时间序列问题,如果是普通的问题,比如中国的首都在哪里,可以直接给出答案.但是有时间顺序的问题就不容易回答了,要考虑句子所在的时间顺序.因此,对这类问题的数据需要额外的文本来标记时间顺序,为每个句向量的前 3 位添加标记,再设立$\Phi(x,y,y')$函数,训练后,函数可以按照 y与y'的时间顺序选择最合适的y.

此模型在bAbI数据集中极少遇见未知词,此模型应用在其他数据集中时要考虑到未知词对模型的影响,采用的方法有两种:(1) 通过考虑临近的词,猜测这个未见单词的含义;(2) 每个单词不再单独表示,使用 3 个单词为一组的形式表示,即任何一个单词的左、右单词都与其成为一组.单词维度从$|W|$变为词组维度$3|W|$,任何未知词都可用这个方式表示.同时,为了适应这种表示方式,嵌入式矩阵的 D 维度从 $3|W|$增加到 $5|W|$.在训练模型时,采用类似于“dropout”的方式进行训练.在某些时间内,假装从没有见过某个词,而完全使用其左右的相关词代替.中国科学院的研究人员专门对答案是未知词的情况做了研究[36],文献[36]设计了新的层级存储结构和最大k池(k-max pooling)的寻址方式,在4个包含有大量未知词的数据集中进行了测试,取得了很好的实验效果.

2.2.2 弱监督学习和关注度机制

尽管强监督模型[17]以两段监督的方式在bAbI上取得了良好的表现,但是现实情况中,很多问答数据集只有问题与答案这样的数据对,难以提供关于记忆的标签.针对这一问题,Sukhbaatar 等人又提出了一种端到端的弱监督模型[26,37],显著减少了训练时间和步骤,应用范围更广泛.当前,人工智能领域有两个重要问题:一是如何解决连续数据上的长距离依赖问题;二是如何在自然语言理解中建立多级计算的步骤,即,将上下文关联的信息在计算中联系起来.

图4(a)和图4(b)是 Sukhbaatar 在文献[26]中提出的弱监督模型基本结构和多跳结构,其基本训练和应用方法如下:模型应用的数据集以一组离散的数据形式$\{x_1,x_2,\ldots,x_n\}$输入到模型之中,这组数据中包含故事文本 x_i,一个问题q,一个答案ans;模型将所有的x都写入到一个可读写的缓冲池中,然后将问题q和答案ans也表示为同类型的向量;将x_i和q定义为输入向量,已经标识的ans定义为目标向量,设置运算层级(hops).弱监督模型支持多层级运算,为多次检索故事文本x_i与问题q的相关性提供信息.通过backpropagation误差反向传播的训练方法,在多个训练层级中反复地训练,最终用x_i和q通过模型的计算输出答案ans.

- 输入方式

首先,将文本数据集存储到模型的记忆细胞中.如,将故事全文本$\{x_1,x_2,\ldots,x_n\}$(维度为 $V\times1$)逐个与嵌入式矩阵A(维度为$d\times V$)做矩阵内积运算后,将数据集x_i转为向量m_i(维度为$d\times1$)存入到存储细胞中;同时,将问题q与另外一个嵌入式矩阵B(维度为$d\times V$)做矩阵运算后转化为向量u(维度为$d\times1$),用向量u与图4(a)中Input模块下的每个句子向量m_i做内积,再用Softmax函数归一化后得到一组权重,这就是关注度机制.通过此机制,可以清晰地知道问题q与每个句子x_i的相关度,如公式(5)所示.

$$p_i=Softmax(u^Tm_i) \tag{5}$$

其中,归一化公式: $Softmax(z_i)=e_i^z\Big/\sum_j e_j^z$.

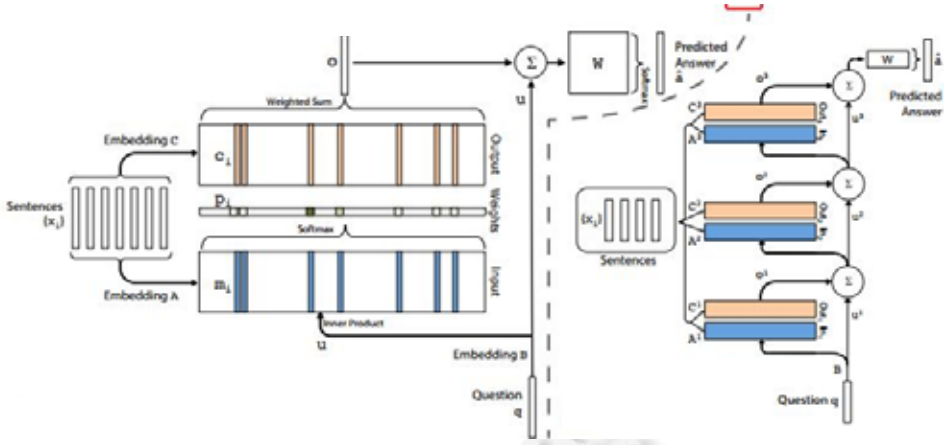
- 输出方式

文本$\{x_1,x_2,\ldots,x_n\}$同时与嵌入式矩阵C(维度为$d\times V$)做矩阵内积运算,存入到另一层记忆细胞中表示为c_i,将权重p_i与c_i做加权平均得到输出O,O即代表支撑因子,如公式(6)所示.

$$o=\sum_i p_ic_i \tag{6}$$

以上公式使用的权重归一化方式正是关注度机制(attention mechanism).在文献[38,39]中,有关注度机制在

文本研究的详细介绍.文献[26]中,关注度机制代表一个权重,越大的权重代表对应位置的 m_i 越重要.



(a) 单层弱监督模型 (b) 多跳弱监督模型

Fig.4

图 4

关注度机制在神经网络领域有很长的历史,早期被应用在图像识别领域[40,41],而比较有影响力的是 Google DeepMind 团队的文献[42].他们在 RNN 模型上使用关注度机制进行图像分类,通过关注度机制去学习图像要处理的部分,每次运算时,都会根据前一个状态学习到需要关注的位置和需要注意的部分像素,而非全部图像.这样做的目的是减少任务复杂度,更贴近人类的注意力机制.特别是在文献[43]中,Gregor 等人将关注机制用于生成图片的描述,使用 CNN 编码测试图片,用关注度机制和 LSTM 来对关注度的权重值可视化,从而在生成词语的同时解释模型在图片上的关注位置.Moritz 等人也是利用关注机制可视化的方式[44],通过先读入一个文本,再生成一个答案的训练,实现观测神经网络模型在寻找答案时关注了文章的哪些方面.

近年来,有名的关注度机制应用是神经网络的机器翻译[45],在 2014 年的神经网络翻译文献[15]中,Google 团队还在使用 RNN 的编码-解码(encode to decoder)模型,文献[45]在新一代的神经网络翻译模型上加入了关注度机制,通过关注度机制,把源语言端的每个词学到的表达和被预测的词联系起来,并将模型的数据加权组合用 Softmax 函数得到一个概率分布,如公式(5)所示,这样就可以表示源语言和目标语言是怎样对齐的.通过关注度机制的使用,机器翻译取得了令人振奋的结果[45].

- 预测答案

C 模型层中,最终的输出是 O,它代表了概率统计上故事文本 x_i 和问题 q 间最有相关性的句子向量;利用 O 向量和问题向量 u 作为输入,答案 a 向量作为输出可以再构建一个神经网络模型,其权值矩阵为 W;最后一层以 Softmax 为输出,如公式(7)所示.

$$a'=Softmax(W(o+u)) \tag{7}$$

a'作为输出值,与目标值 a 之间的误差可以用来训练整个模型.纵观整个模型,可以看到数据的流通非常平滑,并且是全程可微的,此模型是可以用误差反向传播进行训练,最终优化 A,B,C,W 这 4 个权值矩阵,Sukhbaatar 在文献[26]中使用交叉熵损失函数和随机梯度法对 4 个权值矩阵进行训练.

实验结果表明,单层模型不足以找到所有包含支撑因子的相关句子[26].因此,弱监督模型引入多跳计算的方式对故事全文本 $\{x_i\}$ 进行反复的检索.多跳记忆的方式是将前一层的输出作为后一层的输入,目的是不只寻找与 u 相关的句子,也寻找与已知支撑因子 m_i 相关的句子.这种模型的组织方式与 RNN 非常相似,因此也可以将弱监督模型看成是 RNN 模型的一种.如图 4(b)所示,将 C 层的输出 o 和问题向量 u 作为下一层的输入,如公式(8)所示.

$$u^{k+1}=u^k+o^k \tag{8}$$

最终输出如公式(9)所示.

$$a'=Softmax(Wu^{k+1})=Softmax(W(o^k+u^k)) \tag{9}$$

多跳计算的方式下,每一层都有其自己的嵌入式矩阵,如 A^k,C^k 用来将输入 x_i 转化为待存储的向量 m_i.

上式中,A 和 C 在不同层级别下有不同的初始化方式,最开始,A,B,C,W 都是采用随机初始化的方式,在新增加的计算层中,文献[26]采用两种方式优化模型,减少训练参数.

- 邻接共享矩阵:将上下两层用来转化输入数据,并将存储的矩阵参数共享,前一层的输出权重矩阵 C 就是下一层的输入权重矩阵 A,比如 $A^{k+1}=C^k$.不仅是这里共享,在最终输出答案的 W 矩阵也和 C 矩阵进行共享,$W=C^K$.
- 矩阵逐层共享:一种类似于 RNN 权重共享的模式:$A^1=A^2=\ldots=A^K$ 和 $C^1=C^2=\ldots=C^K$.其应用方式与单层没有区别,但是却有更好的信息提取能力.

弱监督模型在句子表达上尝试使用了位置编码(position encoding)方式,在实验中比词袋模型表现得更好.

- 词袋模型与 A 的运算方式如公式(10)所示.

$$m_i=\sum_j Ax_{ij} \tag{10}$$

- 位置编码与 A 的运算方式如公式(11)所示.

$$m_i=\sum_j l_j\cdot Ax_{ij} \tag{11}$$

- l_j 的计算方式如公式(12)所示.

$$l_{kj}=(1-j/J)-(k/d)(1-2j/J) \tag{12}$$

J 代表句子中的单词数目,d 代表嵌入矩阵维度,k 代表多跳计算的次数.位置编码是一种更适应多跳计算的表达方式,在后文的情感分析文献[27]中也被使用.图 4(b)设计了多跳计算,它可以在存储记忆中反复检索信息,将上下文关联信息联系起来,因此在长程记忆中至关重要.

弱监督模型尝试了语言建模实验[26],即从序列单词中预测下一个单词,分别在 Penn Treebank[46]和 Text8 corpora[47]这两个数据集上进行了实验.文献[26]的实验结果显示:增加多跳记忆有助于模型表现的提高,平均表现也超过 LSTM[26].

2.2.3 强监督模型与弱监督模型比较

强监督模型和弱监督模型都曾经应用在 bAbI 数据集上,并取得了不错的效果.强监督模型的平均错误率是 6.7%,弱监督模型使用位置编码表达句子向量和 3 跳的计算方式,取得的最好平均错误率是 12.4%.在准确率上的差异源于二者模型结构的不同以及训练阶段的要求不同.

- 首先,弱监督模型相对于强监督模型,少一个训练步骤,即强监督模型在训练中使用了记忆标签,能够更准确地把握哪些是包含支撑因子的记忆向量,而弱监督模型则仅依靠关注度机制把握含有支撑因子的记忆向量.
- 其次,强监督模型的两个模块分别使用各自的损失函数进行训练,而弱监督模型通过连续模型处处可微的优势全程使用一个损失函数,用 backpropagation 的方式对数据进行端到端的训练,所以比较而言,强监督模型训练更加充分.
- 最后,弱监督模型提出的新神经网络结构能够将计算和存储深度耦合,在输出最终结果之前用多跳的方式循环读取存储数据.这种模型既可以看成 RNN 的一种新模型,也可以看成强监督模型的连续形式.

现实情况中,很多自然语言训练数据只有输入输出对,难以提供更多的监督信息.这有利于弱监督模型体现其端到端的训练优势,因此,弱监督模型更加适用于多种类型的任务,本文第 3 节详细介绍了弱监督模型在多种类型数据集上的应用.但是在 bAbI 数据集上可以看出,其准确率低于强监督模型,因为弱监督模型没有充分利用记忆标签.在某些数据集上,如 SimpleQuestions 等知识库类数据集,因为知识库三元组的表示方式使得数据集自身带有记忆标签,所以强监督模型更适用于这类情况,不但可以取得较高的准确率,还能扩展出迁移学习的能力.综上所述,记忆神经网络应用在自然语言数据集时需要依据数据集的具体情况来选择合适的神经网络模型,才有可能得到较好的效果.

3 记忆神经网络应用现状

3.1 强监督模型的发展

自然语言理解中,问答是比较重要的方向.文献[17]中使用的 bAbI 问答数据集是为了验证强监督模型而构建的合成数据集,合成数据的优点是,用简单模型构建的大数据胜过用真实数据构建的小数据集.发展合成数据集的好处是可以依托仿真的数据集开发模型,并最终应用在真实数据集上.Weston 等人在文献[29]中详细讲述了 bAbI 数据集的构建过程和 20 个子集合的问答种类,并进一步改进了强监督模型.bAbI 数据集的目的并不是替代真实数据,而是用来评估模型的表现,使模型能够渐渐地应对复杂模型.

3.1.1 读取方式的改进

Weston 等人在 2015 年的文献[29]中研究如何改进模型的读取方式来获得性能提升,实验结果表明,改进模型[29]比旧模型[17]在 bAbI 的子集任务上有更好的表现.新模型的主要改进之处如下.

(1) 原始模型不支持 3 个支撑因子以上的回答,在新实验[29]中设置了自适应记忆寻找机制.具体方式是:在检索时,只有检索到正确的答案才停止在记忆细胞中寻找.

(2) 多单词答案.文献[17]模型中,所有的回答都是一个单词的回答,改进方式是:在答案预测模块 R 的字典中增加一个预测词 W_C,只有预测模块成功输出 W_C 时才停止在预测模块中查找.

(3) 改变句子的向量表示方式.文献[17]中使用词袋模型表示,在改进的方法中,使用基于词袋的 N 元模型(a bag-of-N-grams)表示[29],这种非线性的句子表示方式使得 MemNN 在读取上有更准确的表现.

新模型[29]至少在 5 个 bAbI 的数据子集上有超过最初文献[17]的表现.

Bordes 则尝试将强监督模型应用于大数据的文本理解[30],为此创建了一个大数据集 SimpleQuestions[30]包含了 10 万个问题对,在这个数据集上,有效地验证了强监督模型具有复杂推理能力,开发并测试了强监督模型的迁移学习能力.SimpleQuestions 数据集是基于 Google 知识库 Freebase 而构建,它将 Freebase 进行简单的预处理后构建成问答对,用来回答一些简单的事实,甚至是一个回答列表,形式见表 2.

Table 2 An example of SimpleQuestions

表 2 SimpleQuestions 样例

问题	知识表示三元组
What American cartoonist is the creator of Andy Lippincott?	(andy_lippincott,character_created_by,garry_trudeau)
Which forest is Fires Creek in?	(fires_creek,containedby,Nantahala_national_forest)
What is an active ingredient in childrens earache relief ?	(childrens_earache_relief,active_ingredients,capsicum)
What does Jimmy Neutron do?	(jimmy_neutron,fictional_character_occupation,inventor)
What dietary restriction is incompatible with kimchi?	(kimchi,incompatible_with_dietary_restrictions,veganism)

注:第 1 列是问题;第 2 列代表三元关系,既是读入的故事文本又是支撑因子,下划线部分代表答案

妥善组织基于 Freebase 的知识库,形成足够大的问答数据集就可以在实际应用中涵盖很大的知识范围.在 SimpleQuestions 数据集之前,常用的数据集是 WebQuestions[48],强监督模型在 WebQuestions 上的测试也取得了较高的评分,相对于 Yang 等人在实验[49]中得到的最高分 41.3%,强监督模型达到了 42.2%的高分,这进一步说明了强监督模型在问答问题上的有效性.但是由于 WebQuestions 是个较小的数据集,并不能证明强监督模型在足够大的数据集上依然有较好的效果.而且 WebQuestions 数据集过小,也不能应用在迁移学习上.基于 Freebase 构建的 SimpleQuestions 是较大的数据集,则拥有足够多的问答对.Bordes 等人[30]在 3 个数据集 FB2M,FB5M,Reverb 上测试了强监督模型的大范围问答能力和迁移学习能力.具体做法是:先用 FB2M 和 FB5M 训练模型,训练之后保存权重,同时,也将这两个数据集的信息保存在记忆体之中.迁移学习分为两部分实现:第 1 步,由于 FB2M 和 FB5M 的支撑因子涵盖了很多 Reverb 的信息,可以直接回答部分 Reverb 数据集中的问题;第 2 步,使用强监督中的 I 模块将 Reverb 中的包含支撑因子的信息进行预处理,然后通过 G 模块将信息链接到已经存在的独立记忆细胞 m_i 之下,采用的文本相似性的方式与已经存储的信息进行链接.此处的关键点在于构建评分函数,如何运用评分函数是迁移学习的关键.文献[30]在训练 FB2M 和 FB5M 使用的评分函数是余弦近似函数,其

公式(13)为

$$S_{OA}(q,y)=\cos(W_V g(q),W_S f(y)) \tag{13}$$

其中,q 代表自然语言问句,y 代表记忆.文献[30]使用 n-gram 词袋模型方法,用 N_V 维度的多热(multi-hot)向量 $g(q)$ 来表达问句 q.N_V 代表字典大小,字典包含所有 Freebase 知识库实体和所有问题中出现的单词,N_V 也作为 W_V 的维度之一.同理,使用词袋模型(bag-of-symbol)的方法,用 N_S 维度的多热向量 $f(y)$来表示每一条知识并作为记忆.N_S 为知识库实体和实体关系大小之和,N_S 也作为 W_S 的维度之一.文献[30]将问题和记忆共同投影到一个低维嵌入式空间,通过余弦相似度作为得分函数来寻找记忆的支撑因子,即公式(13).这里需要训练的就是两个权值矩阵 W_V 和 W_S,这种多任务的训练与文献[17]相似.

当迁移到 Reverb 数据集时,使用评分函数,如公式(14)所示.

$$S_{RVB}(q,y)=\cos(W_V g(q),W_{VS} h(y)) \tag{14}$$

文献[30]通过实体链接和实体别名匹配等方式来匹配已有记忆中的实体和新知识库 Reverb 里的实体,新知识库中剩下的实体和所有的关系都用词袋模型表示,因此可以用一个 N_V+N_S 维的向量 $h(y)$来表示新知识并将其存储到记忆中,其中,矩阵 W_{VS} 直接由之前训练好的 W_V 和 W_S 拼接而成.Bordes 等人[30]的实验结果表明,强监督模型具有增强记忆功能,可以快速吸收新数据,并继续训练,而不是放弃原有训练模型而重新开始训练.

通过已知文本和问题得到问题的支撑因子,是强监督模型的核心,有两个重要方面影响这个核心:第一,从文本和问题得到支撑因子的神经网络函数,又称为记忆细胞检索过程;第二,如何存储并有效检索记忆细胞中的信息,良好的检索方式和存储结构直接影响训练速度和训练效果.Bordes 等人的实验[30]结果表明,强监督模型在大数据文本理解上具有潜力,但是随着记忆细胞的增多,强监督模型的寻址时间也线性增大,导致整体训练时间增加,特别是在 SimpleQuestions 这样的包含了 10 万对信息的大数据集上,训练时间很长.

3.1.2 记忆细胞的改进

针对大数据集上训练时间增加的情况,Chandar 等人提出了一种优化的强监督模型——分层记忆网络(hierarchical memory network,简称 HMN)[31],引入最大内积搜索(K-MIPS)来训练模型,并在 SimpleQuestions 数据集上尝试了新的检索和记忆存储结构,提升了强监督模型在大数据集上的扩展能力.分层记忆网络特点如下:

(1) 开发了层次结构的记忆细胞,其中有簇结构、哈希结构和 PCA 树结构;
(2) 将记忆细胞分解为不同的结构子集,用最大内积检索(MIPS)方式检索记忆细胞,以此训练读取模块;
(3) 实验结果表明,将精确的 MIPS 算法作为一种软关注模型时,整体的计算代价非常巨大;而采用近似的 MIPS 算法,可以在速度提升和扩展性上取得一定的平衡,代价是牺牲一定的准确性.

在文献[17,30]的模型中,记忆细胞是以多维数组的方式保存故事文本,在 HMN 中,采用层级方式保存文本信息,信息按照一定的规则分成各种子集,然后,与读取记忆体的 O 模块的检索方式进行耦合,加快访问速度.O 模块的读取方式有最基本的逐个读取方式,还有软关注机制的方式[26,33]和硬关注方式[32],这两种关注机制在小数据集上有优势,随着数据集的增大,性能都会下降.

针对大型训练集,使得读取模块具有最大内积读取能力是有效的训练方式,如 K-MIPS 训练方法直接寻找最接近的子集[31],K 代表子集中优先选中的 K 个条目.给定一系列句子 $\{x_1,x_2,\ldots,x_n\}$和一个问题 q,找到最大内积的前 K 个条目的 K-MIPS,如公式(15)所示.

$$\text{argmax}^{(K)}_{i\in\chi} q^T x_i \tag{15}$$

除了最大内积训练方式以外,还可以采用近似方式,如最大余弦值(MCSS)等方式,当输入数据向量 x_i 都拥有类似的形式时,最大余弦等同于 MIPS.以上训练方式的优点是:当模型使用 $Softmax(K)$时,自然就是将关注点集中在几个记忆细胞之上,这些受到重点关注的记忆细胞能够以较强的梯度进行学习.在大数据训练集上,记忆细胞的贡献的梯度非常小,文献[30]的实验结果表明,这会减缓训练速度.使用近似 K-MIPS 的方式可以在加速算法的同时不损失效率,这也是文献[31]的主要贡献.

3.2 弱监督模型的发展

弱监督模型得益于其端到端的训练方式,非常容易应用在现有的数据集上,比如新闻分类、问答、情感分

析、影视评论、对话、推荐系统等等.多跳的计算方式更是让其在数据集的检索上有更好的表现.

3.2.1 多领域的应用

Dodge 等人尝试将弱监督模型应用在多种文本领域[28],如端到端的对话数据集、推荐系统数据集、问答和推荐数据集、Reddit[50]的回复数据集和联合训练集.他们的研究将当前主流问答类模型进行对比实验,比如问答系统[48,51]、SVD[52,53]、信息检索模型[54−57],结果显示,弱监督模型在对话数据集、问答数据集、推荐数据集上都有超越其他模型的良好表现.可见,采用存储记忆方式的神经网络模型在特征提取和自然语言理解上达到了一个新高度.

弱监督模型在情感分析领域[27]也得到了应用.哈尔滨工业大学唐都钰博士将情感分析作为一个文本分类问题看待,用机器学习的方法训练文本分类器[27].分类器的性能通常依赖于情感词典和文本特征等信息,常用的模型如递归神经网络和长短记忆神经网络,在多篇文献中[58−61]都增加了在情感分析方面的关注.传统的递归神经网络模型可以应用在情感分析中[58,61],并且擅长捕捉背景信息,但是训练时间过长,且不能明确地区分哪个词对整个句子情感有重要影响.真实情况下的情感分析中,只有一部分信息对整个句子的情感倾向有影响,其中,句子中的某些实词和动词对情感倾向的影响是明显低于某些形容词和副词.而加入了关注度机制的弱监督模型却很容易在这类问题上有很好的表现[27].

如将一个句子词向量分成两部分看待:第 1 部分关键情感词向量,可以是 1 个词也可以是多个词的平均值,用词向量表示;第 2 部分上下文词向量,即在句子中除了关键情感词之外的所有词.那么,分析一个句子的情感倾向问题围绕着这两部分进行.这两部分之间的相互影响决定了句子的情感方向.在应用的数据集 SemEval 2014[62]上,可以将这两部分作为目标输入,将已经人工标识好的情感倾向作为目标输出,以这对数据应用于模型进行训练.为了研究不同词向量的影响,文献[62]分别为 4 种不同的词向量表达做了实验.

为了适应情感分析,模型对弱监督模型进行改进,原模型的故事文本替换为上下文单词(context word),原模型的问题替换为关键情感词(aspect word).文献[27]中使用了多跳计算,包含多个计算层,用来反复提取记忆细胞中的信息.每个计算层包含一个关注度层和一个线性层,这两个层的输入都是关键情感词向量(aspect vector),将关注度层和线性层的输出结果求和作为下一个计算层的输入,最后一层将包含了关键情感词(aspect word)的信息作为情感分类的特征送入到 Softmax 中分类.文献[27]的情感分析模型如图 5 所示.

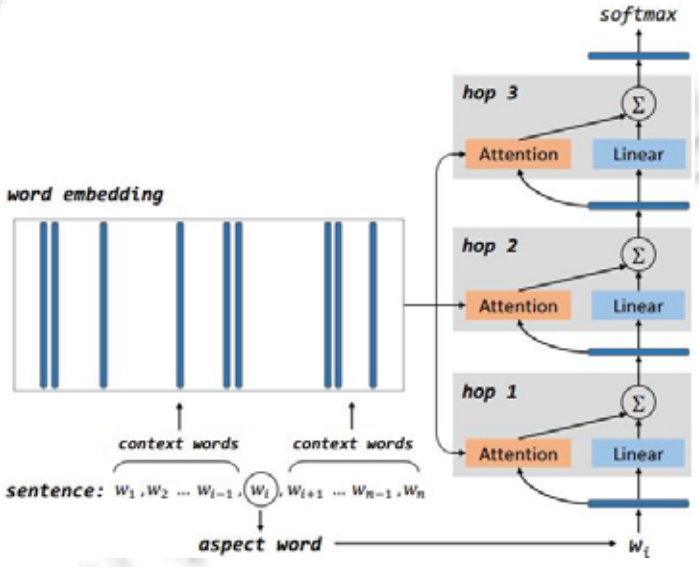


Fig.5 A deep memory neural network with three computational layers (hops) for aspect level sentiment classification

图 5 3 层(跳)深度记忆神经网络计算情感倾向分类

图 5 中,Context words 代表上下文词向量,Aspect word 代表关键情感词向量.这两个向量代表模型的输入.Context words 词向量被表示为记忆细胞 m_i,Aspect word 词向量被表示为 w_i,并使用了多跳计算的方式.

文献[27]中还提出了关注度机制的不同用法,分别从内容和位置两个方面设计了关注度机制:第一,内容关注度机制,通过训练模型可以自动地对每个上下文词的情感贡献进行评分,从此体现出上下文词与整体情感的相关性;第二,内容关注度机制忽视了关键情感词与上下文词位置上的关系,通常情况下,与关键情感词位置较近的上下文词对相应的情感倾向判断更重要,所以文献[27]又设计了位置关注度机制.实验结果显示,这两种关注度模型在单跳计算模式下都有超越 LSTM 的表现,在准确率提升的情况下,计算时间得到了很大的缩短.在单跳计算模式下,4 种位置关注度机制的准确率最低为 72%,而内容关注度机制则为 76%,略高于全部位置关注度机制.从 2 跳提升到 8 跳的过程中,几种模型准确率很接近;在 9 跳时,所有模型都达到了 81%的准确率.

3.2.2 编码与解码的模型

弱监督模型也可以用于语音助手的工作[34],常见的语音助手有 Microsoft's Cortana 和 Apple's Siri.这两种助手都可以将口音指令翻译成机器可理解的操作,如使用者表达:(just send email to bob about fishing this weekend(1)).转换成为执行语意则是,发送邮件(联系人="bob",主题="fishing this weekend"(2)).通常,针对这样的语意转换使用的是两种 RNN 模型:第一,直接使用一个 RNN 单层或多层进行端到端的训练;第二,使用编码-解码(encode-decode)的方式,先将表达文本(1)编码(encode)成为一个固定长度向量 f,然后再将向量 f 解码(decode)为目标语句(2).所用的方式都是通过 RNN 记住一些转换规则,而弱监督模型经过适当的修改也可以处理此类任务.弱监督模型通常使用一个嵌入式矩阵 A,将文本 S_i 逐句地转换成向量 m_i,$m_i=A\Phi(S_i)$.在这个应用中,将嵌入式矩阵 A 都替换成 RNN,将输入文本 S_i 编码,将最后的输出 Softmax 替换成 RNN,用来解码.文献[34]的模型结构如图 6 所示.

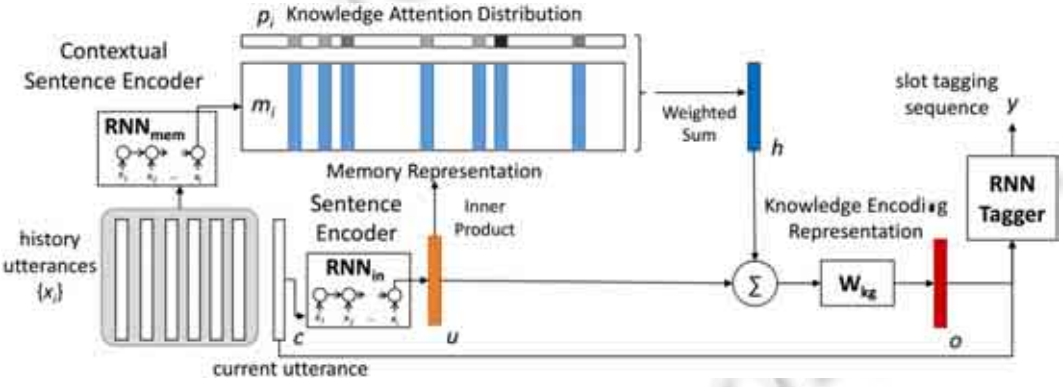


Fig.6 Illustration of the proposed end-to-end memory neural network model for multi-turn SLU

图 6 端到端记忆神经网络的多级语音理解模型说明

文献[34]用全部的口语指令作为文本,而用当前输入作为问题,Microsoft's Cortana 和 Apple's Siri 的输出作为答案进行训练.在文献[34]的实验中可以比较两种 RNN 方式和弱监督模型,实验结果显示,弱监督模型保持了较高的准确率.而在内容理解的精度值和 $F1$ 值上,文献[34]的模型略高于弱监督模型,体现了文献[34]模型的可用性和鲁棒性更好.由此可见,长程记忆在语言模型上的优势非常明显.针对不同的应用而适当改变,弱监督模型也可以得到非常好的表现.

3.2.3 软关注机制和硬关注机制

RNN 有较强的语言建模能力,优势主要体现在动词和介词上,而对实体和实物的表现很差,而实体和实物常常是回答的核心.Hill 等人尝试将弱监督模型和改进的记忆神经网络模型应用到更复杂的问答领域中,如儿童读物(CBT)和 CNN 问答集[32].在儿童问答读物测试中,与测试人员的比较已经达到了很高的水平,例如实体、实物、动词、介词等几类词的预测回答上,测试人员的平均水平是 81%左右,而 Hill 等人的改进模型平均准确率

在 66.6%左右,此结果已经超越过去全部的问答模型,包括曾经在语言建模领域有较好表现的 LSTM[32].

CBT(children's book test)语料库数据来自于 Project Gutenberg,为保证故事叙述的连续性,上下文作用更加突出.每篇文章只选用 21 句话,前 20 句作为故事文本,用 S 表示,S_i 代表每一行,将其中第 21 句去掉一个单词后作为问题 q 表示,而被去掉的单词作为答案,用 a 表示,并且给定 10 个候选答案,用 C 表示,每个候选答案从原文中随机抽取,文献[32]的实验数据如图 7 所示.

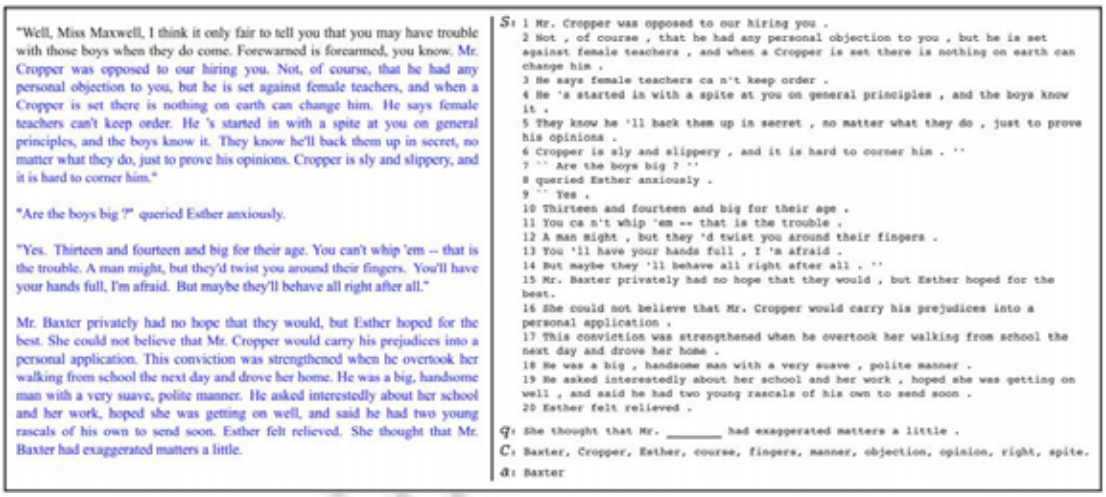


Fig.7 A named entity question from the CBT

图 7 命名实体的 CBT 语料库样例

Hill 等人的主要工作体现在如下两个方面.

- 第一,开发了自监督模型.
- 第二,利用了多种记忆向量的表示方式:(1) 词汇式向量(lexical memory),每个单词作为一个单独的记忆向量;(2) 窗口式向量(window memory),使用文章中任意位置出现的候选答案 c_i 作为句子中心,向左、右各扩展出 $D/2$ 长度的单词,以总长 D 作为窗口宽度抽取句子;(3) 句子式向量(sentential memory),如图所示的 20 个句子,每个句子作为一个单独的记忆向量.

在 CBT 的实验中,弱监督模型的多跳计算方式仅在词汇式向量模型下才有较好的表现.在测试数据中,弱监督模型的命名实体预测准确率仅为 43.1%,而使用硬关注机制(hard attention)的改进模型,准确率则为 66.6%.原因是软关注机制(soft attention)依赖于权重求和计算输出支撑因子的方式[26],如公式(16)所示.

$$m_{o1}=\sum_{i=1,\ldots,n}\alpha_i m_i,\alpha_i=e^{c_i^T q}\Big/\sum_j e^{c_j^T q}/,i=1,\ldots,n \tag{16}$$

其中,m_{o1} 代表通过 Softmax 函数后的权值求和,相当于每个 m_i 的概率.

Hill 等人在改进的记忆神经网络模型中使用窗口式向量和自监督模式实现了硬关注机制:首先,将包含 c_i 的 10 个窗口向量存入记忆细胞中,其中,包含答案的句子自动地被标注为含有支撑因子的向量;然后,使用类似于强监督模型训练方法,以最大内积的方式和随机梯度下降法训练嵌入式矩阵 A;之后,用 A 和 q 在记忆细胞中找到包含支撑因子的记忆向量 m_{o1}.这种方式称为自监督模式,具体方法如公式(17)所示.

$$m_i=A\Phi(S_i),m_{o1}=\arg\max_{i=1,\ldots,n} m_i^T q \tag{17}$$

这种依靠答案来标注记忆向量的训练方式相对于软关注机制用权重来寻找支撑因子,硬关注机制利用候选答案作为标签,能够比软关注机制更加精准、快速地找到相关支撑因子.实验中还发现:改进的记忆神经网络模型更适用于 CBT,CNN QA 这类多选择答案或语言建模数据集,特别是在寻找实体和实物作为答案的训练对

中,表现超过以往 LSTM 模型.该模型在 CNN QA 语料集上进行了测试,识别新闻文章中命名实体的准确率达到 69.4%.

3.2.4 键值记忆模型

直接阅读文本方法也有发展,文献[17,32]都曾经提出在文本中直接截取有完整含义的句子方法,文献[17]方法的弊端是每次都要训练一个抽取模型,并且模型的有效性无法保证,特别是在大数据集上训练成本过高.文献[30]提出了用知识库(knowledge base)三元组的方式将大数据集的信息分解,构建问答数据集.但是知识库的规则过于严格,包含的信息也不如原始文件充分.Miller 在文献[33]中提出了键值记忆神经网络(key-value memory neural network),这个新模型可以直接阅读文本,键值记忆神经网络与弱监督模型[26]的结构基本相同,但其在寻址和输出阶段采用不同的运算方式.文献[33]的模型如图 8 所示.

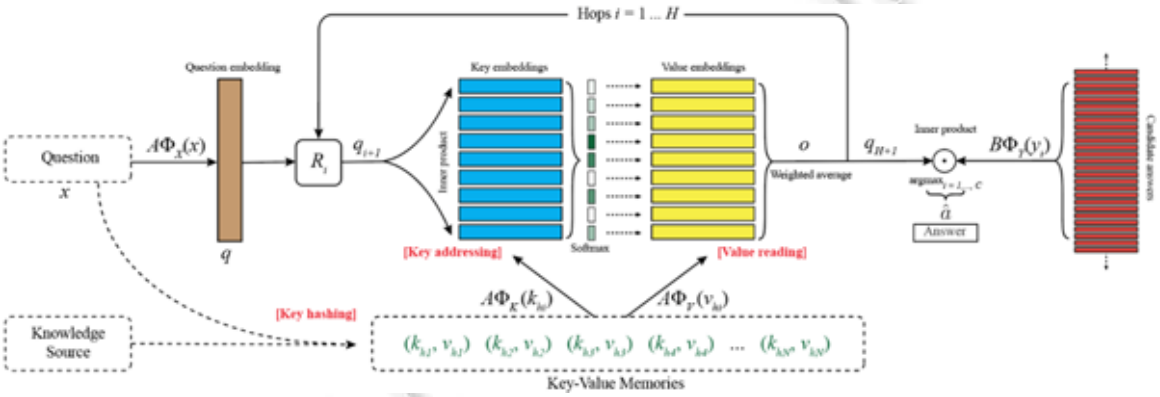


Fig.8 Key-Value memory neural networks

图 8 键值记忆神经网络

待检索的记忆细胞分成了两部分,即 Key 和 Value.在初始的弱监督模型中,相同信息与两个不同的嵌入式矩阵 A 和 C 运算,求得包含支持因子的整个句子.键值记忆模型中,只与 1 个嵌入式矩阵 A 进行运算,可以将与问题有关的部分与问题 x 紧密联系起来,如公式(18)所示.

$$P_{hi}=soft\max(A\Phi_X(x)\cdot A\Phi_K(K_{hi})) \tag{18}$$

以加权总和后的最大概率为 VALUE 的选择依据,这与文献[26]基本相同,如公式(19)所示.

$$O=\sum_i P_{hi}\cdot A\Phi V(v_{hi}) \tag{19}$$

如果需要进行多次检索和多跳计算,则 $q^2=R^1(q+o)$,比如 $q^2=R^1(q+o)$,其中,R 代表可训练矩阵.在检索过程中逐渐建立一个循环,如公式(20)所示.

$$P_{hi}=Soft\max(q_{j+i}^T A\Phi_K(K_{hi})) \tag{20}$$

最后,将与答案有关的部分与后面的候选集合联系起来,通过最大内积训练输出矩阵,最终得到答案.

文献[33]为了比较知识库、信息抽取、直接阅读维基百科文档这 3 种类型文件,设计了 4 种不同的 Key-Value 方式,比如:

(1) 知识库三元组:将主体和关系作为 Key,将客体作为 Value,通常对三元组以客体作为回答对象.

(2) 句子层面:将文档分割成为多个句子,每个句子既为 Key 也为 Value,此方法与弱监督模型相似.

(3) 窗口方式:以文档中的每个实词为中心,向左也向右开一个窗口,将窗口作为 Key、实词作为 Value.

(4) 窗口方式和中心编码:与方式(3)基本相同,区别是中心实体词与其他词采用不同的嵌入式矩阵运算.

该实验构建了新语料库 WIKIMOVIES,实验结果表明:直接利用知识库的方式好于直接从文档中读取,直接读取文档好于信息抽取.这样的结论给日后进行文本阅读实验提供了很好的参考.值得一提的是,Key-Value 模式下,直接读取文档就能获得很好的效果,不一定要将文档进行拆分.此外,文献[33]也在 WIKIQA 上进行了实

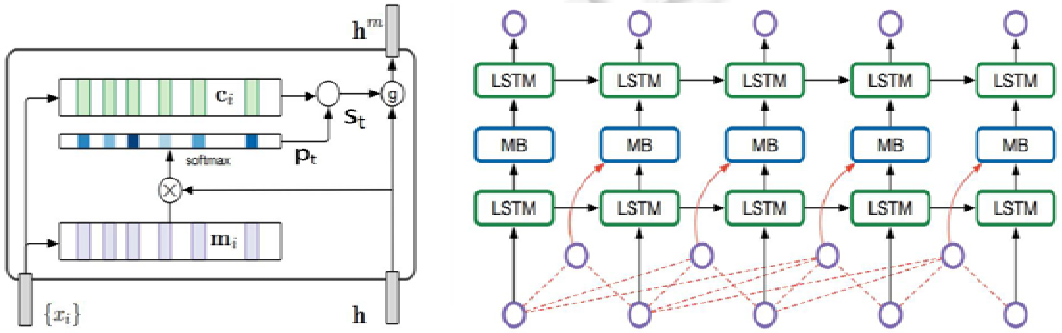
验,Key-Value 模型也取得了很好的效果.

Key-Value 模型继承了强监督的分段监督做法,又利用了弱监督端到端训练的能力.吸收了两个模型的优点,简化了强监督的训练步骤,扩大了弱监督搜索记忆细胞的能力.更重要的是,模型可以直接阅读文本中,不需要拆分文本也可以进行自然语言处理,这为以后的大语料库监督学习奠定了一些基础.

4 基于长程记忆的机器学习模型

自 2014 年以来,学术界和产业界陆续开发了基于记忆和关注度机制的机器学习模型,如 RNN 的改进模型——递归记忆神经网络(recurrent memory neural network,简称 RMN)[63]、Google DeepMind 开发的神经图灵机(NTM)[16]、MetaMind 的动态记忆神经网络(dynamic memory neural network,简称 DMN)[64].

RMN 是 LSTM 的一种创新结构,在 LSTM 模型中加入记忆模块(memory block,简称 MB).MB 结构借鉴于弱监督模型,扩大了 LSTM 在语言建模和完成句子上的能力.RMN 并没有在文献[63]中描述问答和对话等测试,仅展示了在语言建模上超过 LSTM 的基线水平的测试实验,比如英语、德语、意大利语.RMN 在完成句子测试中,准确率也超过 LSTM.文献[63]MB 的结构如图 9(a)所示,RMN 的结构如图 9(b)所示.



(a) MB 的图形化说明 (b) 4 连词 RMN 展开的图形化说明

Fig.9

图 9

从图 9(a)中可以看出,MB 的输入由两部分组成:在 t 时刻,第 1 部分为输入 $\{x_i\}$(包含 x_t 之前的 n 个最近的输入 $x_{t-n},\ldots,x_t$);第 2 部分为 LSTM 输出的隐藏层 h_t.与弱监督模型结构类似,x_i 与嵌入式矩阵 M 和 C(维度同为 $V\times d$)进行运算,以输入向量 m_i、输出向量 c_i 的形式保存在两个矩阵 M 和 C 中,用输入向量 m_i 和 h_t 进行计算,得到关注度的权重 $p_t=softmax(m_i\times h_t)$,之后,用 p_t 与输出向量 c_i 计算出一个上下文相关向量 s_t.s_t 可以看成是 x_i 加入关注度贡献后的重新表示.最终用一个类似于 GRU 的门机制来控制 s_t 和 h_t 的输出量,限制从 LSTM 到 MB 的输出.从图 9(b)中可以直观地看出 MB 与 LSTM 的结合方式,图中展示出这个 MB 的记忆块大小为 n=4,输入包含当前 t 时间单词和前 3 个单词,最后一个 LSTM 负责将 MB 中保存的信息循环输出.这种两个 LSTM 一个 MB 的结构称为 RMR.RMN 最终实现了两个目的:第一,获得更好的预测能力;第二,更好地理解 LSTM 模型隐含层的信息.

NTM 与 MemNN 同时在 2014 年被提出来,MemNN 的提出是基于自然语言处理需要,NTM 则是在一个更高的层面构建神经网络模型,实现图灵机的功能.NTM 有两个重要组件:神经网络控制器和内存模块.控制器通过输入输出向量与外界交互信息,控制器通过一个有选择性的矩阵实现对内存模块进行读写.更重要的是,NTM 中每个组件都是可微分的,这更容易使用梯度下降法训练.

Graves 等人在实验[16]中验证了 NTM 能够回忆起一个包含任意信息的长序列,但是低于 128 位.NTM 有远超 LSTM 记忆能力的表现.在学习嵌套函数的实验中,NTM 也比 LSTM 要快得多.文献[16]从实验中证明了 NTM 的架构可以从样本数据中学习简单的算法,并应用在样本之外的数据上.

DMN[64]是一种端到端的神经网络模型,在数据集 bAbI 的测试上,它的评分几乎达到了强监督模型的效果,

并且高出弱监督模型许多.得益于端到端的的训练方式,文献[64]还在情感分类、序列建模、词性标注上进行了实验,DMN 在这些实验上都有很好的表现.

DMN 构建了一种动态的记忆存储和读取方式,而不是 MemNN[17,26]上的静态结构.DMN 提出了一种新的关注机制使用方式,能够在检索中不断地提高检索效率.在文本表达方式上也与 Chen 等人的文献[34]近似,不使用嵌入式矩阵转化文本,转而使用 RNN-GRU 的方式进行编码(encode),将文本序列$\{s_1,s_2,\ldots,s_n\}$和问题 q 转化为向量.这种方式更有利于在场景式记忆模块(episodic memory module)上存储和检索.DMN 模型先是将$\{s_1,s_2,\ldots,s_n\}$问题 q 用 RNN 编码成向量,再用包含关注度机制的 RNN-GRU 模型计算出文本与问题的相似度,用关注度机制得到每一个 s_i 的相关性概率 e_i^1,当 $e_i^1=1$ 时,表示相关性高;值越低,相关性就越低.全部的 e_i^1 组合成 m^1.然后,再用本次获得的相关句子 m^1 联合 q 重新在记忆中进行迭代检索,进一步得到更多的相关性句子.迭代方式如公式(21)所示.

$$m^i=GRU(e^i,m^{i-1}) \tag{21}$$

m^i 代表 q 以及与其相关性高的 e^i.如此迭代下去,获得既与 q 相关的句子 s_i,又与 s_i 相关的句子 s_j,最终所有与问题 q 相关的句子都可以得到.之后,将问题与全部相关句子用 RNN-GRU 的解码(decode)运算得到最终答案.文献[64]中的 DMN 示意结构如图 10 所示.

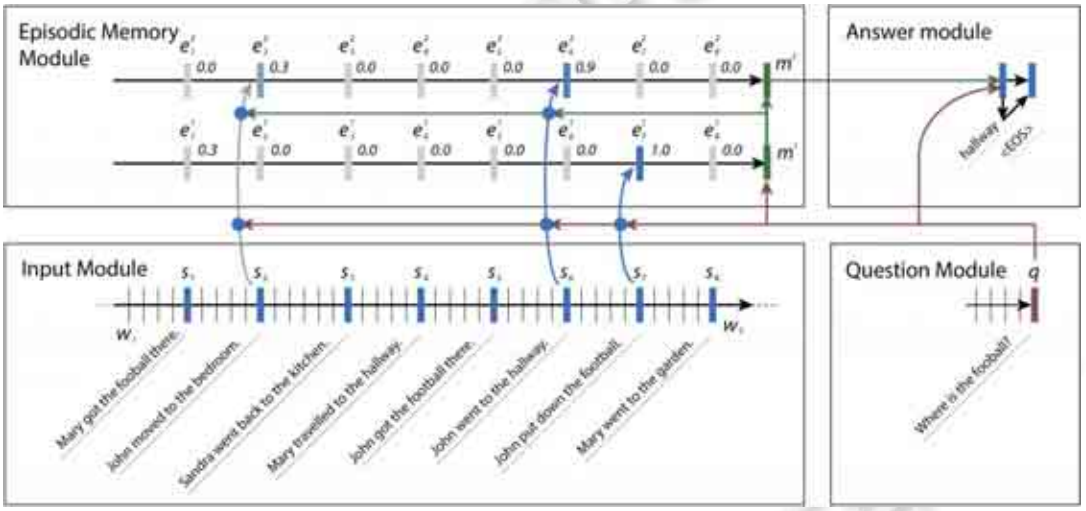


Fig.10 Dynamic memory networks model

图 10 DMN 模型

从图 10 中可以看到:当问题(where is the football?)在第 1 次检索时,第 7 句(John put down the football)因为含有 football 被关注度模型将概率 e_7^1 标记为 1,即被检索到与问题相关;第 2 次检索时,问题 q 与第 7 句同时与记忆(episodic memory module)进行相关性检索,凡是与 John 相关的都被赋予一定的概率.这意味着通过连续检索可以得到足够多的信息,这些信息构成了与答案相关的支撑因子,这正是 DMN 的连续检索模式特色.

DMN 是基于动态记忆的全新神经网络模型,它的记忆与检索存在着交互,随着检索过程的持续,记忆的可读范围不断地变化,以帮助检索到更多的相关信息.该模型将 RNN-GRU 应用于数据的运算和检索,并利用关注度机制直观地表示了支撑因子上概率的变化.DMN 相对于 MemNN 在记忆模块的使用上有更大的灵活性,同时,由于关注度机制的频繁使用,在大数据下,DMN 可能要耗费大量的运算时间.

带有记忆元素的 3 个神经网络模型 MemNN,NTM,DMN 都共享着 5 个元素:记忆模块、输入模块、读取模块、写模块、输出模块,所不同的是,MemNN 的记忆模块几乎不可擦写,它更多地关注于如何更好地读取记忆.但如果记忆模块的容量有限,写模块就有很大的用处,写模块可以确定哪些区域重写、哪些区域擦除,这也是

NTM 与 MemNN 的不同之处.DMN 的读取方式相对于 MemNN 更加灵活,检索准确度更高.这 3 个模型都有端到端的结构,使用了关注度机制,并且是全程可微的,也都使用误差反向传播的方式进行训练,这是三者的共同之处.在未来的发展中,这 3 个模型会互相借鉴,在各自的模型上进行模型创新,并在应用范围上进行领域创新.

5 总结与展望

随着近两年 MemNN 的发展,利用外部存储形式的机器学习方式已经成为机器学习领域中一个不可或缺的方向,然而存储型神经网络的研究还是处于起步阶段,与 RNN 相比还年轻.目前,MemNN 所应用的范围还局限在自然语言理解领域,在机器学习的热门领域,如图像分析和音频识别等方向上应用较少.尽管 MemNN 在文本分析领域中端到端的学习方式有进展,但是在无监督学习上依然缺少相关实验.

在无监督学习中,让 MemNN 自动地从文本中学习到表征文本本质的特征,则会让研究人员更好地利用计算机实现文本的理解.鉴于在问答数据集和情感分析数据集上的良好表现,相信日后 MemNN 在无监督学习的研究中自主地理解文本信息将会是一个热点方向.当前的深度学习研究类似于学习一项运动技能,如学投篮,虽然学会了某项技能,但不涉及大量的知识.自然语言理解等领域要求我们必须记住事实,必须存储某些知识,在这一点上,记忆神经网络模型相对于传统神经网络模型有较大优势.利用这种优势,可以让模型读入大量文本,如新闻文本,把它们全部聚类,就会自动分成几十个不同的组,如娱乐、科技、政治等,然后,通过这种方式提取某类文件的特征,将其应用在文件分类之中.也可以使用类似于 Google 的虚拟人脑[65]的方式,通过大量文本作为训练集,训练一个深度自编码器网络.之后,在完全没有标签的情况下,自动地从训练集中学习到某些名词在自然语言中的概念.Iyyer 等人[66]提出了一种无监督学习模型——关系建模网络(relationship modeling network),并提出通过结合深度循环自编码器和字典学习来对关系描述符进行学习,实现对小说中复杂人物关系的梳理.

相对于先前的机器学习模型在训练结束后就与原始数据进行割离,不再具有重新访问的能力,MemNN 依然保存着原始的训练数据,这种情况下,MemNN 可以进行某种程度的自监督训练,利用新输入的数据和原有数据进行重新训练,这是也一种增强学习的方式.针对以上问题,已有实验结果表明,联合训练后的模型准确率高于单独训练.

鉴于 MemNN 在问答系统的迁移学习上有良好的表现,可见迁移学习方法在大数据系统上有应用前景.依托于 MemNN 的记忆和寻址能力,它将具有更强的知识迁移能力.相对于 LSTM 模型,MemNN 的记忆大小理论上是不受到限制的,因此,MemNN 可以在公共知识库上进行迁移学习.

在记忆神经网络的发展中,首先实现了模型创新,出现了两种既互为联系又相互独立的模型——强监督模型和弱监督模型,分别依靠自身的优势应用在更广阔的领域之中,实现了应用上的创新,应用创新又反过来促进两种模型各自的模型创新.与此同时,也出现了融合两种模型优点的新模型.由此可见,随着研究的深入,记忆神经网络具有广阔的应用前景.

References:

[1] Hinton GE, Osindero S, Teh Y. A fast learning algorithm for deep belief nets. In: Proc. of the Neural Computation. 2006. [doi: 10.1162/neco.2006.18.7.1527]

[2] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the Advances in Neural Information Processing Systems. 2012. 1097−1105.

[3] Hausknecht M, Stone P. Deep recurrent Q-learning for partially observable MDPs. In: Proc. of the Computer Science. 2015. https://arxiv.org/abs/1507.06527

[4] Pollack J. The induction of dynamical recognizers. In: Proc. of the Machine Learning. 1991. 227−252.

[5] Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. In: Proc. of the Computer Science. 2012. 212−223.

[6] Xiong W, Droppo J, Huang X, Seide F, Seltzer M, Stolcke A, Yu D, Zweig G. The Microsoft 2016 conversational speech recognition system. arXiv preprint arXiv:1609. 03528, 2016. https://arxiv.org/abs/1609.03528

[7] Saon G, Sercu T, Rennie SJ, Kuo HKJ. The IBM 2016 English conversational telephone speech recognition system. arXiv 2016. https://arxiv.org/abs/1604.08242

[8] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-Based learning applied to document recognition. Proc. of the IEEE, 1998, 86(11):2278−2324. [doi: 10.1109/5.726791]

[9] Elman JL. Finding structure in time. In: Proc. of the Cognitive Science. 1990. 179−211. [doi: 10.1016/0364-0213(90)90002-E]

[10] Mikolov T, Karafiát, Burget L, Černocký JH, Khudanpur S. Recurrent neural network based language model. In: Proc. of the Interspeech. Conf. of the Int'l Speech Communication Association (INTERSPEECH 2010). 2010. 1045−1048.

[11] Hochreiter S, Schmidhuber J. Long short-term memory. In: Proc. of the Neural Computation. 1997. 1735−1780.

[12] Sundermeyer M, Schluter R, Ney H. LSTM neural networks for language modeling. In: Proc. of the Interspeech. 2012. 194−197.

[13] Mikolov T. Statistical language models based on neural networks [Ph.D. Thesis]. Brno University of Technology, 2012.

[14] Mozer MC, Das S. A connectionist symbol manipulator that discovers the structure of context-free languages. In: Proc. of the NIPS. 1993. 863−870.

[15] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proc. of the Advances in Neural Information Processing Systems. 2014. 3104−3112.

[16] Graves A, Wayne G, Danihelka I. Neural turing machines. In: Proc. of the Computer Science. 2014. https://arxiv.org/abs/1410.5401

[17] Weston J, Chopra S, Bordes A. Memory networks. In: Proc. of the Int'l Conf. on Representation Learning (ICLR 2015). 2015. https://arxiv.org/abs/1410.3916

[18] Shang LF, Lu ZD, Li H. Neural responding machine for short-text conversation. In: Proc. of the ACL. 2015. https://arxiv.org/abs/1503.02364

[19] Palangi H, Deng L, Shen YL, Gao JF, He XD, Chen JS, Song XY, Ward R. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. IEEE/ACM Trans. on Audio Speech & Language Processing, 2016,24(4):694−707. [doi: 10.1109/TASLP. 2016.2520371]

[20] Lin TN, Horne BG, Tiňo P, Giles CL. Learning long-term dependencies in narx recurrent neural networks. IEEE Trans.on Neural Networks, 1996,7(6):1329−1338. [doi: 10.1109/72.548162]

[21] Wang D, Nyberg E. A long short-term memory model for answer sentence selection in question answering. In: Proc. of the Meeting of the Association for Computational Linguistics and the Int'l Joint Conf. on Natural Language Processing. 2015. 707−712. [doi: 10.3115/v1/P15-2116]

[22] Tan M, Santos CD, Xiang B, Zhou BW. LSTM-Based deep learning models for non-factoid answer selection. In: Proc. of the Computer Science. 2015. https://arxiv.org/abs/1511.04108

[23] Gao H, Mao J, Zhou J, Huang Z, Wang L, Xu W. Are you talking to a machine? Dataset and methods for multilingual image question answering. In: Proc. of the Computer Science. 2015. 2296−2304.

[24] Das S, Giles CL, Sun GZ. Learning context-free grammars: Capabilities and limitations of a recurrent neural network with an external stack memory. In: Proc. of the 14th Annual Conf. of Cognitive Science Society. 1992. 791−795.

[25] bAbI. 2014. https://github.com/facebook/bAbI-tasks

[26] Sukhbaatar S, Weston J, Fergus R. End-to-End memory networks. In: Proc. of the Advances in Neural Information Processing Systems. 2015. https://arxiv.org/abs/1503.08895v5

[27] Tang DY, Qin B, Liu T. Aspect level sentiment classification with deep memory network. In: Proc. of the EMNLP. 2016. [doi: 10.18653/v1/D16-1021]

[28] Dodge J, Gane A, Zhang X, Bordes A, Chopra S, Miller AH, Szlam A, Weston J. Evaluating prerequisite qualities for learning end-to-end dialog systems. In: Proc. of the ICLR. 2016. https://arxiv.org/abs/1511.06931

[29] Weston J, Bordes A, Chopra S, Mikolov T. Towards AI-complete question answering: A set of prerequisite toy tasks. arXiv preprint arXiv:1502.05698, 2015. https://arxiv.org/abs/1502.05698

[30] Bordes A, Usunier N, Chopra S, Weston J. Large-Scale simple question answering with memory networks. In: Proc. of the Computer Science. 2015. https://arxiv.org/abs/1506.02075

[31] Chandar S, Ahn S, Larochelle H, Vincent P, Tesauro G, Bengio Y. Hierarchical memory networks. arXiv:1605.07427, 2016. https://arxiv.org/abs/1605.07427

[32] Hill F, Bordes A, Chopra S, Weston J. The Goldilocks principle: Reading children's books with explicit memory representations. In: Proc. of the Computer Science. 2015. https://arxiv.org/abs/1511.02301

[33] Miller AH, Fisch A, Dodge J, Karimi AH, Bordes A, Weston J. Key-Value memory networks for directly reading documents. arXiv: 1606.03126v2, 2016. https://arxiv.org/abs/1606.03126

[34] Chen YN, Hakkani-Tür D, Tur G, Gao JF, Deng L. End-to-End memory networks with knowledge carryover for multi-turn spoken language understanding. In: Proc. of the Meeting of the Int'l Speech Communication Association. 2016. [doi: 10.21437/Interspeech. 2016-312]

[35] Karpathy A, Johnson J, Li FF. Visualizing and understanding recurrent networks. arXiv preprint arXiv:1506.02078, 2015. https:// arxiv.org/abs/1506.02078

[36] Hierarchical memory networks for answer selection on unknown words. arXiv preprint arXiv:1609.08843, 2016. https://arxiv.org/ abs/1609.08843

[37] Sukhbaatar S, Szlam A, Weston J, Fergus R. Weakly supervised memory networks. arXiv preprint arXiv:1503.08895, 2015. https:// arxiv.org/abs/1503.08895v2

[38] Luong T, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. 2015. 1412−1421. [doi: 10.18653/v1/D15-1166]

[39] Rush AM, Chopra S, Weston J. A neural attention model for abstractive sentence summarization. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. 2015. 379−389.

[40] Larochelle H, Hinton G. Learning to combine foveal glimpses with a third-order Boltzmann machine. In: Proc. of the Int'l Conf. on Neural Information Processing Systems. 2010. 1243−1251.

[41] Denil M, Bazzani L, Larochelle H, de Freitas N. Learning where to attend with deep architectures for image tracking. In: Proc. of the Neural Computation. 2012. 2151−2184. [doi: 10.1162/NECO_a_00312]

[42] Mnih V, Heess N, Graves A, Kavukcuoglu K. Recurrent models of visual attention. In: Proc. of the Computer Science. 2014. 2204− 2212.

[43] Gregor K, Danihelka I, Graves A, Rezende DJ, Wierstra D. DRAW: A recurrent neural network for image generation. In: Proc. of the Computer Science. 2015. 1462−1471.

[44] Moritz HK, Kočiský T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P. Teaching machines to read and comprehend. In: Proc. of the Computer Science. 2015.

[45] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). 2015. https://arxiv.org/abs/1409.0473

[46] Marcus MP, Marcinkiewicz MA, Santorini B. Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics, 1993,19(2):313−330.

[47] Mikolov T, Joulin A, Chopra S, Mathieu M, Ranzato MA. Learning longer memory in recurrent neural networks. In: Proc. of the Computer Science. 2014. https://arxiv.org/abs/1412.7753

[48] Berant J, Chou A, Frostig R, Liang P. Semantic parsing on freebase from question-answer pairs. In: Proc. of the EMNLP. 2013. 1533−1544.

[49] Yang MC, Duan N, Zhou M, Rim HC. Joint relational embeddings for knowledge-based question answering. In: Proc. of the EMNLP. 2014. 645−650. [doi: 10.3115/v1/D14-1071]

[50] Reddit. 2015. http://files.pushshift.io/reddit/comments/

[51] Bordes A, Chopra S, Weston J. Question answering with subgraph embeddings. In: Proc. of the Computer Science. 2014. [doi: 10. 3115/v1/D14-1067]

[52] Koren Y, Bell R. Advances in Collaborative Filtering. Recommender Systems Handbook. Springer-Verlag, 2015. 77−118. [doi: 10. 1007/978-1-4899-7637-6_3]

[53] Koren Y. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In: Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2008. 426−434. [doi: 10.1145/1401890.1401944]

[54] Lee IC, Michael K, Dave K, Satinder S, Peter S. Cobot in LambdaMOO: A social statistics agent. In: Proc. of the AAAI/IAAI. 2000. 36−41.

[55] Jafarpour S, Burges CJC, Ritter A. Filter, rank, and transfer the knowledge: Learning to chat. In: Proc. of the Advances in Ranking. 2010. https://core.ac.uk/display/21317606

[56] Ritter A, Cherry C, Dolan WB. Data-Driven response generation in social media. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2011. 583−593.

[57] Sordoni A, Galley M, Auli M, Brockett C, Ji YF, Mitchell M, Nie JY, Gao JF, Dolan B. A neural network approach to contextsensitive generation of conversational responses. In: Proc. of the NAACL. 2015. [doi: 10.3115/v1/N15-1020]

[58] Dong L, Wei FR, Tan CQ, Tang DY, Zhou M, Xu K. Adaptive recursive neural network for target-dependent twitter sentiment classification. In: Proc. of the ACL. 2014. 49−54.

[59] Lakkaraju H, Socher R, Manning C. Aspect specific sentiment analysis using hierarchical deep learning. In: Proc. of the NIPS Workshop on Deep Learning and Representation Learning. 2014. https://www.mendeley.com/research-papers/aspect-specific-sentiment-analysis-using-hierarchical-deep-learning-1/

[60] Nguyen TH, Shirai K. Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language. 2015. [doi: 10.18653/v1/D15-1298]

[61] Tang DY, Qin B, Feng XC, Liu T. Target-Dependent sentiment classification with long short term memory. In: Proc. of the Computer Science. 2015. https://arxiv.org/abs/1512.01100v1

[62] Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H, Androutsopoulos I, Manandhar S. Semeval-2014 task 4: Aspect based sentiment analysis. In: Proc. of the 8th Int'l Workshop on Semantic Evaluation. 2014. 27−35.

[63] Tran K, Bisazza A, Monz C. Recurrent memory networks for language modeling. In: Proc. of the NAACL-HLT. 2016. [doi: 10.18653/v1/n16-1036]

[64] Ankit K, Irsoy O, Ondruska P, Iyyer M, Bradbury J, Gulrajani I, Zhong V, Paulus R, Socher R. Ask me anything: Dynamic memory networks for natural language processing. In: Proc. of the Computer Science. 2015. https://arxiv.org/abs/1506.07285

[65] Le QV. Building high-level features using large scale unsupervised learning. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. IEEE, 2013. 8595−8598.

[66] Iyyer M, Guha A, Chaturvedi S, Boyd-Graber J, Daumé H III. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In: Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016. 1534−1544. [doi: 10.18653/v1/N16-1180]

梁天新(1984－),男,黑龙江齐齐哈尔人,博士生,主要研究领域为机器学习,神经网络,社会计算.

张永俊(1986－),男,博士生,主要研究领域为机器学习,数据挖掘,异构信息网.

杨小平(1956－),男,博士,教授,博士生导师,主要研究领域为信息系统工程,电子政务,网络安全技术.

朱艳丽(1975－),女,博士生,副教授,主要研究领域为机器学习,知识图谱.

王良(1963－),男,博士,副教授,CCF 高级会员,主要研究领域为智能科学,数据库管理系统,数据库系统评价和性能优化.

许翠(1995－),女,硕士生,主要研究领域为机器学习,神经网络.