融合动态卷积注意力的机器阅读理解研究

吴春燕[1],李 理[1],黄鹏程[1],刘知贵[1,2],张小乾[2]
(1.西南科技大学 计算机科学与技术学院,四川 绵阳 621000;
2.西南科技大学 信息工程学院,四川 绵阳 621000)

摘 要:针对机器阅读理解在采用长短期记忆神经网络和注意力机制处理文本序列信息时,存在特征信息提取不足和预测结果准确性不高的问题,提出了一种融合动态卷积注意力的片段抽取型机器阅读理解模型。该模型考虑到LSTM的当前输入和之前的状态相互独立,可能会导致上下文信息丢失,采用Mogrifier作为编码器,让当前输入与前一个状态充分交互多次,增强上下文和问题中的显著结构特征并减弱其次要特征;其次,由于静态卷积的卷积核相同,只能提取固定长度文本的特征,这可能对机器更好的理解文本产生阻碍,通过引入动态卷积,采用多个不同卷积核的一维卷积来捕获上下文和问题的局部结构,弥补注意力机制只有全局捕获能力的缺点。在SQuAD数据集上的实验结果表明,与其他模型相比,该方法有效提升了模型在特征信息提取和答案预测方面的能力。

关键词:机器阅读理解;片段抽取;答案预测;长短期记忆神经网络;动态卷积
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2023)07-0160-07
doi:10.3969/j.issn.1673-629X.2023.07.024

Study on Machine Reading Comprehension Hybriding Dynamic Convolution Attention

WU Chun-yan[1],LI Li[1],HUANG Peng-cheng[1],LIU Zhi-gui[1,2],ZHANG Xiao-qian[2]
(1.School of Computer Science and Technology,Southwest University of Science and Technology,
Mianyang 621000,China;
2.School of Information Engineering,Southwest University of Science and Technology,Mianyang 621000,China)

Abstract:To solve the problems of insufficient feature information extraction and low accuracy of prediction results when using long short-term memory and attention mechanism to process text sequence information in machine reading comprehension,we propose a span-extracting machine reading comprehension model hybriding dynamic convolution attention. Considering that the current input and the previous state of LSTM are independent of each other,which may lead to the loss of context information,the Mogrifier is adopted as the encoder,which makes the current input fully interact with the previous state several times,so as to enhance the significant structural features in the context and the problem and weaken the secondary features. Secondly,because the convolution kernel of static convolution is the same,only the features of fixed length text can be extracted,which may hinder the machine from better understanding the text. By introducing dynamic convolution,one-dimensional convolution of multiple different convolution kernels is used to capture the local structure of the context and the problem,which makes up for the disadvantage that the attention mechanism has only global capture ability. Experimental results on SQuAD datasets show that compared with other models,the proposed method can effectively improve the model's ability in feature information extraction and answer prediction.

Key words:machine reading comprehension;span-extracting;answer prediction;long short-term memory;dynamic convolution

0 引 言

机器阅读理解(Machine Reading Comprehension,MRC)要求机器阅读并理解人类自然语言文本,在此基础上,回答跟文本信息相关的问题[1]。该任务通常被用来衡量机器理解自然语言的能力,可以帮助人类从大量文本中快速聚焦相关信息,降低人工获取信息的成本。作为自然语言处理(Natural Language Processing,NLP)的研究方向之一,机器阅读理解近年

收稿日期:2022-08-26 **修回日期**:2022-12-28
基金项目:国家自然科学基金(62102331,62176125,61772272)
作者简介:吴春燕(1998-),女,硕士研究生,研究方向为深度学习、自然语言处理;通信作者:刘知贵(1966-),男,教授,硕、博士导师,研究方向为计算机应用技术、控制理论应用及自动化装置。

来已受到工业界和学术界广泛的关注。

机器阅读理解模型的研究历史可以追溯到 20 世纪 70 年代[2],当时的研究人员已经意识到机器阅读理解可以作为测试计算机语言理解能力的一种方法。其中最具代表性的是 1977 年 Lehnert 提出的 QUALM 问答程序[3],该程序专注于实用主义,为机器阅读理解提供了发展的远景。然而由于其规模小、领域特殊等限制,使得该系统无法推广到更广泛的领域。受限于当时数据集和技术的发展,这一领域的研究进展缓慢。直到二十世纪初,随着社会的发展和进步,一些用于阅读理解的大规模数据集相继被提出,如 Mctest[4]、Stanford Question Answering Dataset (SQuAD)[5]、RACE[6]等,这些数据集使得研究者们能够用深层神经网络结构模型解决阅读理解任务[7]。

目前机器阅读理解主要采用的深度学习技术包括卷积神经网络(Convolutional Neural Networks,CNN)、循环神经网络(Recurrent Neural Network,RNN)和注意力机制。其中 CNN 擅长提取局部特征;注意力机制旨在关注全局结构特征;RNN 则在序列建模中表现优异。然而,目前的工作仅聚焦在使用 RNN 和注意力机制对文本进行全局建模,忽略了对文本局部结构的捕获,导致模型对文本理解不足,回答问题不准确。针对这一问题,该文提出了一个融合动态卷积注意力的机器阅读理解模型。主要工作内容如下:

(1)采用改进的长短期记忆网络(Long Short-Term Memory,LSTM)——Mogrifier 作为编码器,让输入与前一个状态进行多次交互,防止上下文信息流失。

(2)将动态卷积和注意力机制结合同时提取文章局部和全局结构特征,增强基线模型文本建模的能力,提高模型性能。并在公共数据集 SQuAD 上进行实验,分析实验结果和模型结构对模型性能的影响。

1　相关工作

1.1　任务定义

机器阅读理解任务通常被定义为[8]:给定长度为 m 的文章 P,即 $P=\{p_1,p_2,\cdots,p_m\}$,长度为 n 的问题 $Q=\{q_1,q_2,\cdots,q_n\}$,模型需要通过学习函数 F 使 $F(P,Q)\rightarrow A$,从中提取连续子序列 $A=\{a_i,\cdots,a_{i+k}\}$(其中 $1\leqslant i\leqslant i+k\leqslant m$)作为问题 Q 的正确答案。训练数据为文章、问题、答案组成的三元组 $<P,Q,A>$。

1.2　基于深度学习的机器阅读理解相关研究

为了捕获文章和问题的语言特征,注意力机制、循环神经网络及其变体长短期记忆网络和门控循环单元(Gated Recurrent Unit,GRU)和卷积神经网络在模型中表现出优异的性能。早期的模型中采用简单的注意力机制,如 Hermann 等人[9]提出的 Attentive Reader,通过计算问题和文章之间的注意权重得到它们的交互信息,Kadlec 等[10]提出的 Attention Sum Reader 和 Chen 等[11]提出的 The Stanford Attentive Reader 模型在一定程度上提升了文本相似度的计算能力。然而,这些模型中的注意力无法理解文本的深层含义。

针对这一问题,研究者们开始对深层注意力进行研究。Seo 等[12]提出的 BiDAF 模型同时计算文章到问题和问题到文章两个方向的注意力权重,以获得它们之间更深层的交互信息,达到增强模型的语义表示能力的目的。Chen 等[13]通过将词性等语法特征融入词嵌入层,丰富词的向量表示,经过模型处理得到答案。Wang 等[14]提出 R-Net 模型,使用门控的基于注意力的循环网络来计算文章和问题的相似度,以获得问题感知的文章表示,之后通过自匹配的注意力机制改善文章表示,实现整个文章的有效编码。Huang 等[15]提出了 Fusion- Net 模型,通过单词级注意力、句子级注意力等不同层次的特征注意融合作为输入,同时使用所有层的表示,达到更好的文本理解。Yu 等[16]提出了一个不含 RNN 网络的架构,仅由注意力和卷积组成的机器阅读理解模型 QANet,虽然带来了训练和推理速度的提升,但无法表示出句子深层次的含义。

以上提出的模型往往采用注意力机制、RNN(LSTM、GRU)和卷积三者中的部分组合对上下文和问题进行交互建模,虽然注意力可以解决长距离的依赖关系,但深层次的注意往往过度集中在单个标记上,而忽略局部信息的利用,难以表示长序列;RNN 由于其顺序特性,不能并行处理,使得模型在训练和推理方面都很耗时;卷积由于其窗口滑动的特性,只能捕捉文章和问题的局部特征。因此,如何利用它们的优点以构建更有效的语言特征提取模型是当前研究的重点任务。

2　DCAM 模型结构设计

针对片段抽取型机器阅读理解目前存在的语义信息提取不足等问题,提出一种融合动态卷积注意力的机器阅读理解模型(hybriding Dynamic Convolution Attention mechanisms machine reading comprehension Model,DCAM),经过编码层获得文本的序列表示,再通过问题对文章的注意权重得到融合问题特征的文章表示,采用结合动态卷积的注意力机制捕获问题感知的文章表示中的局部和全局关系,利用自注意力机制进一步挖掘文本之间的联系,经过两层双向 LSTM 建模后传入输出层,得到预测答案的起始位置。该模型一共包含词嵌入层、编码器层、多注意力层以及答案输出层四个部分,其整体结构如图 1 所示。

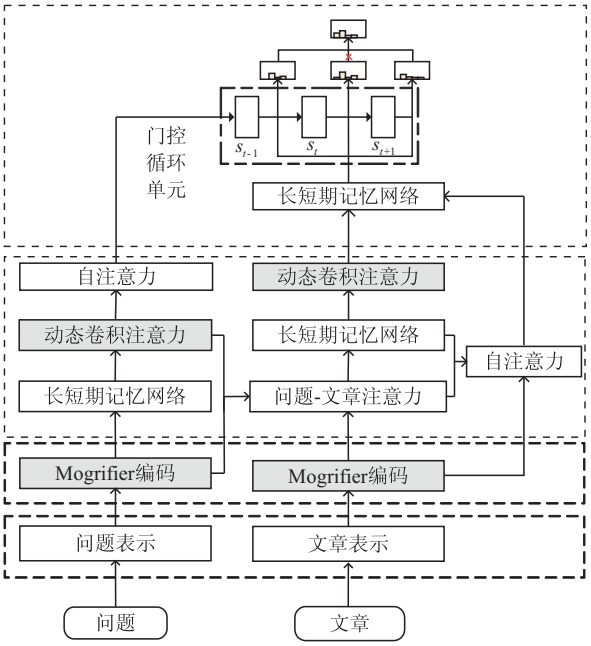

图 1 融合动态卷积注意力的机器阅读理解结构

2.1 词嵌入层

该层旨在将文章 P 和问题 Q 中的词表示特征映射到高维空间,获取词嵌入的一种典型技术是将单词嵌入与其他特征嵌入连接起来作为最终的词向量表示。单词嵌入使用预训练的 300 维 GloVe 向量[17]来表示 Q 和 P,其中文章 P 中的每个单词 p_i 额外使用三种特征嵌入和问题增强嵌入,特征嵌入分别为 9 维词性标签嵌入、8 维命名实体识别嵌入和 3 维二进制精确匹配特征嵌入;问题增强嵌入由问题表示经过一个 280 维的单层神经网络得到。

在词嵌入层,文章中的每个标记 $\boldsymbol{p}_i$ 表示为一个 600 维向量,问题中的每个标记 $\boldsymbol{q}_j$ 表示为一个 300 维向量。为了解决维度不匹配的问题,采用两层独立的全连接位置前馈网络,将段落和问题词汇编码映射到相同数量的维度。

$$\mathrm{FFN}(x) = \boldsymbol{W}_2\mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{x} + b_1) + b_2 \tag{1}$$

其中,$\boldsymbol{x}$ 为段落和问题的词汇编码,$\boldsymbol{W}_1$、$\boldsymbol{W}_2$、b_1、b_2 为需要学习的参数。

通过词嵌入层输出得到 P 中每个标记最终的词汇矩阵 $\boldsymbol{E}^p \in \mathbb{R}^{d\times m}$ 和 Q 中每个标记最终的词汇嵌入矩阵 $\boldsymbol{E}^q \in \mathbb{R}^{d\times n}$,其中 d 表示全连接神经网络隐藏层的大小,m 表示文章 P 的长度,n 表示问题 Q 的长度。将文章和问题的词汇矩阵作为下一个模块的输入。

2.2 编码器层

由于原始 LSTM 中输入 x 和之前的状态 $\boldsymbol{h}_{\text{prev}}$ 是完全独立的,可能导致上下文信息的流失。该模型在编码层采用 Melis 等[18]提出的 Mogrifier 替代传统的 LSTM,将输入 $\boldsymbol{x}$ 与之前的状态 $\boldsymbol{h}_{\text{prev}}$ 进行多次交互,再输入到各个门里进行运算。其结构如图 2 所示。

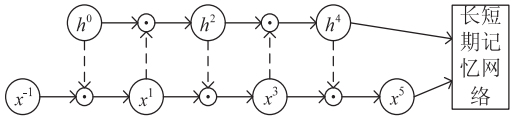

图 2 Mogrifier 结构

在普通的 LSTM 计算之前,交替地让 $\boldsymbol{x}$ 与 $\boldsymbol{h}_{\text{prev}}$ 交互,即:

$$\mathrm{Mogrify}(\boldsymbol{x}, c_{\text{prev}}, \boldsymbol{h}_{\text{prev}}) = \mathrm{LSTM}(\boldsymbol{x}^{\uparrow}, c_{\text{prev}}, \boldsymbol{h}_{\text{prev}}^{\uparrow}) \tag{2}$$

其中,$\boldsymbol{x}^{\uparrow}$、$\boldsymbol{h}_{\text{prev}}^{\uparrow}$ 为 $\boldsymbol{x}^i$、$\boldsymbol{h}_{\text{prev}}^i$ 中上标最大的值,c_{prev} 为状态。该过程如公式(3)(4)所示。

$$x^i = 2\sigma(\boldsymbol{Q}^i\boldsymbol{h}_{\text{prev}}^{i-1}) \odot x^{i-2} \quad \text{for odd } i \in [1,2,\cdots,r] \tag{3}$$

$$h_{\text{prev}}^i = 2\sigma(\boldsymbol{R}^i\boldsymbol{x}^{i-1}) \odot h_{\text{prev}}^{i-2} \quad \text{for even } i \in [1,2,\cdots,r] \tag{4}$$

其中,$\odot$ 表示哈达玛积,$\boldsymbol{Q}^i$ 和 $\boldsymbol{R}^i$ 为随机初始化矩阵,r 是交互轮数,若 $r = 0$,则为普通的 LSTM。

文章和问题都使用一个两层的 Mogrifier 将词汇嵌入投射到上下文嵌入,再拼接一个预训练的 600 维上下文 CoVe 向量[19]$\boldsymbol{C}^p$、$\boldsymbol{C}^q$,作为上下文编码层的最终输入,并将第一个上下文编码层的输出作为第二个编码层的输入。为了减少参数大小,在每个 Mogrifier 层上使用一个 maxout 层[20]来缩小矩阵的维度。通过连接两个 Mogrifier 层的输出,得到文章 P 的最终表示 $\boldsymbol{H}^p \in \mathbb{R}^{2d\times m}$ 和问题 Q 的最终表示 $\boldsymbol{H}^q \in \mathbb{R}^{2d\times n}$,其中 d 为 Mogrifier 的隐藏层大小。

$$\boldsymbol{H}^p = \mathrm{BiMogrifier}(\boldsymbol{E}^p; \boldsymbol{C}^p) \tag{5}$$

$$\boldsymbol{H}^q = \mathrm{BiMogrifier}(\boldsymbol{E}^q; \boldsymbol{C}^q) \tag{6}$$

其中,;表示向量/矩阵串联运算符。

2.3 多注意力层

注意力机制作为一种权重分配机制,可以对重要的语义信息分配较多的注意力。在阅读理解任务中,文章和问题中不同的词对问题的回答的影响是不同的,因此在模型中采用多注意力机制来识别文章和问题中哪些词与答案最相关,该层是模型的核心部分。

2.3.1 互注意力机制

将 Mogrifier 的输出作为该层的输入,首先利用点积注意力计算 Q 和 P 中词汇标记的对齐矩阵,并使用该矩阵得到问题感知的段落表示。

$$\boldsymbol{M} = \mathrm{dropout}(f_{\text{query2passage\_attention}}(\hat{\boldsymbol{H}}^q, \hat{\boldsymbol{H}}^p)) \in \mathbb{R}^{m\times n} \tag{7}$$

其中,$\boldsymbol{M}$ 为一个对齐矩阵。$\hat{\boldsymbol{H}}^p$ 和 $\hat{\boldsymbol{H}}^q$ 分别表示 $\boldsymbol{H}^p$ 和 $\boldsymbol{H}^q$ 通过单层神经网络 $\mathrm{ReLU}(\boldsymbol{W}_2 x)$ 转换得到。

2.3.2 动态卷积注意力机制

由于单独的注意力机制会受到分散权重的影响,不适合长序列表征学习。而结合卷积的注意力机制[21]混合了逐点变换、卷积和自注意力机制,可以并

行学习文本的多角度多层次序列表示。因此,在基线模型中加入结合动态卷积的注意力机制(Dynamic Convolution Attention,DCA),该注意力机制包含三个主要部分:捕获全局特征的自注意力机制,捕获局部特征的动态深度可分离卷积,以及用于捕获标记特征的位置前馈网络。该模块获取前一层的输出矩阵 $\boldsymbol{M}$ 作为输入,并以融合的方式生成输出表示:

$$\boldsymbol{C} = \boldsymbol{M} + \mathrm{Att}(\boldsymbol{M}) + \mathrm{Conv}(\boldsymbol{M}) + \mathrm{Pointwise}(\boldsymbol{M}) \tag{8}$$

其中,Att 表示自注意力机制,Conv 表示动态卷积,Pointwise 表示位置前馈网络。图 3 表示了该注意力的详细结构。

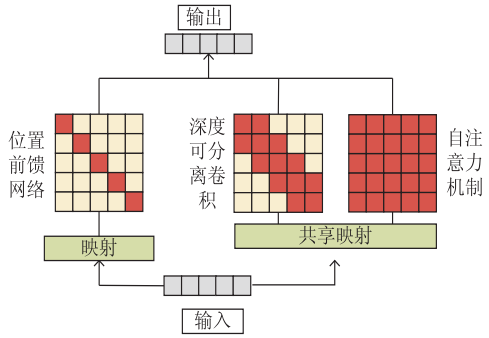

图 3　动态卷积注意力机制结构

自注意力机制负责学习全局语境的表征。对于前一层的输入序列 $\boldsymbol{M}$,它首先将 $\boldsymbol{M}$ 进行线性变换产生键 $\boldsymbol{K}$、查询 $\boldsymbol{Q}$ 和值 $\boldsymbol{V}$,然后使用自注意力机制来获得输出表示:

$$\mathrm{Att}(\boldsymbol{M}) = \sigma(Q\boldsymbol{W}^{Q}, K\boldsymbol{W}^{K}, V\boldsymbol{W}^{V})\boldsymbol{W}^{O} \tag{9}$$

其中,$\boldsymbol{Q} = \mathrm{Linear}_1(\boldsymbol{M})$,$\boldsymbol{K} = \mathrm{Linear}_2(\boldsymbol{M})$,$\boldsymbol{V} = \mathrm{Linear}_3(\boldsymbol{M})$,$\boldsymbol{W}^{O}$、$\boldsymbol{W}^{Q}$、$\boldsymbol{W}^{K}$ 和 $\boldsymbol{W}^{V}$ 为权重矩阵,σ 是键、查询和值对之间的点积生成,见公式(10):

$$\sigma(Q_1, K_1, V_1) = \mathrm{softmax}\left(\frac{Q_1 K_1^{\mathrm{T}}}{\sqrt{d_k}}\right)V_1 \tag{10}$$

为了和自注意力在相同的映射空间中学习上下文序列表示,选取深度方向卷积[22]的变体—动态卷积[23]进行卷积运算。每个卷积子模块包含多个内核大小不同的单元,用于捕捉不同范围的特征。卷积核大小为 k 的卷积单元的输出为:

$$\mathrm{Conv}_k(\boldsymbol{M}) = \mathrm{Depth\_conv}_k(V_2)\boldsymbol{W}^{\mathrm{out}} \tag{11}$$

$$V_2 = \boldsymbol{M}\boldsymbol{W}^{V} \tag{12}$$

其中,$\boldsymbol{W}$、V 和 $\boldsymbol{W}^{\mathrm{out}}$ 是参数,$\boldsymbol{W}^{V}$ 是逐点映射变换矩阵。

含有多个卷积核的卷积运算如公式(13)所示:

$$\mathrm{Conv}(\boldsymbol{M}) = \sum_{i=1}^{n}\frac{\exp(\alpha_i)}{\sum_{j=1}^{n}\exp(\alpha_j)}\mathrm{Conv}_{k_i}(\boldsymbol{M}) \tag{13}$$

为了学习单词级表示,卷积注意力在每一层连接一个自注意力网络和一个位置前馈网络。

$$\mathrm{Pointwise}(\boldsymbol{M}) = \max(0, \boldsymbol{M}\boldsymbol{W}_3 + b_3)\boldsymbol{W}_4 + b_4 \tag{14}$$

其中,$\boldsymbol{W}_3$、b_3、$\boldsymbol{W}_4$ 和 b_4 是映射参数。

2.3.3　自注意力机制

通过上下文信息表征 $\boldsymbol{H}^{p}$ 和通过卷积注意得到的问题感知表示 $\boldsymbol{H}^{q} \cdot C$ 的简单连接来表示从文章中提取的所有信息。

$$U^{p} = \mathrm{concat}(\boldsymbol{H}^{p}, \boldsymbol{H}^{q}C) \in \mathbb{R}^{4d\times n} \tag{15}$$

通常一篇文章可能包含数百个单词,很难完全捕获长距离依赖关系。于是采用一个独立的自注意力层进一步捕获文章中的远距离依赖关系。

$$\hat{\boldsymbol{U}}^{p} = \boldsymbol{U}^{p}\,\mathrm{drop}_{\mathrm{diag}}(f_{\mathrm{attention}}(\boldsymbol{U}^{p}, \boldsymbol{U}^{p})) \tag{16}$$

最后根据多注意力层收集到的所有信息,使用 BiLSTM 生成历史记忆,作为答案预测模块的输入。

$$\boldsymbol{M} = \mathrm{BiLSTM}([\boldsymbol{U}^{p}; \hat{\boldsymbol{U}}^{p}]) \tag{17}$$

其中,;表示向量/矩阵串联运算符。

2.4　答案输出层

利用记忆网络输出答案。将记忆网络的初始状态向量初始化为:

$$s_0 = \sum_j \alpha_j \boldsymbol{H}_j^{q} \tag{18}$$

其中,$\alpha_j = \dfrac{\exp(\omega_4 \cdot \boldsymbol{H}_j^{q})}{\sum_{j'}\exp(\omega_4 \cdot \boldsymbol{H}_{j'}^{q})}$,$\omega_4$ 为需要学习的参数。

在时间步 $\{0,1,\cdots,T-1\}$ 的范围内,第 t 步的状态定义为:

$$s_t = \mathrm{GRU}(s_{t-1}, x_t) \tag{19}$$

其中,$\boldsymbol{x}_t$ 由前一个状态 s_{t-1} 和历史记忆 $\boldsymbol{M}$ 计算而来。

$$x_t = \sum_j \beta_j \boldsymbol{M}_j \tag{20}$$

$$\beta_j = \mathrm{softmax}(s_{t-1}\boldsymbol{W}_5\boldsymbol{M}) \tag{21}$$

其中,$\boldsymbol{W}_5$ 为要学习的权重矩阵。

最后使用双线性函数来查找每个推理步骤 $t \in \{0,1,\cdots,T-1\}$ 的答案范围的起点和终点。

$$P_t^{\mathrm{begin}} = \mathrm{softmax}(s_t\boldsymbol{W}_6\boldsymbol{M}) \tag{22}$$

$$P_t^{\mathrm{end}} = \mathrm{softmax}([s_t; \sum_j P_{t,j}^{\mathrm{begin}}\boldsymbol{M}_j]\boldsymbol{W}_7\boldsymbol{M}) \tag{23}$$

其中,$\boldsymbol{W}_6$、$\boldsymbol{W}_7$ 为权重矩阵,;表示向量/矩阵串联运算符。

根据答案预测模块输出的一对起点和终点,可以从文章中提取答案片段。该模型通过利用所有 T 步输出的平均值作为最终的预测答案起始点,使得答案的输出不依赖于具体某一步起始点的产生。

$$P^{\mathrm{begin}} = \mathrm{avg}([P_0^{\mathrm{begin}}, P_1^{\mathrm{begin}}, \cdots, P_{T-1}^{\mathrm{begin}}]) \tag{24}$$

$$P^{\mathrm{end}} = \mathrm{avg}([P_0^{\mathrm{end}}, P_1^{\mathrm{end}}, \cdots, P_{T-1}^{\mathrm{end}}]) \tag{25}$$

为了防止各个模块之间信息的丢失以及模型过拟合的发生,在所有模块的最后一层添加一个随机丢弃层,丢弃率设置为 0.4,使模型不依赖于特定的步骤或模块来预测答案。

3 实 验

3.1 实验数据

实验数据采用斯坦福大学发布的 SQuAD 数据集,该数据集用于片段抽取型阅读理解任务,共包含 107.7 K 个(文章,问题,答案)三元组。其中 87.5 K 个问答对作为训练集,10.1 K 个问答对作为验证集,10.1 K 个问答对作为测试集。表 1 为 SQuAD 数据集的样例。

表 1 SQuAD 数据集样例展示

主题:American_Broadcasting_Company
文章:The American Broadcasting Company (ABC) (stylized in its logo as abc since 1957) is an American commercial broadcast television network that is owned by the Disney \ u2013ABC Television Group, a subsidiary of Disney Media Networks division of The Walt Disney Company. The network is part of the Big Three television networks. The network is headquartered on Columbus Avenue and West 66th Street in Manhattan, with additional major offices and production facilities in New York City, Los Angeles and Burbank, California.
问题 1: What company owns the American Broadcasting Company? 答案 1: The Walt Disney Company 问题 2: In what borough of New York City is ABC headquartered? 答案 2: Manhattan ……

3.2 实验设置

实验代码基于 Python 语言及其第三方库,深度学习环境采用 Pytorch 框架。并在具有两个 1 660 Super 的 GPU 上进行训练,单模型训练时占用显存约 12 GB,通常训练 10 个 epoch 至收敛,整个模型完成训练约需 12 个小时。模型训练过程中,使用的部分参数设置如表 2 所示。

表 2 参数设置

参数	值
embedding_dim	300
epoch	50
batch_size	16
Mogrifier 隐藏层	150
随机丢弃率	0.4
学习率	0.001

3.3 评价指标

将提出的模型在 SQuAD 数据集上进行评估。斯坦福大学官方给定了两个度量准则用以评估模型的性能。

(1)精确匹配(Exact Match,EM)。如果预测答案等于真实答案,EM 值为 1,否则为 0。

(2)F1 Score。预测答案和真实答案之间精确率(precision)和召回率(recall)的调和平均值。

即:

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

3.4 实验结果及分析

3.4.1 对比实验

为了评估模型的效果,将提出的模型与以下几个模型进行实验对比:

DCN+模型[24]采用双向 LSTM 对文章和问题进行编码,利用堆叠的自注意力机制和互注意力机制捕获结构特征。

R-Net 模型[14]使用双向 LSTM 作为编码器,并在注意力机制中加入门控机制来筛选出对回答问题相关性强的语义信息部分。

FusionNet[15]采用双向 LSTM 对文章和问题编码,融合多层注意力来理解文章和问题浅层和深层含义。

QANet[16]采用自注意力机制和深度可分离卷积对文章和问题进行编码,同时使用互注意力机制计算文章和问题的相似度来确定与回答最相关的信息。

SAN 模型[25]为选取的基线模型,采用双向 LSTM 作为编码器,并使用互注意力机制和自注意力机制捕获文章的结构,通过记忆网络来预测答案的起止位置。

选择的对比模型与提出模型的结构及采用的注意力机制对比如表 3、表 4 所示。

表 3 模型结构对比

模型	RNN	CNN	注意力
DCN+	√		√
R-Net	√		√
QANet		√	√
FusionNet	√		√
SAN	√		√
DCAM	√	√	√

表 4 各模型结构中的注意力机制比较

模型	双向注意力	自注意力	类型
DCN+	2	2	并行
R-Net	1	1	串行
QANet	1	1	串行
FusionNet	3	1	并行
SAN	1	1	并行
DCAM	1	2	并行

以上可以看出所提模型与其他模型的不同之处,即所提模型在不影响训练速度的情况下,将 RNN、CNN 和注意力机制融为一体,特别是采用了动态卷积代替普通卷积,并与注意力机制结合,从多层次和多角度提取文本特征,这样可以更好地利用它们的优势,提

高文本之间的交互程度。

对比模型与 DCAM 的实验结果如表 5 所示。

表 5　对比实验结果　　%

模型	EM	F1
DCN+	74.50	83.10
R-Net	72.30	80.60
QANet	73.60	82.70
FusionNet	76.00	83.90
SAN	75.93	83.74
DCAM	76.74	84.30

从表 5 可以看出,所提模型由于在编码层采用的结构增强了文章和问题在低层语义表示的交互;此外,添加的卷积结构也加深了机器对文章局部结构的理解,弥补注意力机制只能捕获全局结构的不足。在 SQuAD 数据集上的 EM 值和 F1 值分别达到了 76.74%、84.30%,相比基线模型 SAN 其 EM 值和 F1 值分别提高了 0.81 百分点和 0.56 百分点。同时在 SQuAD 数据集上的表现也均优于其他对比模型,这得益于 DCAM 将三种结构的优势相结合。实验结果表明该模型在阅读理解任务上的有效性。

3.4.2　消融实验

DCAM 模型是在 SAN 上进行的改进。为了验证改进模块对模型性能的影响,设计消融实验比较改进模块之后模型的 EM 值和 F1 值大小。其实验对比结果如表 6 所示。

表 6　消融实验结果　　%

模型	EM	F1
SAN	75.93	83.74
SAN+DCA	76.34	84.00
SAN+Mogrifier	76.41	84.15
SAN+Mogrifier + DCA	76.74	84.30

从表 6 可以看出,使用结合卷积的注意力机制在 EM 值和 F1 值上分别提升了 0.41 百分点和 0.26 百分点,采用改进的 LSTM 作为编码器在 EM 值和 F1 值上分别提升了 0.48 百分点和 0.41 百分点,同时改进则获得 0.81 百分点的 EM 值提升和 0.56 百分点的 F1 值提升。结果表明改进的两个模块均能够加深模型对文章和问题的理解,提高模型回答问题的准确率。

3.4.3　卷积注意力大小对模型性能的影响

一般来说,卷积核越小,所需的参数量和计算量越小。卷积核越大,其感受野越大,相应的参数量和计算量也越大。但多层小卷积核堆叠不仅可以减少计算量,还能达到大卷积核一样的感受野。这里采用实验里常用的卷积核大小 1、3、5 来探索不同卷积核大小对模型性能的影响,结果如表 7 所示。

表 7　卷积核大小　　%

卷积核大小	EM	F1
3	76.35	84.20
5	75.96	83.80
3、5	76.44	84.10
1、3、5	76.74	84.30

从表 7 的结果可以看出,卷积核越大,模型的 EM 值和 F1 值越低;卷积核越多,模型的 EM 值和 F1 值越高。当只采用单个卷积核时,卷积核大小为 3 的卷积比卷积核大小为 5 的卷积对模型的 EM 值和 F1 值提升较大,分别高出 0.39 百分点和 0.40 百分点;当采用多个卷积核组合时,不同的卷积组合对模型性能均有提升,但 1、3、5 的卷积组合对模型的 EM 值和 F1 值提升更高,相较于 3、5 的卷积组合分别提升了 0.30 百分点和 0.20 百分点。因此,在实验过程中选择卷积核大小为 1、3、5 的组合。

4　结束语

为解决语义表示能力差、信息冗余、信息丢失等问题,提出了一种融合动态卷积注意力的机器阅读理解模型。该模型利用 Mogrifier 加强相关文本之间的特征表示,借助注意力机制捕获文章和问题中的相关信息,结合动态卷积注意力进一步捕获文章的局部和全局结构。在 SQuAD 阅读理解数据集上进行了实验验证,结果表明引入动态卷积和多注意力机制的模型能够有效提高机器阅读理解的准确性,具有一定的应用价值。在未来的研究工作中,可以考虑与大规模的预训练模型(如 BERT、RoBERTa 等)相结合,进一步提升机器阅读理解模型的性能。

参考文献:

[1] 刘高军,李亚欣,段建勇. 基于混合注意力机制的中文机器阅读理解[J]. 计算机工程,2022,48(10):67-72.

[2] 包　玥,李艳玲,林　民. 抽取式机器阅读理解研究综述[J]. 计算机工程与应用,2021,57(12):25-36.

[3] HIRSCHMAN L,LIGHT M,BRECK E,et al. Deep read:a reading comprehension system[C]//Proceedings of the 37th annual meeting of the association for computational linguistics. Maryland:Association for Computational Linguistics,1999:325-332.

[4] RICHARDSON M,BURGES C J C,RENSHAW E. Mctest:a challenge dataset for the open-domain machine comprehension of text[C]//Proceedings of the 2013 conference on empirical methods in natural language processing. Washington:Association for Computational Linguistics,2013:193-203.

[5] RAJPURKAR P,ZHANG J,LOPYREV K,et al. SQuAD:

100, 000 + questions for machine comprehension of text [C]//Proceedings of the 2016 conference on empirical methods in natural language processing. Austin: Association for Computational Linguistics, 2016: 2383-2392.

[6] LAI G, XIE Q, LIU H, et al. RACE: large-scale reading comprehension dataset from examinations[C]//Proceedings of the 2017 conference on empirical methods in natural language processing. Copenhagen: Association for Computational Linguistics, 2017: 785-794.

[7] 唐竑轩,武恺莉,朱朦朦,等. 基于双向注意力机制的多文档神经阅读理解[J]. 计算机工程,2020,46(12):43-51.

[8] LIU S, ZHANG X, ZHANG S, et al. Neural machine reading comprehension: methods and trends[J]. Applied Sciences, 2019, 9(18): 3698.

[9] HERMANN K M, KOCISKY T, GREFENSTETTE E, et al. Teaching machines to read and comprehend[J]. Advances in Neural Information Processing Systems, 2015, 28: 1693-1701.

[10] KADLEC R, SCHMID M, BAJGAR O, et al. Text understanding with the attention sum reader network[C]//Proceedings of the 54th annual meeting of the association for computational linguistics. Berlin: Association for Computational Linguistics, 2016: 908-918.

[11] CHEN D, BOLTON J, MANNING C D. A thorough examination of the cnn/daily mail reading comprehension task[C]//Proceedings of the 54th annual meeting of the association for computational linguistics. Berlin: Association for Computational Linguistics, 2016: 2358-2367.

[12] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension[EB/OL]. (2016-11-05)[2022-10-23]. https://arxiv.org/abs/1611.01603.

[13] CHEN D, FISCH A, WESTON J, et al. Reading wikipedia to answer open-domain questions[C]//Proceedings of the 55th annual meeting of the association for computational linguistics. Vancouver: Association for Computational Linguistics, 2017: 1870-1879.

[14] WANG W, YANG N, WEI F, et al. Gated self-matching networks for reading comprehension and question answering[C]//Proceedings of the 55th annual meeting of the association for computational linguistics. Vancouver: Association for Computational Linguistics, 2017: 189-198.

[15] HUANG H Y, ZHU C, SHEN Y, et al. FusionNet: fusing via fully-aware attention with application to machine comprehension[EB/OL]. (2017-11-16)[2022-10-23]. https://arxiv.org/abs/1711.073 41.

[16] YU A W, DOHAN D, LUONG M T, et al. Qanet: combining local convolution with global self-attention for reading comprehension[J]. arXiv: 1804.09541, 2018.

[17] PENNINGTON J, SOCHER R, MANNING C D. Glove: global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing. Doha: Association for Computational Linguistics, 2014: 1532-1543.

[18] MELIS G, KOCISKY T, BLUNSOM P. Mogrifier LSTM[J]. arXiv: 1909.01792, 2019.

[19] MCCANN B, BRADBURY J, XIONG C, et al. Learned in translation: contextualized word vectors[C]//Proceedings of the 31st international conference on neural information processing systems. California: Curran Associates Inc., 2017: 6297-6308.

[20] GOODFELLOW I, WARDE-FARLEY D, MIRZA M, et al. Maxout networks[C]//Proceedings of the 30th international conference on machine learning. Georgia: JMLR. org, 2013: 1319-1327.

[21] ZHAO G, SUN X, XU J, et al. Muse: parallel multi-scale attention for sequence to sequence learning[J]. arXiv: 1911.09483, 2019.

[22] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE, 2017: 1251-1258.

[23] WU F, FAN A, BAEVSKI A, et al. Pay less attention with lightweight and dynamic convolutions[C]//International conference on learning representations. Vancouver: Open-Review. net, 2018: 1-14.

[24] XIONG C, ZHONG V, SOCHER R. Dynamic coattention networks for question answering[C]//Proceedings of the 5th international conference on learning representations. Ithaca: arXiv. org, 2017: 1-14.

[25] LIU X, SHEN Y, DUH K, et al. Stochastic answer networks for machine reading comprehension[C]//Proceedings of the 56th annual meeting of the association for computational linguistics. Melbourne: Association for Computational Linguistics, 2018: 1694-1704.