浙江理工大学学报，2021，45(3)：400-407
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2021.03.014

基于自注意力和门控机制的答案选择模型

陈巧红，李妃玉，贾宇波，孙　麒
（浙江理工大学信息学院，杭州 310018）

摘　要：针对现有答案选择方法语义特征提取不充分和准确性差的问题，引入自注意力和门控机制，提出了一种答案选择模型。该模型首先在问题和答案文本内部利用层叠自注意力进行向量表示，并在自注意力模块中让单词和位置分开进行多头注意力；然后将答案句通过卷积神经网络(Convolutional neural network，CNN)得到的向量表示输入注意力层，根据问题生成与问题相关的答案表示，并通过门控机制融合两种表示；最后计算问题和答案文本的相关性分数，得到候选答案的排名和标注。结果表明：该模型与双向长短时记忆网络模型、自注意力模型和基于注意力的双向长短时记忆网络模型相比，在WebMedQA数据集上平均倒数排名分数分别提高了8.37%、4.79%和2.03%，预测答案正确率也有提高。这表明提出的模型能够捕获更丰富的语义信息，有效提升了答案选择的性能。

关键词：答案选择；层叠自注意力；注意力机制；门控机制
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1673-3851 (2021) 05-0400-08

Answer selection model based on self-attention and gating mechanism

CHEN Qiaohong, *LI Feiyu*, *JIA Yubo*, *SUN Qi*
(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: In view of the problem of insufficient semantic feature extraction and poor accuracy of existing answer selection methods, self-attention and gating mechanism are introduced to propose an answer selection model. In the question and answer texts, this model firstly adopted cascading self-attention for vector representation, and then separated the words and positions in the self-attention module for multi-head attention. After that, the answer sentences were input into the attention layer, through the vector representations obtained through the convolutional neural network (CNN). The answer representation related to the question was generated based on the question, and the two representations were merged through gating mechanism. Finally, the correlation score between question and answer texts was calculated to obtain the ranking and labeling of the candidate answer. The experimental results show that compared with the bidirectional long short-term memory (BiLSTM) model, the self-attention model and the attention-based BiLSTM model, the mean reciprocal rank scores on the WebMedQA dataset grew by 8.37%, 4.79% and 2.03% respectively, and the predicted accuracy of answers was also improved, suggesting that the proposed model can capture more abundant semantic information and effectively improve the performance of answer selection.

Key words: answer selection; cascading self-attention; attention mechanism; gating mechanism

收稿日期：2021－02－01　　网络出版日期：2021－03－30
基金项目：浙江理工大学中青年骨干人才培养经费项目
作者简介：陈巧红(1978－　)，女，浙江临海人，副教授，博士，主要从事计算机辅助设计及机器学习技术方面的研究。

0 引　言

随着问答系统和社区问答论坛的发展，用户对快速、准确地获取信息的需求明显增加，答案选择逐渐成为开发智能问答系统的关键。传统的问答系统，包括问题处理模块和答案处理模块，侧重点在于语料中词汇的语义特征提取。但是语料中不只包含词汇的语义信息，还隐含了单词的位置信息以及问答对之间的语义关系。答案选择通过提取问答对的深层特征来反映问题与候选答案之间的匹配关系，是开发更加高效的问答系统的基础，具有重要的研究意义[1]。

经典的答案选择方法局限于问题和答案文本的浅层信息，通过手工构建一些浅层特征来对候选答案进行打分。这样仅考虑词级特征，会忽略结构信息，如忽略词序、句法信息和文本间的关系，从而直接影响问题与正确答案的匹配度。为此研究人员探索了各种答案选择方法，并取得了一定的研究成果。Heliman 等[2]提出了一种使用 33 种句法特征对问答对进行分类的逻辑回归模型，来解决答案选择中的语义匹配问题。Chan 等[3]将问题和答案文本通过 word2vec[4]转化为特征后，分别通过一个多层 CNN 得到长度固定的输入，然后再输入到长短时记忆网络(Long short-term memory，LSTM)中，从而利用上下文信息实现答案选择，解决了特征工程无法精确匹配正确答案的问题。Zhou 等[5]针对文本序列的长期依赖性特点，提出了一种对答案句采用 LSTM 的模型，结果发现 LSTM 相比 CNN 可以更快地收敛。Wakchaure 等[6]使用 CNN-LSTM 充当编码器，通过条件随机场从多个问答对中预测最有用的一对作为最终的输出。各种基于 CNN 和 LSTM 的神经网络模型都表现出针对答案选择任务的高性能，这些深度学习方法通过各种类型的深度神经网络从低级表示中提取重要特征，摆脱了仅关注浅层的单词级特征的限制。Tan 等[7]提出了一种基于注意力的 LSTM 模型用于答案选择，该模型使用 BiLSTM 对答案文本进行编码，并采用注意力机制，根据问题生成答案的表示，量化关注问题对答案的影响，解决了无法充分提取问答对的语义关系的问题。熊雪等[8]提出了一种基于层叠注意力机制的答案选择模型，采用 BiLSTM 与词匹配方法对问题和答案文本进行特征提取，并融合句内注意力机制，同时考虑句子内部的结构和句子之间的逻辑关系，其匹配度优于基础神经网络。Shao 等[9]提出了一种基于 Transformer 的神经网络模型，只采用自注意力机制学习词与词之间的长程依赖关系，提取全局特征，使问题答案的表示包含的信息更加丰富，得到了较高的准确率。谢正文等[10]提出了一种构建词级交互矩阵的方法，通过问题文本的信息获取答案文本重要的部分，更精确地完成问题和答案之间的语义匹配。

分析上述答案选择方法发现：由于语料库通常存在问题短答案长的问题，很难找到两个句子的结构化表示形式之间的最佳匹配，利用特征工程匹配答案的准确度仍然较低。基于神经网络的答案选择，能获取更精确的语义匹配关系，但不能很好地捕获词与词之间的相互关系以及词的位置信息，而且基于神经网络的模型分别对问题和答案文本进行编码。当代表答案时，会忽略问题所涉及的信息；反之亦然。注意力机制着重考虑上下文信息和句子中不同词之间的相互关系，通过分配不同的权重来增强特征，捕获了词与词之间的相互关系，但是忽略了词自身的特征。采用自注意力机制[11-12]学习词自身的特征以及问答对之间的复杂语义关系，能够更充分利用地特征之间的关系。但是仅采用自注意力机制提取全局信息，缺少局部信息和位置信息，无法学习句子的序列关系。如果词的位置信息缺失和词自身的特征未得到充分挖掘，那么答案选择的性能就会受到一定的限制。

针对在答案选择任务中缺少词的位置信息以及词自身的特征没能得到充分挖掘的问题，本文提出了一种基于层叠自注意力和门控机制的答案选择模型。该模型利用自注意力作为基础模型，在问题和答案文本交互前降低噪声词的影响并提取词的顺序特征，对答案文本采用基于注意力的卷积神经网络来获取包含问题信息的答案表示，并通过门控机制融合答案的全局信息表示和包含问题语义信息的局部特征表示。由于该模型融合了整体和局部的特征，增强了模型的表达能力，问题和正确答案的匹配度将得到提升。

1 本文模型

本文将答案选择任务视为一个有效的句子匹配和序列标注问题，对于给定的问题，计算问题与候选答案间的相关性分数，并标注候选答案是否为标准答案，最终考量标准答案的排名值。

本文提出的基于自注意力和门控机制的答案选择模型，共包含两个输入层、两个自注意力层、一个

卷积层、一个注意力层和一个门控层。首先通过自注意力层获得问题和答案的句向量表示;其次通过卷积注意力获得与问题交互后的答案的句向量表示,并通过门控机制将两部分答案表示相结合;最后通过余弦相似度函数计算问题与候选答案的匹配度,将得到的相似度分数输入到 Softmax 层中。整体模型如图 1 所示,其中:$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_L]$为输出词嵌入矩阵,$\boldsymbol{x}_i$ 为第 i 个词的词向量表示,L 为输入的问题和答案文本长度;$\boldsymbol{H}_{\mathrm{q}}[\boldsymbol{h}_{1,\mathrm{q}},\boldsymbol{h}_{2,\mathrm{q}},\cdots,\boldsymbol{h}_{L,\mathrm{q}}]$、$\boldsymbol{H}_{\mathrm{a}}=[\boldsymbol{h}_{1,\mathrm{a}},\boldsymbol{h}_{2,\mathrm{a}},\cdots,\boldsymbol{h}_{L,\mathrm{a}}]$分别为问题和答案经过编码器得到的输出向量组成的矩阵,$\boldsymbol{h}_{i,\mathrm{q}}$、$\boldsymbol{h}_{i,\mathrm{a}}$分别为问题和答案中第 i 个词和所有词的相关信息;$\boldsymbol{v}_{\mathrm{q}}$、$\boldsymbol{v}_{\mathrm{a}}$ 分别为问题和答案的最终向量表示;Sim_i $(i=1,2,\cdots,5)$为问题与其 5 个候选答案的相似性。

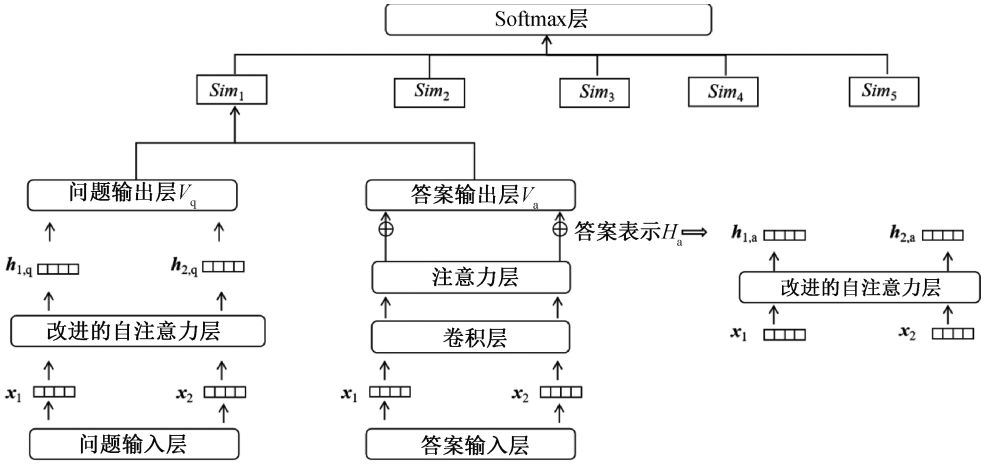

图 1 基于自注意力和门控机制的答案选择模型

1.1 问题答案句向量表示

本文首先采用自注意力获取句子的向量表示。自注意力是 Transformer 的一个核心模块,通过学习输入序列自身的特征信息来提取全局信息,从而产生更好的性能[13]。自注意力作用在同一个句子内部,即作用在问题句或答案句自身。对于自注意力来说,在答案任务中的询问(Query, $\boldsymbol{Q}$)是句子中的一个词,而键(Key, $\boldsymbol{K}$)和值(Value, $\boldsymbol{V}$)是句子中所有的词以及词的表示。自注意力机制重点在于学习输入序列自身的内部信息,每个词可以观察到序列中其他词的信息,并通过注意力交互使其他词产生不同大小的权重,最终自注意力层的输出将涵盖序列所有词的语义信息,从而实现双向编码上下文的效果。同时,自注意力机制使得序列中任何两个词之间的距离为 1,因而可以同时观测序列中所有位置的词,解决了循环神经网络所存在的长时序依赖问题。

改进的自注意力模型包括多头注意力、前馈神经网络、残差连接以及层归一化,其结构如图 2 所示。本文将位置相关性和单词相关性分别在多头注意力模块进行计算,通过解除单词和位置间的关联,帮助模型学习整个句子的准确表示,并达到加快预训练损失收敛的效果。其中:$\boldsymbol{P}=[\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_L]$为位置嵌入矩阵,$\boldsymbol{p}_i$ 为第 i 个位置向量表示;$\boldsymbol{C}=[\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_L]$、$\boldsymbol{C}^*=[\boldsymbol{c}_1^*,\boldsymbol{c}_2^*,\cdots,\boldsymbol{c}_L^*]$为多头注意力的中间输出;$\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_L]$为通过残差及层归一化得到的每个词的上下文信息和位置信息的表示;$\boldsymbol{R}=[\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_L]$为通过前向反馈神经网络计算得到的残差;$\boldsymbol{C}$、$\boldsymbol{C}^*$、$\boldsymbol{R}$、$\boldsymbol{Y}$ 维度均为 $D\times L$,D 为自注意力机制的内部表示维度且值与词向量的维度相同。

在答案选择任务中,自注意力层的输入是问题或答案文本的词序列的矩阵表示,以问题为例介绍本文应用自注意力的计算过程。首先仅使用词嵌入作为输入,即将原始词序列转换为矩阵表示形式,得到词序列的矩阵表示 $\boldsymbol{X}$;然后加入位置嵌入 $\boldsymbol{P}$。对于词嵌入,将本文使用的数据集分词后作为语料,经过 GloVe 词转向量工具[14]预训练得到。对于位置嵌入,本文采用固定的位置编码,其位置编码 $\boldsymbol{p}_i$ 的第 j 维可用式(1)所示:

$$\boldsymbol{P}(i,j)=\begin{cases}\sin(i/10000^{2j/E}), & j \text{ 为偶数},\\ \cos(i/10000^{2j/E}), & j \text{ 为奇数}\end{cases} \tag{1}$$

其中:E 为词向量的维度。

自注意力层的输出 $\boldsymbol{H}$ 是输入句中各个词的上下文表示,计算过程如式(2)—(5)所示:

$$\boldsymbol{C}=f_{\mathrm{M}}(\boldsymbol{X})+f_{\mathrm{M}}(\boldsymbol{P}) \tag{2}$$

$$\boldsymbol{Y}=f_{\mathrm{L}}(\boldsymbol{X}+\boldsymbol{C}) \tag{3}$$

$$\boldsymbol{R}=f_{\mathrm{F}}(\boldsymbol{Y}) \tag{4}$$

$$\boldsymbol{H}=f_{\mathrm{L}}(\boldsymbol{Y}+\boldsymbol{R}) \tag{5}$$

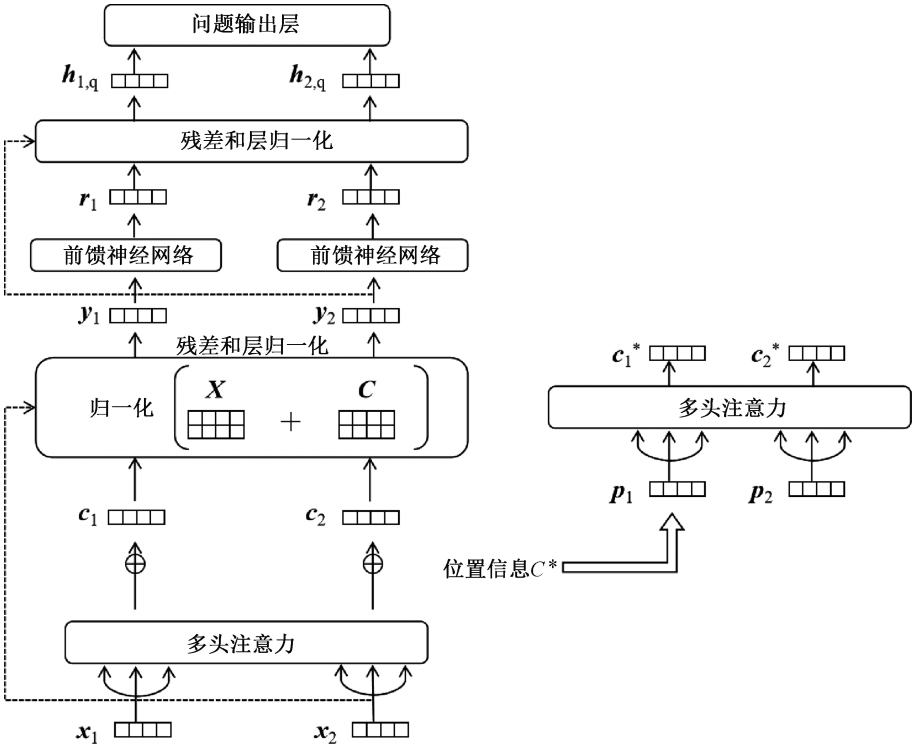

图 2　改进自注意力模型结构

其中：f_M、f_L、f_F 分别为计算多头注意力、层归一化和前馈网络的函数。

自注意力模型结构核心是多头注意力，其结构如图 3 所示。本文将计算得到的信息映射回原始空间，通过多头注意力机制让每个词同时在不同的子空间获取其他词的信息，增强了注意力机制的性能。

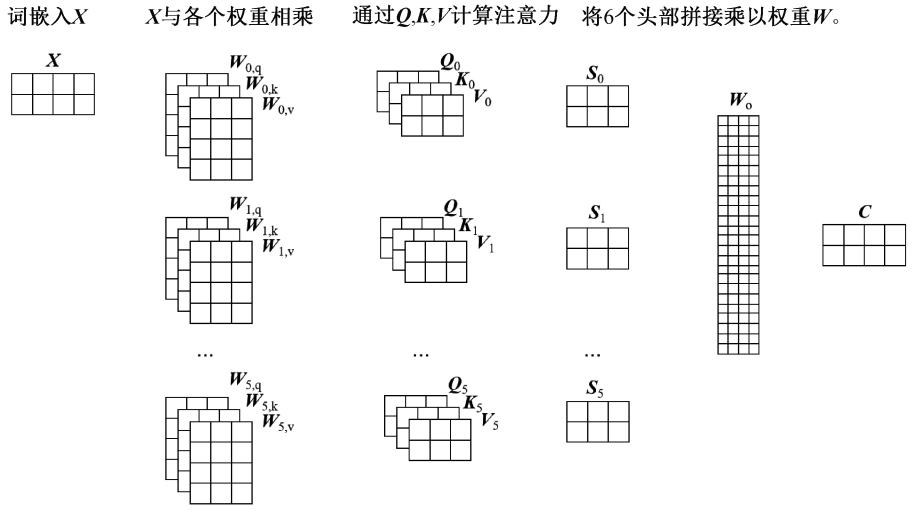

图 3　多头注意力结构

本文将注意力机制运作的过程，看作通过句子中的一个词 $\boldsymbol{Q}$，匹配句子中所有的词 $\boldsymbol{K}$，并得到词的表示 $\boldsymbol{V}$ 的过程。在进行多头注意力的过程中，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 会被切割成 n 个部分，具体表示为 $\{\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i \mid i=1,2,\cdots,n\}$，其中 D 能被 n 整除。单词相关性的 $\boldsymbol{Q}_i$、$\boldsymbol{K}_i$、$\boldsymbol{V}_i$ 通过三个参数矩阵 $\boldsymbol{W}_{i,q}$、$\boldsymbol{W}_{i,k}$、$\boldsymbol{W}_{i,v}$ 变换得到，位置相关性的 $\boldsymbol{Q}_i^*$、$\boldsymbol{K}_i^*$、$\boldsymbol{V}_i^*$ 通过三个参数矩阵 $\boldsymbol{U}_{i,q}$、$\boldsymbol{U}_{i,k}$、$\boldsymbol{U}_{i,v}$ 变换得到。位置相关性和单词相关性分别在每个头部采用自注意力机制，然后将每个头部得到的结果串联并通过 $\boldsymbol{W}_o$ 映射，得到每个词包含上下文信息和位置信息的表示。其计算过程如式(6)—(10)所示：

$$[\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i] = [\boldsymbol{W}_{i,q}, \boldsymbol{W}_{i,k}, \boldsymbol{W}_{i,v}] \cdot \boldsymbol{X} \tag{6}$$

$$[\boldsymbol{Q}_i^*, \boldsymbol{K}_i^*, \boldsymbol{V}_i^*] = [\boldsymbol{U}_{i,q}, \boldsymbol{U}_{i,k}, \boldsymbol{U}_{i,v}] \cdot \boldsymbol{P} \tag{7}$$

$$\boldsymbol{S}_i = f_{\mathrm{A}}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i) + f_{\mathrm{A}}(\boldsymbol{Q}_i^*, \boldsymbol{K}_i^*, \boldsymbol{V}_i^*) \quad (8)$$

$$f_{\mathrm{A}}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V} \quad (9)$$

$$\boldsymbol{C} = \boldsymbol{W}_{\mathrm{o}} \cdot \mathrm{Concat}(\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_n) \quad (10)$$

其中:d_{k} 为矩阵 $\boldsymbol{K}$ 中每个列向量的维度,$\boldsymbol{S}_i$ 为第 i 个头部的输出,$\boldsymbol{W}_{\mathrm{o}}$ 为参数矩阵,f_{A} 为计算权重系数的函数。

在自注意力模型的每一层均采用两个残差连接,分别为多头注意力机制子层外部以及全连接前馈网络子层外部,它们都模拟了残差网络的结构但没有共享参数,并且跟随着层归一化(LayerNorm)处理。层归一化在计算每个词之间是独立的,即只在 D 这一维上起作用。假设某个词的中间表示为 $\boldsymbol{Z} = [z_1, z_2, \cdots, z_D]$,其中 z_i 是 $\boldsymbol{Z}$ 的第 i 维,LayerNorm 的计算过程如式(11)—(13)所示:

$$\mu = \frac{1}{D}\sum_i^D z_i \quad (11)$$

$$\sigma = \sqrt{\frac{1}{D}\sum_i^D (z_i - \mu)^2} \quad (12)$$

$$f_{\mathrm{L}}(\boldsymbol{Z}) = \boldsymbol{\gamma} \odot \frac{(\boldsymbol{Z} - \mu)}{\sigma} + \boldsymbol{\beta} \quad (13)$$

其中:μ、σ 分别为均值和方差,$\boldsymbol{\gamma}$、$\boldsymbol{\beta}$ 为可学习的参数向量,$\odot$ 为对应位置元素相乘。

前馈神经网络(FeedForwardLayer)是使用 ReLu 激活函数的两层全连接网络,其计算过程如式(14)所示:

$$f_{\mathrm{F}}(\boldsymbol{Y}) = \boldsymbol{W}_{\mathrm{c}} \cdot \max(0, \boldsymbol{W}_{\mathrm{b}} \cdot \boldsymbol{Y} + \boldsymbol{b}_{\mathrm{b}}) + \boldsymbol{b}_{\mathrm{c}} \quad (14)$$

其中:$\boldsymbol{W}_{\mathrm{c}}$、$\boldsymbol{W}_{\mathrm{b}}$、$\boldsymbol{b}_{\mathrm{b}}$、$\boldsymbol{b}_{\mathrm{c}}$ 为网络参数,维度分别为 $D \times D_f$、$D_f \times D$、D_f、D,D_f 为全连接网络的内部表示维度。

1.2 答案的注意表示

答案中可能包含许多无用的信息,故采用注意力机制去学习答案的句子向量,根据问题内容在答案中寻找相匹配的信息,当匹配到相关信息后,则对这些词向量赋予较大的权重,使其包含更多与问题相关的信息量。

经过自注意力后得到问题的表示 $\boldsymbol{H}_{\mathrm{q}} = [\boldsymbol{h}_{1,\mathrm{q}}, \boldsymbol{h}_{2,\mathrm{q}}, \cdots, \boldsymbol{h}_{L,\mathrm{q}}]$,其中 $\boldsymbol{h}_{1,\mathrm{q}}$ 包含问题中所有信息的汇总,将其作为最终的问题句向量表示 $\boldsymbol{v}_{\mathrm{q}} = \boldsymbol{h}_{1,\mathrm{q}}$。对初始答案向量 $\boldsymbol{X}_{\mathrm{a}} = [\boldsymbol{x}_{1,\mathrm{a}}, \boldsymbol{x}_{2,\mathrm{a}}, \cdots, \boldsymbol{x}_{L,\mathrm{a}}]$ 卷积后得到词及周围词信息的局部特征表示 $\boldsymbol{v}_{\mathrm{ac}}$,通过计算得到一个 Softmax 权值,该权重相当于各个词与询问词的相关性,与 $\boldsymbol{v}_{\mathrm{ac}}$ 相乘后得到带问题信息权重的答案注意表示。其计算过程如式(15)—(17)所示:

$$\boldsymbol{v}_{\mathrm{ac}} = \tanh(\boldsymbol{W}_{\mathrm{f}}\boldsymbol{x}_{i,\mathrm{a}} + \boldsymbol{B}_{\mathrm{f}}) \quad (15)$$

$$\boldsymbol{v}_{\mathrm{at}} = f_{\mathrm{A}}(\boldsymbol{v}_{\mathrm{ac}}, \boldsymbol{v}_{\mathrm{q}}) = \sum_i \alpha_i \cdot \boldsymbol{v}_{i,\mathrm{ac}} \quad (16)$$

$$\alpha_i \propto \exp(\boldsymbol{m}^{\mathrm{T}} \cdot \tanh(\boldsymbol{W}_{\mathrm{at}} \cdot \boldsymbol{v}_{\mathrm{at}} + \boldsymbol{W}_{\mathrm{q}} \cdot \boldsymbol{v}_{\mathrm{q}})) \quad (17)$$

其中:$\boldsymbol{W}_{\mathrm{f}}$ 为卷积核的参数,$\boldsymbol{B}_{\mathrm{f}}$ 为偏置参数,$\boldsymbol{v}_{\mathrm{ac}}$ 为经过一次卷积得到的答案的特征表示,$\boldsymbol{v}_{\mathrm{at}}$ 为经过注意力得到的答案的注意表示,α_i 为通过 softmax 函数计算得到的答案中第 i 个词的注意力权重,$\boldsymbol{W}_{\mathrm{at}}$、$\boldsymbol{W}_{\mathrm{q}}$ 为可学习的权重参数,$\boldsymbol{m}$ 为注意力参数。

1.3 门控融合两种表示

经过自注意力后得到答案的表示 $\boldsymbol{H}_{\mathrm{a}} = [\boldsymbol{h}_{1,\mathrm{a}}, \boldsymbol{h}_{2,\mathrm{a}}, \cdots, \boldsymbol{h}_{L,\mathrm{a}}]$,其中 $\boldsymbol{h}_{1,\mathrm{a}}$ 包含答案中所有信息的汇总,将其作为经过自注意力得到的包含上下文信息和位置信息的输出,即答案的第一部分向量表示 $\boldsymbol{h}_{1,\mathrm{a}}$;经过卷积注意力后,得到包含问题信息的输出,即答案的第二部分向量表示 $\boldsymbol{v}_{\mathrm{at}}$。本文设计一个门控机制,当 g_i 开放时,即 g_i 值为 1 时,当前词的表示和上下文信息很相关;当 g_i 关闭时,即 g_i 值为 0 时,当前词的表示和上下文信息无关,此时 g_i 可以将一些与上下文无关的信息过滤,得到最终的答案句向量表示 $\boldsymbol{v}_{\mathrm{a}}$。其计算过程如式(18)—(20)所示:

$$\overline{\boldsymbol{h}_{1,\mathrm{a}}} = \frac{\sum_i^L \boldsymbol{h}_{1i,\mathrm{a}}}{L} \quad (18)$$

$$g_i = \mathrm{Sigmoid}(\boldsymbol{W}_{\mathrm{e}} \cdot (\boldsymbol{h}_{1i,\mathrm{a}} \odot \overline{\boldsymbol{h}_{1,\mathrm{a}}}) + \boldsymbol{b}_{\mathrm{e}}) \quad (19)$$

$$\boldsymbol{v}_{\mathrm{a}} = (1 - \boldsymbol{G}) \odot \boldsymbol{v}_{\mathrm{at}} + \boldsymbol{G} \odot \boldsymbol{h}_{1,\mathrm{a}} \quad (20)$$

其中:$\boldsymbol{h}_{1i,\mathrm{a}}$ 为答案第 i 个词的具有上下文信息的向量表示,$\overline{\mathbf{h}_{1,a}}$ 为答案句中所有词表示的平均,$\boldsymbol{W}_{\mathrm{e}}$ 为可学习的权重参数,$\boldsymbol{b}_{\mathrm{e}}$ 为偏移项,g_i 为 $\boldsymbol{G}$ 的第 i 列,$\boldsymbol{G}$ 为作用在输入 $\boldsymbol{h}_{1,\mathrm{a}}$ 上的全局信息门。门控的作用是通过 $\boldsymbol{G}$ 调节 $\boldsymbol{v}_{\mathrm{at}}$ 与 $\boldsymbol{h}_{1,\mathrm{a}}$ 的贡献比例得到融合特征 $\boldsymbol{v}_{\mathrm{a}}$。

1.4 计算相关性

本文采用余弦相似度来计算问题与候选答案之间的相关性。对给定的问题答案对,其匹配度的计算如式(21)所示:

$$Sim = \cos(\theta) = \frac{\boldsymbol{v}_{\mathrm{q}} \cdot \boldsymbol{v}_{\mathrm{a}}}{\|\boldsymbol{v}_{\mathrm{q}}\| \ \|\boldsymbol{v}_{\mathrm{a}}\|} = \frac{\sum_{i=1}^{L} v_{i,\mathrm{q}} \times v_{i,\mathrm{a}}}{\sqrt{\sum_{i=1}^{L} (v_{i,\mathrm{q}})^2} \times \sqrt{\sum_{i=1}^{L} (v_{i,\mathrm{a}})^2}} \quad (21)$$

其中:$\boldsymbol{v}_{\mathrm{q}}$、$\boldsymbol{v}_{\mathrm{a}}$ 分别为问题和答案的最终向量表示,$v_{i,\mathrm{q}}$,$v_{i,\mathrm{a}}$ 为问题和答案的各分量。

2　实验设置与结果分析

2.1　数据集

本文实验使用的语料是 2018 年 He 等[15]发布的大规模的综合中文医学问答(webMedQA)数据集。

webMedQA 是从与健康有关的专业咨询网站(百度医生或 120Ask)收集的相关问答。webMedQA 中每行文档都包含问题所属类别、问题编号、问题和若干候选答案。对数据集中采用答案的问题进行过滤,得到 65941 个问题。清理所有的 web 标签、链接和乱码字节,删除标准答案过长和有多个标准答案的问题,获得了 63284 个有效问题,针对问题抽取 4 个否定答案,对于否定答案少于 4 个的问题,从其他问题中抽取答案进行补充,对问题答案对进行分词和统计词表等操作,最终得到 316420 组问题答案对。然后,按照每个问题类别中 8∶1∶1 的比例将数据集分为训练集、开发集和测试集。本文使用的分词工具是 jieba 分词[16],词转向量工具是 GloVe。语料库中的一个问题及其候选答案与标注的示例见表 1。

表 1　语料示例

问题	答案	标注
喉咙老是有痰还有血丝最近喉咙老是有痰,今天还痰里血丝,是怎么回事?	您的情况可能就是咽喉炎引起的症状,与上火也是有关系的。	1
	根据你的描述,出现食欲不振,胃胀,局部疼痛,考虑是胆囊炎的症状。	0
	你好考虑是咽喉炎,上呼吸道感染有炎症引起的有痰等症状。	0
	您好,您的情况初步考虑是咽部炎症导致的多痰和出血。	0
	您好,你描述的情况应该是咽喉炎造成的,与上火有关,需要及时合理的,建议到医院拍个胸片。	0

2.2　评价指标

本文使用 a_{top1} 和 *MRR* 作为评价指标。a_{top1} 是预测结果正确的概率,计算相关性分数最大的一组问答对的答案作为标准答案,如结果是标准答案,即预测正确。*MRR* 是平均倒数排名,通过检索正确答案在候选答案中的排名来评估答案选择模型的性能。计算 a_{top1} 和 *MRR* 的数学公式定义为式(22)—(23)所示:

$$a_{\text{top1}}/\% = \frac{1}{|N|}\sum_{i}^{|N|} label_i \times 100 \tag{22}$$

$$MRR/\% = \frac{1}{|N|}\sum_{i}^{|N|} \frac{1}{rank_i} \times 100 \tag{23}$$

其中:N 为测试集中问题答案对的数量,$label_i$ 为第 i 个预测答案是标准答案的预测结果,预测正确则为 1,预测错误则为 0。$rank_i$ 为在同一个问题的候选答案集合中标准答案的排名。

2.3　实验及对比分析

本文实验在 Windows 10 操作系统,12 Gi 显存 Nvidia GeForce GTX 1080Ti 显卡,Intel(R)Xeon(R)CPU E5-2620 v4,PyTorch1.0.0 框架上完成。通过进行多组对比实验,来分析本文提出的模型在答案选择任务上的性能,对比的模型主要包括 CNN 模型和 BiLSTM 模型。本文提出的基于自注意力和门控机制的答案选择模型,首先通过 GloVe 获取词向量表示;其次采用自注意力机制获取问题与答案的句向量表示,采用卷积注意力获取答案的第二种向量表示,并通过门控机制融合两种表示;最后计算问题答案句向量的相关性,得到问题与候选答案的匹配度,挑选匹配度最高的答案作为问题的最佳答案。其中,在利用自注意力获取句向量时,采用 6 层自注意力;在利用卷积注意力获取句向量时,采用 1 层卷积层、1 层注意力层。模型的问题输入和答案输入长度都为 120,对于句子词数超过 120 的,删除多余词;对于句子次数不足 120 的,添加占位符进行补全,输出也是长度固定的向量。当模型训练结束后,返回候选答案的排名和标注。

本文利用 5 种模型进行对比实验,分别为多尺度的 CNN 模型、BiLSTM 模型、基于 CNN 和 BiLSTM 的模型、基于注意力的 BiLSTM 的模型和基于自注意力的模型。5 种对比模型和本文模型的实验结果见表 2。结果表明:本文模型与 BiLSTM 模型、自注意力模型和基于注意力的 BiLSTM 模型相比,在 webMedQA 数据集上 *MRR* 分别提高了 8.37%、4.79%和 2.03%。

表 2　模型结果的评价指标

模型	$a_{\text{top1}}/\%$	$MRR/\%$
Multi-scale CNN[17]	40.03	60.53
BiLSTM	58.13	66.55
CNN+BiLSTM	53.17	68.10
BiLSTM+注意力模型[2]	66.00	72.89
自注意力模型	58.82	70.13
本文模型	59.33	74.92

利用 CNN 模型进行实验时,实验采用不同尺度(单个 1×1、3×3、5×5 卷积核)卷积神经网络的特征图来提取不同的局部信息,然后应用最大池化层简化表示,最后计算两个向量的余弦值。通过增

大卷积核加深网络深度,可以一定程度提升特征提取能力,但是由于单个 CNN 通常具有固定滑动窗口,从模型中提取的信息是有限的,针对汉语的复杂结构和表达形式,该模型在本文使用的中文数据集中不能取得比较理想的实验结果。

利用 BiLSTM 模型进行实验时,隐藏层大小为 128,初始学习率为 0.01。由于在每个隐藏层的计算中,接收了上一层的状态,进行上下文信息提取的同时保留了输入序列的顺序特征,使得提取的语义特征更加准确。

利用 CNN-BiLSTM 模型进行实验时,采用 CNN 进行局部语义特征提取,对已提取的局部特征采用 BiLSTM 提取上下文信息,实验结果的评价指标略高于单一模型的实验结果。

利用 BiLSTM 融合注意力模型进行实验时,实验结果的评价指标得到了明显的提升,每一个词的权重计算对提高问题答案匹配度有更大的帮助。实验说明注意力机制能更好地捕捉答案与问题之间的语义关联。

利用自注意力模型进行实验时,实验结果的评价指标要明显优于基础神经网络模型,表明自注意力机制在答案选择方面有很好的性能。实验采用的是与本文相同的参数,但是与本文模型仍有一定差距:由于改进的自注意力模型解除了位置编码与单词之间的关联,收敛速度更快,对位置信息的提取更加准确,且自注意力机制有更短的路径距离,更容易学习到长程依赖关系;前向反馈网络层增强了模型的表达能力;门控机制融合了自注意力提取的特征和卷积注意力提取的特征表示,既包含全局信息,又包含局部特征,提高了特征提取能力。因此本文模型的语义特征提取能力更强,输出信息更加完整,从而使问题与正确答案的匹配程度更高。

为了对比不同自注意力层数和不同头部数量对模型性能的影响,在 webMedQA 数据集上设计了 2 个对比实验。不同自注意力层数对模型性能的影响如图 4 所示。由图 4 可知,层数增加到 4 后,层数对模型表达能力的影响并不大,但是对于答案选择这种注重语义信息的任务,在高层注意力的模型中表现的更好。这是由于在初始注意力层中,每个词的注意力会在自身上;随着注意力层数不断增加,注意力会集中在前几个词或者后几个词;当注意力层数增加到一定后,注意力最终集中在句子的末尾。这表明高层注意力能学习词与词之间的长程依赖关系,捕获到更多的语义信息。

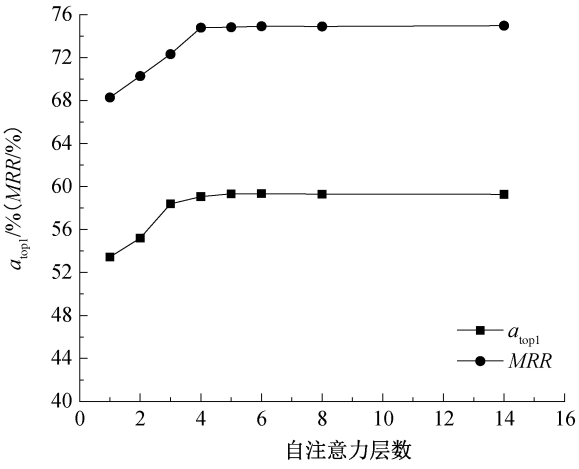

图 4 不同自注意力层数在 webMedQA 上的 a_{top1} 和 *MRR* 指标值

多头注意力机制的头部数量对模型性能的影响如图 5 所示。由图 5 可知:头部数量为 6 时,a_{top1} 和 *MRR* 值最理想;当头部数量超过 6 时,模型的效果并没有提升。虽然理论上,更多的头部数量能在不同的头部中注意到更多的信息,但是由于模型的隐层状态维度是固定的,当头部数量过多时,将使得每个头部的维度过小,从而限制了模型的表达能力。

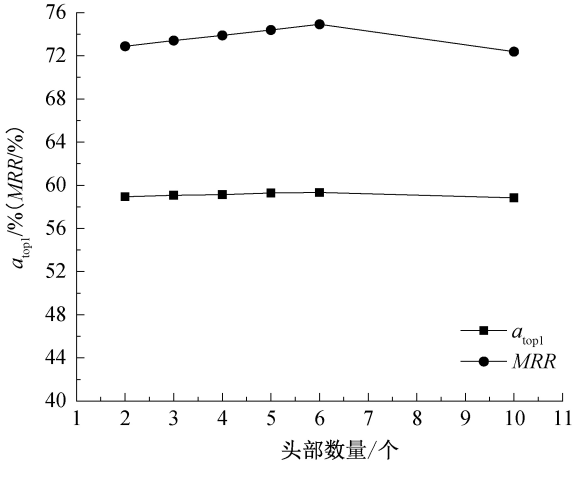

图 5 不同头部数量在 webMedQA 上的 a_{top1} 和 *MRR* 指标值

本文为了使模型更加理想,对模型进行多次调参优化。对于 GloVe,本文使用的词嵌入维度为 300 维;对于编码层,本文设置层数为 6,多头部注意力分为 6 个,输出层使用 ReLu 函数进行激活;对于卷积注意力模块,采用单个 1×1 卷积核,卷积的激活函数采用双曲正切函数 tanh;对于相关性计算,损失函数为平均绝对误差。整个网络的批尺寸为 32,学习率为 0.01,共进行了 20 轮训练,最好的一次实验在语料中 *MRR* 达到了 74.92%。

3 结　论

本文将自注意力和门控机制引入答案选择任务，在问答句内部利用层叠自注意力获取向量表示，使问题与答案文本交互前降低噪声词的影响。在模型中把位置编码从输入层拆开后，解除单词与位置之间的关联，预训练损失收敛更快，对位置信息的提取更准确。通过对比 5 种方法，本文模型在答案选择任务中可以实现更好的性能。实验结果表明，采用门控机制融合两种特征表示，获取的句子表示包含更准确的信息，使模型表达能力更好。关于不同自注意力层数的实验结果表明，层数与模型性能呈正相关，高层注意力捕获到的语义信息更多，针对注重语义信息的答案选择任务，本文提出的模型效果较好，能有效提升预测正确答案的准确率。

参考文献：

[1] Fan H J, Ma Z Y, Li H Q, et al. Enhanced answer selection in CQA using multi-dimensional features combination [J]. Tsinghua Science and Technology, 2019, 24(3): 346-359.

[2] Heilman M, Smith N A. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions [C]//Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics. California, USA: Association for Computing Machinery, 2010: 1011-1019.

[3] Chan W, Zhou X D, Wang W, et al. Community answer summarization for multi-sentence question with group L1 regularization [C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea: Association for Computational Linguistics, 2012: 582-591.

[4] Chen Z W, He Z, Liu X W, et al. Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases [J]. BMC Medical Informatics & Decision Making, 2018, 18(2):53-68.

[5] Zhou X Q, Hu B T, Chen Q C, et al. Answer sequence learning with neural networks for answer selection in community question answering [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing, China: Association for Computational Linguistics, 2015: 713-718.

[6] Wakchaure M, Kulkarni P. A scheme of answer selection in community question answering using machine learning techniques [C]//2019 International Conference on Intelligent Computing and Control Systems (ICCS). Madurai, India: IEEE, 2019: 879-883.

[7] Tan M, dos Santos C, Xiang B, et al. Improved representation learning for question answer matching [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: Association for Computational Linguistics, 2016: 464-473.

[8] 熊雪，刘秉权，吴翔虎. 基于注意力机制的答案选择方法研究[J]. 智能计算机与应用，2018，8(6):90-93.

[9] Shao T H, Guo Y P, Chen H H, et al. Transformer-based neural network for answer selection in question answering[J]. IEEE Access, 2019, 99(7):146-156.

[10] 谢正文，柏钧献，熊熙，等. 基于增强问题重要性表示的答案选择算法研究[J]. 四川大学学报(自然科学版)，2020，57(1)：66-72.

[11] Yu A W, Dohan D, Luong M T, et al. QANet: combining local convolution with global self-attention for reading comprehension [EB/OL]. (2018-04-23) [2021-01-29]. https://arxiv.org/abs/1804.09541v1.

[12] Shen Z, Bello I, Vemulapalli R, et al. Global self-attention networks for image recognition [EB/OL]. (2020-10-14) [2021-01-29]. https://arxiv. org/abs/2010.03019.

[13] Ashish V, Noam S, Niki P, et al. Attention is all you need [C]//Proceedings of the Annual Conference on Neural Information Processing Systems. California, USA: Association for Computing Machinery, 2017: 6000-6010.

[14] Pennington J, Socher R, Manning C, et al. Glove: global vectors for word representation[C]//Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: Association for Computational Linguistics, 2014: 1532-1543.

[15] He J Q, Fu M M, Tu M S, et al. Applying deep matching networks to Chinese medical question answering: a study and a dataset[J]. BMC Medical Informatics and Decision Making, 2019, 19(2): 91-100.

[16] Sugawara H, Takamura H, Sasano R, et al. Context representation with word embeddings for WSD [C]// Computational Linguistics. Singapore: Springer, 2015: 108-119.

[17] Zhang S, Zhang X, Wang H, et al. Chinese medical question answering using end-to-end character-level multi-scale CNNs[J]. Applied Science, 2017, 7(8): 767-783.

(责任编辑：康　锋)