# 基于混合多头注意力和胶囊网络的特定目标情感分析

王家乾，龚子寒，薛云，庞士冠，古东宏
（华南师范大学 物理与电信工程学院，广东 广州 510006）

**摘要：**特定目标情感分析旨在判断上下文语境在给定目标词下所表达的情感倾向。对句子语义信息编码时，目前大部分循环神经网络或注意力机制等方法，不能充分捕捉上下文中长距离的语义信息，同时忽略了位置信息的重要性。本文认为句子的语义信息、位置信息和多层次间的信息融合对该任务至关重要，从而提出了基于混合多头注意力和胶囊网络的模型。首先，使用多头自注意力分别在位置词向量基础上对上下文长句子和在双向 GRU 基础上对目标词进行语义编码；然后，使用胶囊网络在语义信息交互拼接基础上进行位置信息编码；最后，在融入原始语义信息基础上，使用多头交互注意力对上下文与目标词并行融合的方法得到情感预测结果。在公开数据集 SemEval 2014 Task4 和 ACL 14 Twitter 上实验表明，本文模型性能较传统深度学习和标准注意力方法有显著提升，验证了模型的有效性和可行性。

**关键词：**特定目标情感分析；胶囊网络；多头注意力

**中图分类号：**TP391　　**文献标识码：**A

# Aspect-based Sentiment Analysis Based on Hybrid Multi-Head Attention and Capsule Networks

Jiaqian Wang, Zihan Gong, Yun Xue, Shiguan Pang, Donghong Gu
(School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou, Guangdong 510006, China)

**Abstract:** Aspect-based sentiment analysis aims to determine the sentimental polarity expressed by context under a given target. At present, most of methods such as recurrent neural network or attention mechanism can't fully capture semantic information over long distances and ignore the importance of position information. The paper argued the semantic, positional information and multi-level information fusion of sentences are crucial to this task. So we proposed a model based on hybrid multi-head attention and capsule networks. Firstly, multi-head self-Attention was used to encode long sentences based on position word vectors and target words based on Bi-GRU respectively; Then, capsule network was used to encode the position based on the interactive splicing of semantic information; Finally, on the basis of the original semantic information, the results were obtained by integrating context with target entity using multi-head interactive attention. Experiments on SemEval 2014 Task4 and ACL 14 Twitter showed that the performance of this model significantly improved comparing with traditional deep learning and standard attention methods, which verified the effectiveness and feasibility of our method.

**Keywords:** Aspect-based Sentiment Analysis; Capsule Networks; Multi-head attention

## 0　引言

特定目标情感分析（Aspect-based Sentiment Analysis，ABSA）是情感分析领域中的细粒度分析任务，其主要目标是判断句子在不同目标下对应的情感倾向（积极、消极和中性）[1-2]。例如：This mobile phone is beautiful in appearance but expensive in price，就特定目标 appearance 而言，情感极性是积极的；而对于特定目标 price，其情感极性却是消极的。

目前，ABSA 主要采用基于浅层机器学习方法和深度学习的方法。传统浅层机器学习方法[3-5]主要通过人工设计和提取特征（如词袋模型、语义特征、情感词典、语言规则等）来对目标和句子之间的关系进行建模，从而获得大量相关的特

**收稿日期：**2019-06-15　　**定稿日期：**2019-08-15

**基金项目：**国家自然科学基金面上项目（项目编号：61876205），全国统计科学研究项目（项目编号：2016LY98），广东省科技计划项目（项目编号：2016A010101020, 2016A010101021, 2016A010101022），广州市科技计划项目(项目编号：201802010033, 201804010433)

**作者简介：**王家乾（1994—），硕士研究生，主要研究领域为自然语言处理；薛云（1975—），通信作者，博士，教授，主要研究领域为自然语言处理和数据挖掘。

征，最终输入到分类器中得到情感分析结果。这些方法考虑了目标词对情感分类的重要性，但都高度依赖于复杂的特征工程来提升性能，而且特征的设计和提取工作量也十分庞大。

随着深度学习的发展，循环神经网络 RNNs[6]和标准注意力机制[7]被广泛应用于自动学习上下文和目标词的语义特征，同时捕捉在特定目标下上下文中相关的情感特征词，解决了传统使用人工特征提取的缺陷[8-12]。例如，Chen[9]等人通过双向 LSTM 的记忆功能和使用门控循环单元网络方式来组合多个注意力的情感向量的方式来增强记忆网络对长距离句子的记忆功能。Song[11]等人通过在多头注意力基础上设计基于注意力的语义编码网络，摒弃了存在长距离依赖问题的 RNNs 网络。Li[12]等人通过引入位置词向量学习到句子中每个目标词的特定位置，然后进一步编码学习在特定目标词下的上下文语义表示。

尽管这些方法能较好地解决 ABSA 任务，但仍面临着三方面挑战：（1）对句子语义信息编码时，RNNs 的每一个输出状态都依赖于上一个时刻的状态，在语义建模时可能丢失长距离的情感信息和对输入数据不能进行并行计算等问题[13]。同时，标准注意力机制中由于权重值分布过于分散，容易引入过量噪声，难以准确提取足够的上下文情感信息。（2）对句子语义信息编码时，大部分方法忽略了目标词的位置信息对上下文句法结构的重要性，位置词向量[14]的引入仅能浅层引入每个词的位置信息，不能对整个上下文的句法结构进行动态的更新重构。（3）对句子和目标词的信息融合时，大部分基于简单拼接或相乘的组合方式可能丢失了部分原始信息，不能充分融合两者信息；且仅考虑了特定目标对于句子不同成分的影响，忽略了句子对目标实体的影响。受 Ashish Vaswani [14]和 Hinton[15]的启发，本文认为句子的语义、位置信息和多层次信息融合对该任务至关重要，从而提出了基于混合多头注意力和胶囊网络的模型(Hybrid Multi-Head Attention and Capsule Networks, HMAC)来解决上述问题。在公开数据集 SemEval 2014 Task4[2]和 ACL 14 Twitter[16]上的实验表明，我们的模型性能较传统深度学习和标准注意力方法有了显著的提升，验证了本文方法的有效性和可行性。

本文的主要贡献如下：

（1） 本文结合 RNNs 对短距离句子序列信息提取和多头自注意力机制对长距离句子并行语义信息编码的优点，提出了对上下文长句子采用多头自注意力进行语义编码，对特定目标使用双向 GRU 和多头自注意力进行语义编码，充分提取了长短距离句子的语义和情感信息，同时在语义编码时引入了位置词向量对句子的位置信息进行浅层提取。

（2） 针对语义编码层多头注意力和位置词向量对句子向量的句法和序列信息提取能力不足，忽略目标词位置信息在上下文情感词提取中作用的问题，本文结合胶囊网络提取更加丰富的语义信息和句法结构的能力，提出改进动态路由的胶囊网络对句子位置信息编码，使得上下文和目标词间的信息能够充分融合。

（3） 在上下文和目标词融合中，本文提出了两阶信息融合的方法，低阶融合中对两者语义编码后的信息进行交互拼接，从而作为胶囊网络的输入以深层次提取丰富的语义和句法位置信息；高阶融合中对得到的语义和位置编码信息使用多头交互注意力的方式进行融合，使得最终的特征表达充分考虑了目标和句子的紧密联系。

## 1 相关工作

特定目标情感分析是情感分析领域中的细粒度分析任务，是当前 NLP 领域的研究热点之一，其主要目标是判断句子在不同目标下对应的情感倾向（积极、消极和中性），目前深度学习已成为该领域主流的研究方法。

### 1.1 基于深度学习的特定目标情感分析

深度学习模型能够利用分布式表示自动学习并得到目标的相关特征，近年来循环神经网络 RNNs 和递归神经网络等在特定目标的情感分析任务中取得了巨大的成功。Dong[16]为了将目标信息整合到递归神经网络中，提出一种自适应的深度学习方法，能够提取句子中与目标相关的文本信息和句法结构特征。但是，这些基于 RNNs 的方法每一个输出状态都依赖于上一个时刻的状态，在语义建模时可能出现丢失长距离的情感信息和对输入数据不能进行并行计算等问题[13]。

注意力机制[7]通过模拟人脑，能对重要信息给予更多的关注，很好地解决了长距离依赖问题，目前已经广泛应用于各大研究领域。Bahdanau[7]首次在自然语言处理领域中引入注意力机制，在机器翻译任务中可以自动地对句子中相关部分赋予较大权重，使模型的翻译正确率得到显著的提升。然而，简单注意力机制会使注意力权重的分布过于分散容易引入过量噪声，难以准确提取足够的上下文情感信息。因此，Ma[8]、Chen[9] 和 Huang[10]在特定目标情感分析研究中，通过设计多个注意力机制找到特定目标下句子中关键部分，使模型训练时可以更加关注这些重点部分，以提升模型并行计算能力和分类性能。

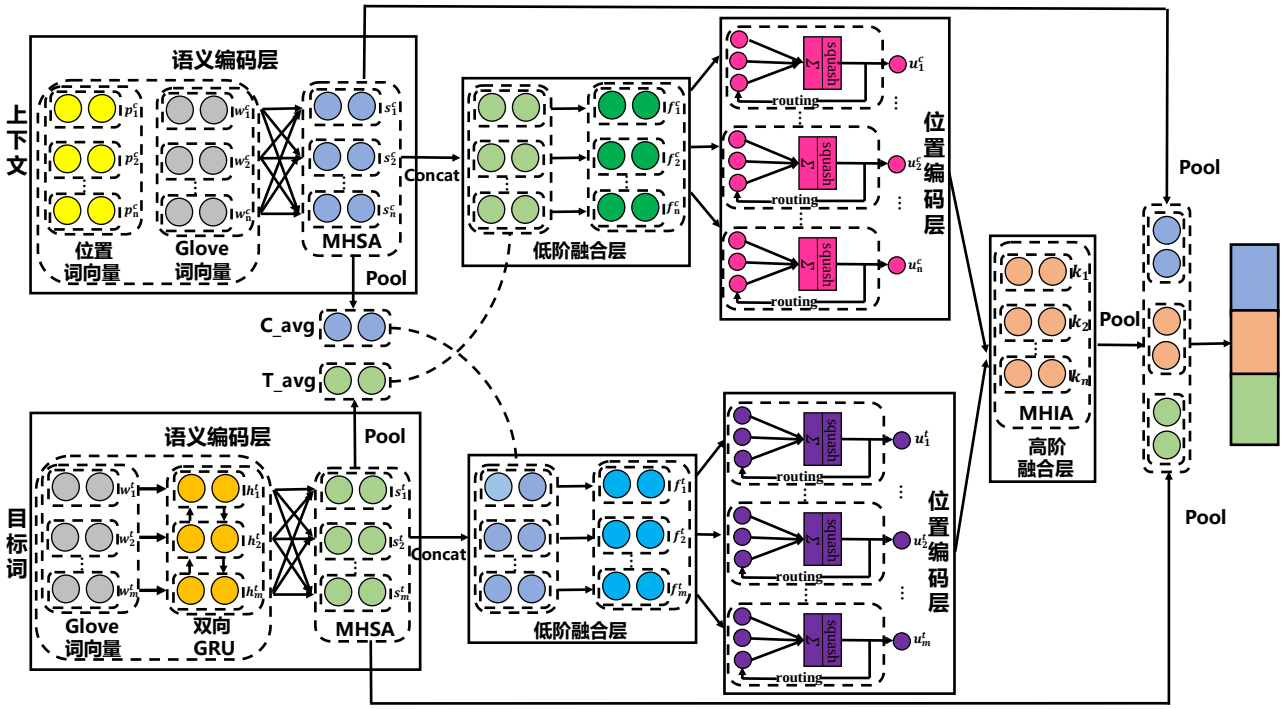


图 1 基于混合多头注意力和胶囊网络模型总体框架图

### 1.2 基于胶囊网络的文本分类方法

胶囊网络是由 Hinton[15]等人提出的一种用于图像分类的复杂神经网络结构。其目的是通过以一种动态路由算法训练的向量值表示胶囊节点来代替传统神经网络中的标量值表示神经元节点，从而可以提取更丰富的位置空间信息。目前，已有学者将胶囊网络用于文本分类和关系抽取等 NLP 研究之中。Du[17]等人提出了基于胶囊网络的混合神经网络方法来提取上下文中隐含的语义信息，且可以有效地对单词的位置、语义和句法结构进行编码，提高情感分类的效果。在多标签学习框架中，Zhang[18]等人提出了基于注意力机制的胶囊网络方法来进行关系提取，其性得到了很大的提升。这些基于胶囊网络的模型不仅在训练过程中特别关注特征信息，而且有效地针对不同的特征调整神经网络的参数，挖掘出更多隐藏的语义和位置信息。特别的，目前还没有研究将胶囊网络用于本文的细粒度情感分析的任务中，因此本文也探索了将胶囊网络用于文本的位置信息编码之中。

## 2 本文方法

在本文模型中，我们假定输入上下文序列为 $v^c=\{v_1^c,v_2^c,...,v_n^c\}$，目标词输入序列为 $v^t=\{v_1^t,v_2^t,...,v_m^t\}$，其中 $v^t$ 是 $v^c$ 的子序列，特定目标情感分类主要目的是预测句子 $v^c$ 在给定目标 $v^t$ 下的情感极性。

图 1 是本文模型的总体框架，主要由 3 部分组成：（1）语义信息编码部分：对词向量嵌入的上下文长句子和目标词进行语义信息编码；（2）位置信息编码部分：利用胶囊网络对语义信息编码的上下文和目标词部分进行深层次位置信息编码；（3）信息融合部分：引入多头交互注意力对上下文与目标词进行深层次信息融合，再与原始语义特征拼接后预测情感极性。

### 2.1 语义信息编码部分

该部分结合了 RNNs 对短距离句子和多头自注意力对长距离句子语义编码上的优点，对上下文长序列 $v^c=\{v_1^c,v_2^c,...,v_n^c\}$ 使用位置词向量和多头自注意力进行编码，对目标词 $v^t=\{v_1^t,v_2^t,...,v_m^t\}$ 使用双向 GRU 和多头自注意力进行编码，充分提取了长短距离句子的语义信息，同时引入的位置词向量也挖掘了句子的浅层位置信息。

#### 2.1.1 Glove 词向量

为了获取输入上下文和目标实体的语义信息，需要先将输入序列转化为向量形式，本文采用预训练的 Glove[19]词向量将每个单词映射到一个低维实值向量中，每个词都可以从 $W^{d_w\times|v|}$ 中得到一个向量 $w_i\in R^{d_w}$，其中 $d_w$ 表示词向量的维度，$|v|$ 表示词的个数，通过查找词嵌入矩阵，分别得到上下文词向量 $w^c=\{w_1^c,w_2^c,...,w_n^c\}\in R^{d_w\times n}$ 和目标词向量 $w^t=\{w_1^t,w_2^t,...,w_m^t\}\in R^{d_w\times m}$。

2.1.2 多头自注意力机制

注意力机制[7]起源于对人类视觉的研究，目前已广泛应用于自然语言处理领域，用来增大重要信息的权重系数，使模型关注到更重要的部分，从而可以提高分类的准确率，其定义如公式 1：

$$Attention(Q,K,V)=soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

其中 Q 表示 Query，K 表示 Key，V 表示 Value，因子 $\sqrt{d_k}$ 起调节作用使得内积不至于太大。

多头注意力机制[14]（Multi-Head Attention, MHA）是注意力机制的完善，是一种能够并行处理不同位置不同表示子空间信息的注意力机制。首先将 Q，K，V 通过参数矩阵进行映射，然后重复进行多次注意力机制，并将结果拼接起来，具体计算如公式 2，示意图如图 2：

$$\begin{aligned} head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \\ MultiHead(Q,K,V) &= Concat(head_1,...,head_h) \end{aligned} \quad (2)$$

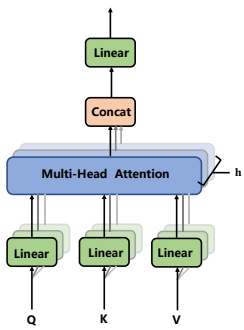


**图 2 多头注意力机制(MHA)**

多头自注意力机制（Multi-Head Self Attention, MHSA）是 MHA 的特殊情况，即输入 $Q=K=V$。本文引入 MHSA 进行语义编码，其在不引入其他冗余外部信息的前提下，寻找序列内部的关联，更好的保留原始句子语义信息，具体计算如公式 3：

$$MHA^{self}=MultiHead(\mathrm{X},\mathrm{X},\mathrm{X}) \quad (3)$$

2.1.3 上下文语义编码层

在对上下文句子进行编码时，将位置嵌入[14]也作为上下文输入的一部分，与 Glove 词向量类似，每个词都可以从 $P^{d_P\times|v_p|}$ 得到一个向量 $p_i\in R^{d_p}$，其中 $d_p$ 表示位置词向量的维度，$|v_p|$ 表示每个词和目标之间可能相对位置的个数，从而得到上下文的位置词向量 $p^c=\{p_1^c,p_2^c,...,p_n^c\}\in R^{d_p\times n}$，其中第 $i$ 个词与目标之间相对偏移量的计算如公式 4：

$$\begin{cases} i-j & i<j \\ i-j-m & j+m<i\le n \\ 0 & j\le i\le j+m \end{cases} \quad (4)$$

其中 $j$ 表示目标第一个词的索引，$m$ 表示目标的长度，$n$ 表示句子的长度。

因此,上下文词向量 $w^c$ 和位置词向量 $p^c$ 拼接后，通过 MHSA 对上下文进行语义编码表示 $s^c=\{s_1^c,s_2^c,...,s_n^c\}\in R^{d_s\times n}$，$d_s$ 表示 MHSA 的维度，其计算如公式 5:

$$s^c=MultiHead([w^c,p^c],[w^c,p^c],[w^c,p^c]) \quad (5)$$

2.1.4 目标词语义编码层

结合 RNNs 对短文本语义学习的优点，我们提出使用双向 GRU 先对目标词向量 $w^t$ 进行初始的特征提取。具体的，前向 GRU 得到隐藏层表示 $(\overrightarrow{h_1^t},\overrightarrow{h_2^t},...,\overrightarrow{h_m^t})$，后向 GRU 得到隐藏层表示 $(\overleftarrow{h_1^t},\overleftarrow{h_2^t},...,\overleftarrow{h_m^t})$，将二者拼接得到最终隐藏层表示 $h^t=(h_1^t,h_2^t,...,h_m^t)$，$h_i^t=[\overrightarrow{h_i^t},\overleftarrow{h_i^t}]$，其中 $h_i^t\in R^{2d_h}$，$d_h$ 表示隐藏层的维度。最后，通过 MHSA 对目标词进行语义编码表示 $s^t=\{s_1^t,s_2^t,...,s_m^t\}\in R^{d_s\times m}$，$d_s$ 表示 MHSA 的维度，其计算如公式 6：

$$s^t=MultiHead(h^t,h^t,h^t) \quad (6)$$

## 2.2 空间位置编码部分

为了充分考虑上下文和目标实体间相互影响，本节结合胶囊网络能够提取丰富语义和位置信息的能力，在语义编码信息 $s^c$ 和 $s^t$ 输入胶囊网络进行位置编码前，将其分别以交互拼接的低阶信息融合方式先进行重构。

2.2.1 上下文与目标词初阶融合层

在解决特定目标的情感分类任务时，目前大部分方法仅考虑了特定目标词对上下文不同成分的影响，忽略了上下文句子对目标词的影响，导致不能充分完整地捕捉两者间的全部重要信息。因此，为了解决上述问题，本文提出了两阶信息融合的方法，该节通过将两者的语义编码信息分别交互拼接的低阶融合方式，最大程度的融合两者的原始信息，影响彼此位置表示的训练过程，从而完整捕捉到目标实体和上下文中的情感信息和位置信息。

对于上下文部分：先对目标词语义信息 $s^t$ 取平均池化得到 $t_{avg}$，如公式 7；$t_{avg}$ 再与上下文 $s^c$ 每个词进行拼接线性激活得到上下文部分低阶融合

信息 $f^c=\{f_1^c,f_2^c,...,f_n^c\}\in R^{d_f\times n}$，$d_f$ 表示初阶信息融合的维度，如公式 8：

$$t_{avg}=\sum_{i=1}^{m}s_i^t/m \tag{7}$$

$$f^c=tanh(W_c[s^c,t_{avg}]) \tag{8}$$

其中，$W_c$ 是一个权重矩阵。

对于目标词部分：先对上下文语义信息 $s^c$ 取平均池化得到 $c_{avg}$，如公式 9；$c_{avg}$ 再与上下文 $s^t$ 每个词进行拼接线性激活得到上下文部分低阶融合信息 $f^t=\{f_1^t,f_2^t,...,f_m^t\}\in R^{d_f\times m}$，$d_f$ 表示初阶信息融合的维度，如公式 10：

$$c_{avg}=\sum_{i=1}^{n}s_i^c/n \tag{9}$$

$$f^t=tanh(W_t[s^t,c_{avg}]) \tag{10}$$

其中，$W_t$ 是一个权重矩阵。

2.2.2 胶囊网络位置编码层

胶囊网络[15]是 Hinton 等人提出的一种复杂神经网络结构。其目的是通过以动态路由算法训练的向量值表示胶囊节点代替传统神经网络中标量值表示神经元节点，提取更丰富的文本信息，对句子的位置、语义和句法结构进行编码[17]。为了解决目标情感分析中语义编码对句子位置信息和句法结构编码不足的问题，本文提出了改进动态路由的胶囊网络对编码后的信息进行深层次的位置编码。

胶囊网络可划分为底层胶囊和上层胶囊两部分，每个胶囊代表一些不同的属性。上层胶囊的输出是由底层胶囊和相对应的权重矩阵共同决定的。具体的，底层胶囊的输入 $u_i^{(l)}$ 即初阶信息融合后的向量 $f^c$ 或 $f^t$，然后 $u_i^{(l)}$ 通过动态路由算法（图 3）的更新方式计算得到上层胶囊的输出 $u_i^{(l+1)}$：

（1）将底层胶囊 $u_i^{(l)}$ 乘以一个参数共享矩阵 $W_j$ 得到 $\hat{u}_{j|i}$，如公式 11；设动态路由算法迭代更新的耦合系数为 $c_{ij}$，$c_{ij}$ 即为 *Softmax* 函数的输出，则 $c_{ij}$ 在胶囊网络输入层与输出层间系数之和为 1，如公式 12：

$$\hat{u}_{j|i}=W_ju_i^{(l)} \tag{11}$$

$$c_{ij}=\frac{\exp(b_{ij})}{\sum_k\exp(b_{ik})} \tag{12}$$

其中，$c_{ij}$ 为概率大小，$b_{ij}$ 为权重大小，$b_{ij}$ 初始化值设为 0。

（2）将得到的 $\hat{u}_{j|i}$ 和耦合系数 $c_{ij}$ 加权求和得到输出向量 $s_j$，如公式 13：

$$s_j=\sum_i c_{ij}\hat{u}_{j|i} \tag{13}$$

其中，$s_j$ 即上层胶囊 $u_i^{(l+1)}$ 的输入。

（3）胶囊网络的核心思想是用向量 $s_j$ 的模长来对比特征的强弱程度即显著性，因此我们提出了一个新的非线性激活函数将向量 $s_j$ 压缩转化为合适长度的输出向量 $u_i^{(l+1)}$，使得 $u_i^{(l+1)}$ 的长度不超过 1，并且保持 $u_i^{(l+1)}$ 和 $s_j$ 同方向，如公式 14：

$$\begin{aligned}u_i^{(l+1)}&=squashing(s_j)\\&=\frac{\|s_j\|^2+epsilon()}{e^{-4}+\|s_j\|^2}\frac{s_j}{\|s_j\|}\end{aligned} \tag{14}$$

其中，$epsilon()=[0,1]$ 代表一个常量值，防止输出为 0；此外，Hinton 提出的胶囊网络中压缩函数是用数值 1 来进行全局压缩[15]，本文通过实验验证发现使用较小的值 $e^{-4}$ 用于放大向量 $s_j$ 的范数能得到更好的效果，$s_j/\|s_j\|$ 则是将向量 $s_j$ 进行单位化。

（4）最后，通过衡量 $u_i^{(l+1)}$ 和 $\hat{u}_{j|i}$ 的相关性来迭代更新参数 $b_{ij}$，如公式 15：

$$b_{ij}=\hat{u}_{j|i}u_i^{(l+1)}+b_{ij} \tag{15}$$

其中，$u_i^{(l+1)}$ 和 $\hat{u}_{j|i}$ 相似性越高，点积值越大即 $b_{ij}$ 越大，底层胶囊与上层胶囊连接的可能性越大；反之，连接可能性则越小，因此权重的大小则可表示对底层胶囊识别的概率。

因此，上下文 $f^c$ 和目标词 $f^t$ 的融合信息经胶囊网络位置编码后分别输出 $u^c=\{u_1^c,u_2^c,...,u_n^c\}\in R^{d_u\times n}$，$u^t=\{u_1^t,u_2^t,...,u_m^t\}\in R^{d_u\times m}$，$d_u$ 表示输出胶囊网络的维度。

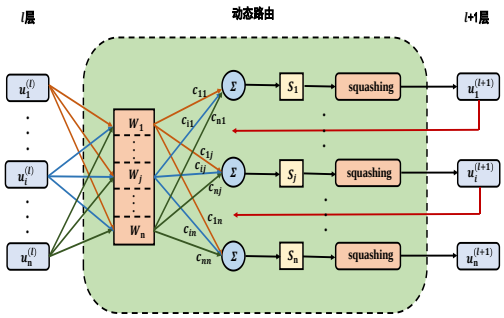


**图 3 动态路由算法示意图**

### 2.3 高阶信息融合部分

#### 2.3.1 多头交互注意力机制

多头交互注意力机制(Multi-Head Interactive Attention, MHIA) 是多头注意力在 $K=V$ 条件下的一般形式。其具体计算如公式 16：

$$MHA^{\text{inter}} = MultiHead(\mathrm{Q},\mathrm{K},\mathrm{K}) \tag{16}$$

因此，针对经胶囊网络位置编码后的 $u^c$ 和 $u^t$ 通过 MHIA 实现上下文与目标实体高阶信息融合 $k=\{k_1,\mathrm{k}_2,...,\mathrm{k}_n\}\in R^{d_k\times n}$，$d_k$ 表示 MHIA 的维度，其计算如公式 17:

$$k = MultiHead(\mathrm{u}^c,\mathrm{u}^t,\mathrm{u}^t) \tag{17}$$

#### 2.3.2 语义编码信息拼接

首先对高阶融合后的 k 取平均池化得到 $\mathrm{k}_{avg}$，如公式 18 所示；为了保留原始信息，将 $\mathrm{k}_{avg}$ 与原始语义编码特征 $c_{avg}$ 和 $t_{avg}$ 拼接，形成了最终的特征表示 o 如公式 19 所示：

$$k_{avg} = \sum_{i=1}^{n}\mathrm{k}_i / n \tag{18}$$

$$\mathrm{o} = [\mathrm{c}_{avg},\mathrm{k}_{avg},\mathrm{t}_{avg}] \tag{19}$$

### 2.4 输出层和模型训练

为了得到句子情感分类的结果，将最终的特征表示输入到 *Softmax* 层中，得到不同目标下情感极性的概率分布为：

$$\begin{aligned} x &= \tilde{W}_O^T\tilde{o}+b_o \\ y &= soft\max(x) \\ &= \frac{\exp(x)}{\sum_{k=1}^{C}\exp(x)} \end{aligned} \tag{20}$$

其中 $\tilde{W}_O^T\in R^{1\times C}$ 为可训练权重，$b_o\in R^C$ 为偏置项，C 表示分类类别，概率最大的类别即为特定目标情感分类的结果。

本文模型采用分类交叉熵总和作为损失函数，定义如公式 21 所示，同时采用反向传播算法进行权值和参数的更新。

$$L = -\sum_{i}\sum_{j=1}^{C} y_i^j \log\hat{y}_i^j + \lambda\|\theta\|^2 \tag{21}$$

其中 $i$ 为第 $i$ 个样本的下标，$j$ 为第 $j$ 种情感类别的下标；$y$ 为句子情感极性的真实分布，$y$ 为句子情感极性的预测分布；C 表示分类类别，$\lambda$ 和 $\theta$ 为正则化的参数。

## 3 实验

### 3.1 数据集介绍和参数设置

本文使用了三个公开的数据集进行实验验证：分别是国际评测 SemEval 2014 Task4[2]数据集（由 Restaurant 餐厅和 Laptop 笔记本电脑评论数据组成）和 Dong 等人开源的 ACL 14 Twitter[16]数据集。其中每条评论由句子、目标词和其对应的情感极性共同组成，旨在判断给定目标词下句子的情感极性（本文仅考虑积极、中性和消极三类），关于数据集的详细统计描述详见表 1。

**表 1 实验数据集**

| 类别 | Twitter | | Restaurant | | Laptop | |
|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test |
| 积极 | 1561 | 173 | 2164 | 728 | 987 | 341 |
| 中性 | 3127 | 346 | 633 | 196 | 460 | 169 |
| 消极 | 1560 | 173 | 805 | 196 | 866 | 128 |
| 共计 | 6248 | 692 | 3602 | 1120 | 2313 | 638 |

本文实验中采用预训练好的 Glove 词向量[19]对上下文句子和目标词进行初始化，向量维度选取 d=300；所有不在词向量词典中的单词都初始化为零向量，偏置都设置为 0；隐藏层大小设置设为 300，位置向量嵌入矩阵维度设为 50。同时，本文模型使用深度学习框架 Keras 实现，在模型训练中所有权重矩阵元素的随机初始化均服从 *glorot* 均匀分布，采用 RMSprop[20]作为模型的优化器；对应的学习率设置为 0.001，批量大小设为 128，Dropout 设为 0.5。为了更好的评估本文模型和基准模型的性能，本实验采用准确率（Accuracy）和 F1 值（macro-F1 measure）作为评价标准。

### 3.2 多头注意力个数选择实验

由于多头注意力中涉及多个 *head* 的注意力，我们实验探索了 *head* 数对 MHA 的影响。因此，我们在 Laptop，Restaurant 和 Twitter 数据集上分别测试了本文 HMAC 模型在参数 head={1,2,3,4,5,6} 的性能，评价指标为准确率（Accuracy），实验结果如图 4 所示。

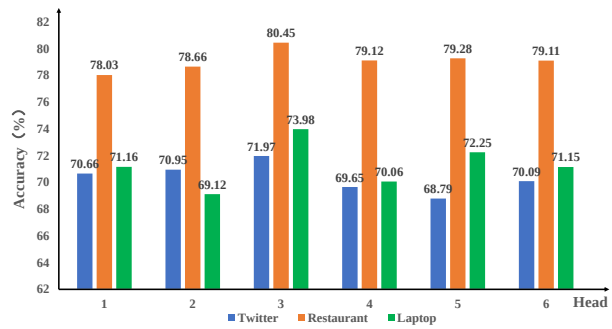


**图 4 三个数据集在 6 个不同 head 下的准确率**

可以观测到：（1）当 $head=3$ 时，三个数据集下分别能得到最高准确率（71.97%，80.45%，73.98%），表明多个 head 通常能得到较好性能，特别当 $head$ =3 时；（2）当 $head$ 较小时（如 1 或 2）性能则较差，表明上下文表示不足以包含重要的情感特征;（3）相反， $head$ 数越大并不一定能获得更好的性能，如在 Laptop 中 $head$ =6 的性能不如 $head$=4 的模型好，因为随着层数的增加，模型参数增加，使得模型难以训练和泛化。因此从整体性能出发，设 $head=3$ 时模型性能较好。

### 3. 3 胶囊网络参数选择实验

#### 3.3.1 动态路由数和维度选择

由于胶囊网络涉及动态路由更新次数和输出胶囊维度，本文实验探索了动态路由次数 $r$ 和输出胶囊维度 $d_u$ 对模型性能的影响。在三个数据集上分别测试了 HMAC 模型在 $r=\{1,2,3,4,5,6$和 $d_u=\{32,64,128,150,256,300\}$ 的性能，评价指标为准确率，图 5 是在 Twitter 数据集上的实验结果。

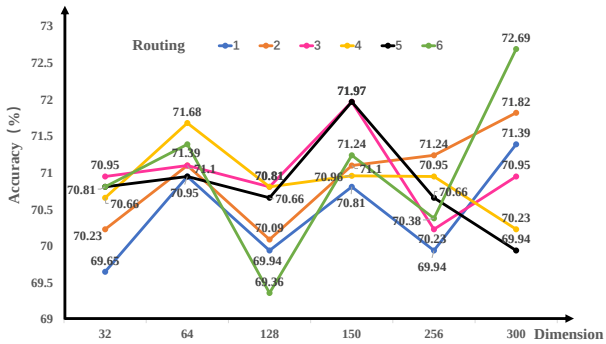


**图 5　Twitter 数据集分别在 6 个不同 r 和 $d_u$ 下的准确率**

可以观测到:(1)整体来看，当 $r=3$ 且 $d_u=150$ 时，模型整体性能较好（71.98%），表明多次动态路由更新和输出胶囊维度较高通常能获得较好的性能；（2）当 $r$ （如 1）和 $d_u$ （如 32 或 64）较小性能则较差，表明胶囊网络中动态路由机制无法很好的更新其耦合参数，且较小的输出维度无法很好的表征每个胶囊输出的概率；（3）反之， $r$ 和 $d_u$ 越大并不一定能获得更好的性能，因为随着动态路由数的增加会导致过拟合和参数增加，且输出胶囊维度过高。此外，图中也出现了当 $r$ 为 6， $d_u$ 为 300 时取得最高 72.69%的精度。因此从整体性能出发，设 $r=3$ 且 $d_u=150$ 时模型性能较好。

#### 3.3.2 Squashing 压缩函数参数选择

由于胶囊网络自身结构涉及到的压缩函数 Squashing 是其提出的主要核心思想和机制[15]，本文实验探索了不同压缩值对输入胶囊向量 $s_j$ 模长压缩的影响，在三个数据集上分别测试了本文模型

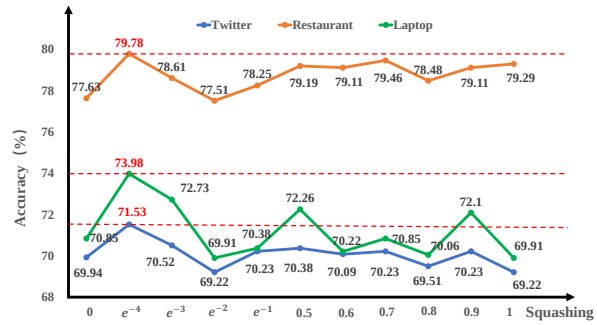


**图 6　不同压缩函数值对应的准确率**

HMAC 在的性能，评价指标为准确率，实验结果如图 6 所示。

可以观测到：（1）当压缩函数值为 $e^{-4}$ 时，同等实验环境下模型在三个数据集上分别能得到最高的准确率 （Twitter = 71.53%，Restaurant = 79.78%，Laptop = 73.98%），表明当压缩函数值趋近 0 时性能较好；（2）当压缩函数值过小为 0 时，其性能会出现明显的下降，分别比最优值下降(1.59%, 2.15%, 3.13%);（3）当压缩函数值在 (0, 1) 间时，其性能都有不同程度的下降。(4) 当取 Hinton 原始的压缩函数值 1 时，其性能整体差于其它压缩值，说明全局压缩不一定是最佳的方式。因此，本文模型对胶囊网络的压缩函数值取值为 $e^{-4}$ 。

### 3. 4 模型对比实验

#### 3.4.1 实验对比模型介绍

为了全面的评估本文 HMAC 模型的性能，我们选择 12 个典型 Baseline 模型比较：

**Feature-based SVM：**一种传统的基于复杂特征工程的支持向量机的方法。

**LSTM:** 标准的单层 LSTM 网络，进行分类的时候没有利用目标的信息。

**TD-LSTM：**以目标为中心，分别从左右两个方向采用 LSTM 网络进行建模，从而能够得到目标的上下文信息。

**Bi-GRU：**采用 Bi-GRU 网络对句子进行建模，并利用最后隐藏层的输出进行预测。

**MemNet：**一种考虑目标词内容和位置的深度记忆网络，多次采用注意力机制来获取上下文词的重要性。

**CNN-MemNet：**在 MemNet 模型的基础上改用多层卷积神经网络，并采用注意力机制来捕获 CNN 的情感特征。

**RAM：**采用双向 LSTM 记忆网络，通过门控递归单元将多个注意力机制的输出组合起来用于句子表示，从而增强了 MemNet。

**ATAE-LSTM：**在输入层将每个词向量和目标词向量进行拼接，从而形成上下文输入，同时在隐藏层将 LSTM 网络的输出和目标词向量进行拼接，然后使用注意力机制得到最终的特征表示。

**表 2 模型总体性能对比表**

（基准模型结果均是从已发表论文中检索得到；“NA”表示论文未报告的结果；“_”下划线表示本文通过复现论文模型得到其为报告的指标结果；前三名的结果使用粗体标出；*表示对比实验中最优的结果）

| 模型 | Twitter（%） | | Restaurant（%） | | Laptop（%） | |
|---|---|---|---|---|---|---|
| | Accuracy | Macro-F1 | Accuracy | Macro-F1 | Accuracy | Macro-F1 |
| Feature-based SVM (Kiritchenko, 2014)[4] | 63.40 | 63.30 | 80.16 | NA | 70.49 | NA |
| LSTM (Tang, 2016a)[6] | 66.50 | 64.70 | 74.28 | **70.90** | 66.45 | 63.90 |
| TD-LSTM (Tang et al., 2016a)[6] | 70.80 | **69.00** | 75.37 | 64.51 | 68.25 | 65.96 |
| Bi-GRU[12] | 65.30 | 62.30 | 80.27 | 70.80 | 73.35 | 70.60 |
| MemNet (Tang, 2016b)[21] | 68.50 | 66.91 | 78.16 | 65.83 | 70.33 | 64.09 |
| CNN-MemNet (Zhang, 2019) [24] | **70.89** | NA | 78.77 | NA | 73.83 | NA |
| RAM (Chen, 2017) [9] | 69.36 | **67.30** | 80.23 | 70.80 | **74.49** | **71.35*** |
| ATAE-LSTM (Wang, 2016) [10] | 68.00 | 66.00 | 77.20 | 67.02 | 68.88 | 65.93 |
| IAN (Ma, 2017) [8] | NA | NA | 78.60 | 63.59 | 72.10 | 66.50 |
| PBAN (Gu, 2018) [22] | NA | NA | **81.16*** | NA | **74.12** | NA |
| LSTM+SynATT (He, 2018) [23] | NA | NA | **80.45** | **71.26** | 72.57 | **69.13** |
| Mul-AT-CNN (Zhang, 2019) [24] | **71.25** | NA | 79.46 | NA | **75.39*** | NA |
| **Our-HMAC** | **72.69*** | **70.93*** | **80.45** | **71.62*** | 73.98 | **69.01** |

**IAN**：以交互的形式学习上下文和目标词的注意力，并分别生成目标词和上下文词的表示。

**PBAN**：使用两个 Bi-GRU 网络分别提取上下文和目标词的特征，并使用双向注意力机制对目标词和上下文间的关系进行建模。

**LSTM+SynATT**：采用基于句法的注意力机制代替传统的注意力机制，结合 LSTM 进行语义建模来进行特定目标情感分析。

**Mul-AT-CNN**：采用多层卷积神经网络并行处理上下文，并对文本进行多次建模，然后通过注意力机制显式学习其情感表示。

### 3.4.2 对比实验结果

准确率(Accuracy)和 F1 值(macro-F1 measure)作为本文对比实验的评价标准，实验结果和模型性能对比情况如表 2。

### 3.4.3 实验结果分析

从表 2 实验结果可看出：本文 HMAC 模型在 Twitter (71.26%, 70.93%)和 Restaurant (80.45%, 71.62%)数据集上整体取得了较优性能，在小数据集 Laptop (73.98%, 69.01%)上性能略差。一方面是因为本模型结合上下文和目标词部分的长短距离特性，分别设计了不同语义和位置信息编码的方法。同时另一方面通过引入了多头交互注意力机制的方法实现了两阶信息融合，充分挖掘出目标词和上下文间的语义关系，最后的实验结果证实了模型的有效性。

（1）本文基于深度学习相比于传统机器学习方法性能较好。表中 Kiritchenko 提出的 Feature-based SVM 模型[4]在依赖大量人工特征提取的基础上使用支持向量机进行分类来提升模型整体性能，在 Twitter，Restaurant，Laptop 三个数据集上分别得到了 63.40%，80.16%和 70.49%的准确率。本文的无人工特征提取深度学习方法相比于机器学习方法分别高出 9.29%，0.29%和 3.49%。说明深度学习适合用于特定目标情感分析的研究。

（2）本文对目标词短文本部分采用双向 GRU 进行语义学习的方法相比于标准 RNNs 的方法性能较好。表中 TD-LSTM 在考虑了目标信息基础上再使用 LSTM 进行语义特征学习，整体性能略高于标准的单层 LSTM 网络，表明了目标信息对提高分类精度起到了重要的作用。特别的我们发现，双向 GRU 达到了很好的性能，在 Restaurant 和 Laptop 上准确率分别达到了 80.27%和 73.35%，分别比 TD-LSTM 高出 4.9%和 5.1%。说明双向 GRU 适用于对短文本进行初步的语义特征提取。

（3）本文使用多头注意力机制语义编码的方法相比标准多注意力机制的方法性能较好。表中 MemNet 通过多个 hops 来简单线性组合不同注意力，提取出上下文中重要情感词，其在三个数据集上的准确率和 F1 值分别为(68.50%, 66.91%), (78.16%, 65.83%), (70.33%, 64.09%)，均远低于本文模型的性能。此外，RAM 通过双向 LSTM 的记忆功能和使用门控循环单元网络方式来组合多个注意力的情感向量的方式来增强 MemNet 模型，其准确率和 F1 值分别为(68.50%, 66.91%), (78.16%, 65.83%), (70.33%, 64.09%)，整体性能均优于 MemNet 模型。特别的，本文模型在 Twitter 和 Restaurant 上分别比 RAM 高(3.33%, 3.63%)和(0.22%, 0.82%)，在 Laptop 上低(0.51%, 2.34%)，说

明本文模型在大数据集上的拟合能获得更好的效果。因此，本文的多头自注意力方法适合进行语义信息编码，特别是对于上下文长距离句子。

（4） 本文提出的两阶信息交互融合方法相比简单交互或拼接方法性能较好。表中 ATAE-LSTM 通过对上下文的输入拼接目标词和引入注意力机制的方式，来增强目标词对整个上下文情感词提取的影响，其在 Restaurant 和 Laptop 上准确率分别达到了 77.20%和 68.88%，比 TD-LSTM 高 1.6%和 0.6%，主要是因为 TD-LSTM 仅考虑了目标信息，同等地对待每一个词在最终结果中起到的作用，不能识别出句子中具有重要信息的词语。IAN 模型在 ATAE-LSTM 基础上同时考虑上下文和目标词之间的相互影响，并设计了交互注意力的信息融合方式，在两个数据集上分别实现了 78.60%和 72.10%的准确率，比 ATAE-LSTM 高 1.40%和 3.22%。本文低阶交互拼接和高阶多头交互注意力的信息融合方式性能均高于 ATAE-LSTM 和 IAN 模型，说明本文信息融合方式适用于特定目标情感分类研究。

（5） 本文基于胶囊网络的位置信息编码方法相比基于标准句法分析的方法性能较好。表中 LSTM+SynATT 在对目标词表示重构后，在注意力机制模型上融入了依存句法分析来挖掘句子的句法信息，其在 Restaurant 和 Laptop 上准确率达到了 80.45%和 72.57%，性能略差于本文模型。此外，PBAN 整体性能较优于本文模型，CNN-MemNet 和 Mul-AT-CNN 基于 CNN 的深度学习方法也取得较好的性能，能够很好的解决信息编码的问题。因此，本文基于胶囊网络的位置编码方法适用于目标情感分类研究。

## 4 结语

特定目标情感分析旨在判断上下文语境在给定实体下所表达的情感倾向。针对 RNNs 的每一个输出状态都需要依赖于上一个时刻的状态，存在对长距离依赖句子的语义建模时可能丢失远距离的情感信息词，且对输入数据不能进行并行处理的问题。本文结合 ABSA 任务中上下文是长句子和目标词是短句子的问题特点，提出了仅对目标词部分保留使用双向 GRU 进行语义特征学习。此外，在此基础上采用注意力机制进行语义编码，但是由于标准注意力机制中权重值分布的过于分散容易引入过量噪声，难以准确提取足够的上下文情感信息。本文提出了使用多头自注意力机制进行语义编码，充分提取句子中的原始语义信息。但是，在语义编码层多头注意力和位置词向量对句子向量的句法和序列状信息提取能力不足，忽略了目标词位置信息对上下文情感词提取的重要性，本文结合胶囊网络能够获取单词的位置与语义信息和句法结构更丰富的信息的能力，使用改进动态路由的胶囊网络来对句子位置信息编码，使得上下文和目标词间的信息能够充分融合。最后，在上下文和目标词融合中，本文提出了两阶信息融合方法，低阶融合中对两者语义编码后的信息进行交互拼接，从而作为胶囊网络的输入来深层次提取丰富的语义和句法位置信息；高阶融合中对得到的语义和位置编码信息使用多头交互注意力的方式进行融合，从而使得最终的特征表达充分考虑了目标和句子的紧密联系。

在 Twitter、Laptop 和 Restaurant 上的实验结果表明，本文提出 MHAC 模型相比基于深度学习模型结果有显著提高，准确率分别达到 72.69%、80.45%和 73.98%；F1 值分别达到 70.93%、71.62%和 69.01%。

尽管当前模型已经达到很高的性能，但未来还有很多研究工作需要开展。首先，将目标信息有效地嵌入到神经网络单元中是一个比较可行的研究方向。其次，利用改进多头注意力机制和胶囊网络这两种大参数量的网络，避免在小数据集上过早发生过拟合的问题，以提升小数据集上情感分类性能，这也将会是我们未来的主要研究重点。

## 参考文献


[1] Gao Y, Zhang Y, Xiao T. Implicit Syntactic Features for Target-dependent Sentiment Analysis[C]//IJCNLP 2017: the Eighth International Joint Conference on Natural Language Processing, 2017: 516-524.

[2] Pontiki M, Galanis D, Pavlopoulos J, et al. Semeval-2014 task 4: Aspect based sentiment analysis [C]//Proc of the 8th International Workshop on Semantic Evaluation, 2014: 27-35.

[3] Vo D, Zhang Y. Target-dependent twitter sentiment classification with rich automatic features[C]//IJCAI 2015: the twenty-fourth international joint conference on artificial intelligence, 2015:1347-1353.

[4] Kiritchenko S, Zhu X, Cherry C, et al. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews [C] //Proc of the 8th International Workshop on Semantic Evaluation, 2014: 437-442.

[5] Long J, Yu M, Zhou M, et al. Target-dependent Twitter sentiment classification[C]//ACL 2011: The 52nd Annual Meeting of the Association for Computational Linguistics, 2011: 151-160.

[6] Tang D, Qin B, Feng X, et al. Effective lstms for target-dependent sentiment classification[C]//COLING 2016: Proc of the International Conference on Computational Linguistics, 2016: 3298-3307.

[7] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate [J]. Computer Science, 2014.[8] Ma D, Li S, Zhang X, et al. Interactive atten-

tion networks for aspect-level sentiment classification[C]//IJCAI 2017: the twenty-sixth international joint conference on artificial intelligence, 2017: 4068–4074.
[9] Chen P, Sun Z, Bing L, et al. Recurrent Attention Network on Memory for Aspect Sentiment Analysis [C]//EMNLP 2017: Proc of the Conference on Empirical Methods in Natural Language Processing, 2017: 452-461.
[10] Wang Y, Huang M, Zhao L, et al. Attention-based lstm for aspect-level sentiment classification[C]//EMNLP 2016: Proc of the Conference on Empirical Methods in Natural Language Processing, 2016: 606-615.
[11] Song Y, Wang J, Jiang T, et al. Attentional encoder network for targeted sentiment classification[C]//arXiv preprint arXiv:1902.09314, 2019.
[12] Li L, Liu Y, Jiang T, et al. Hierarchical Attention Based Position-aware Network for Aspect-level Sentiment Analysis [C]//CoNLL 2018: Proceeding of the 22nd Conference on Computational Natural Language Learning, 2018: 181-189
[13] Li X, Song J, Gao L, et al. Beyond RNNs: Positional Self-Attention with Co-Attention for Video Question Answering[C]//AAAI 2019: The 33rd AAAI Conference on Artificial Intelligence, 2019.
[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//NIPS 2017: Advances in Neural Information Processing Systems, 2017: 5998 6008.
[15] Sabour S, Frosst N, Hinton G. Dynamic routing between capsules[C]//NIPS 2017: Advances in Neural Information Processing Systems, 2017:3856-3866.
[16] Li D, Wei F, Tan C, et al. Adaptive recursive neural network for target-dependent twitter sentiment classification[C]//ACL 2014: The 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 49–54.
[17] Du Y, Zhao X, He M, et al. A Novel Capsule Based Hybrid Neural Network for Sentiment Classification[C]//IEEE Access, 2019: 39321-39328
[18] Zhang N, Deng S, Sun Z, et al. Attention-Based Capsule Networks with Dynamic Routing for Relation Extraction[C]//EMNLP 2018: Proc of the Conference on Empirical Methods in Natural Language Processing,2018: 986-992
[19] Pennington J, Socher R, and Manning C. Glove: Global vectors for word representation[C]// EMNLP 2014: Proc of the Conference on Empirical Methods in Natural Language Processing, 2014: 1532–1543.
[20] Tieleman T, Hinton G. “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude,” [OL]. 2012. COURSERA: Neural Networks for Machine Learning, vol. 4, p. 2.
[21] Tang D, Qin B, Liu T. Aspect level sentiment classification with deep memory network[C]//EMNLP 2016: Proc of the Conference on Empirical Methods in Natural Language Processing, 2016: 214-224.
[22] Gu S, Zhang L, Hou Y, et al. A Position-aware Bidirectional Attention Network for Aspect-level Sentiment Analysis[C]//COLING 2018: Proceeding of the 27th International Conference on Computational Linguistics, 2018: 774-784.
[23] He R, Lee W, Ng T, et al. Effective Attention Modeling for Aspect-Level Sentiment Classification[C]//COLING 2018: Proceeding of the 27th International Conference on Computational Linguistics, 2018: 1121-1131.
[24] Zhang S, Xu X, Pang Y, et al. Multi-layer Attention Based CNN for Target-Dependent Sentiment Classification[J] .Neural Process Lett, 2019.

王家乾（1994—），硕士研究生，主要研究领域为自然语言处理和深度学习。
E-mail：wjq672425265@gmail.com

龚子寒（1996—），硕士研究生，主要研究领域为自然语言处理、数据挖掘、文本情感分析。
E-mail：zihan.gong@m.scnu.edu.cn

薛云（1975—），通信作者，博士，教授，主要研究领域为自然语言处理和挖掘。
E-mail：xueyun@scnu.edu.cn