文章编号：1003-0077(2019)06-0001-11

注意力机制在深度学习中的研究进展

朱张莉[1]，饶 元[1]，吴 渊[1]，祁江楠[1]，张 钰[2]

(1. 西安交通大学 软件学院 社会智能与复杂数据处理实验室，陕西 西安 710049；
2. 陕西师范大学 计算机科学学院，陕西 西安 710119)

摘 要：注意力机制逐渐成为目前深度学习领域的主流方法和研究热点之一，它通过改进源语言表达方式，在解码中动态选择源语言相关信息，从而极大改善了经典 Encoder-Decoder 框架的不足。该文在提出传统基于 Encoder-Decoder 框架中存在的长程记忆能力有限、序列转化过程中的相互关系、模型动态结构输出质量等问题的基础上，描述了注意力机制的定义和原理，介绍了多种不同的分类方式，分析了目前的研究现状，并叙述了目前注意力机制在图像识别、语音识别和自然语言处理等重要领域的应用情况。同时，进一步从多模态注意力机制、注意力的评价机制、模型的可解释性及注意力与新模型的融合等方面进行了探讨，从而为注意力机制在深度学习中的应用提供新的研究线索与方向。

关键词：深度学习；注意力机制；编码器—解码器

中图分类号：TP391 **文献标识码**：A

Research Progress of Attention Mechanism in Deep Learning

ZHU Zhangli[1], RAO Yuan[1], WU Yuan[1], QI Jiangnan[1], ZHANG Yu[2]

(1. Lab of Social Intelligence&Complex Data Processing, School of Software, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China; 2. School of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi 710119, China)

Abstract: The attention mechanism has gradually become one of the popular methods and research issues in deep learning. By improving the source language expression, it dynamically selects the related information of the source language in decoding, which greatly improves the insufficiency issue of the classic Encoder-Decoder framework. On the basis of the issues in the conventional Encoder-Decoder framework such as long-term memory limitation, interrelationships in sequence transformation, and output quality of model dynamic structure, this paper describes a varied aspects on attention mechanism, including the definition, the principle, the classification, state-of-the-art researches as well as the applications of attention mechanism in image recognition, speech recognition, and natural language processing. Meanwhile, this paper further discusses the multi-modal attention mechanism, evaluation mechanism of attention, interpretability of the model and integration of attention with the new model, providing new research issues and directions for the development of attention mechanism in deep learning.

Keywords: deep learning; attention mechanism; Encoder-Decoder

0 引言

近年来，随着深度学习技术的逐步兴起，越来越多的研究人员将深度学习模型引入到自然语言处理(natural language processing, NLP)以及多媒体的内容对象识别任务中。例如，Y Kim 采用卷积神经网络(convolutional neural networks, CNN)解决话题分类任务[1]；I Sutskever 将递归神经网络(recurrent neural networks, RNN)应用到文本生成任务

收稿日期：2018-08-06 **定稿日期**：2018-10-15

基金项目：国家自然科学基金(61741208)；教育部"云数融合"基金(2017B00030)；中央高校基本科研业务费(zdyf2017006)；陕西省协同创新计划(2015XT-21)；西安市碑林区科技创新计划项目(GX1803)；陕西烟草公司科技攻关项目(ST2017-R011)；中央高校建设世界一流大学(学科)和特色发展引导专项资金(PY3A022)；深圳市科技项目(JCYJ20180306170836595)

中[2]；Ma Xuezhe结合多种深度神经网络解决序列标注的问题[3]。这些使用深度学习网络的方法在各领域任务中都取得了比以往研究更好的效果。其中，Encoder-Decoder作为深度学习任务中较为常见的框架，已成功应用于各个领域。但是由于其作为一种通用的框架模型，并不针对某一特定领域而设计，这就导致了该框架本身存在以下亟待解决的问题和技术挑战。

（1）长程记忆能力有限：传统的深度学习模型如RNN在新的时间状态下不断叠加输入序列会导致前面的输入信息变得越来越模糊，即存在长程记忆能力有限问题；CNN虽然在一定程度上可以缓解该问题，但由于滤波器一般不会选择太大，所以并不能很好地解决此问题。特别是如果源句子序列非常长，那么由于梯度更新中衰减较大，导致序列头部的参数无法有效更新，模型难以学到合理的向量表示，并且先输入的内容携带的信息会被后续输入的信息稀释掉，输入序列越长，这个现象就越严重。因此，如何改善模型设计的结构缺陷，成为引入注意力机制需要解决的基本问题。

（2）序列转化过程中的相互关系：从源序列A到目标序列B的转化过程中，解码器基本上沿着源序列的顺序依次接受输入，且输入始终是一段固定的特征向量，并不是按照位置对A进行检索。由于不同的时间步长或者空间位置信息具有明显的差别，利用定长表示无法很好的解决信息损失问题，如何更好地利用注意力机制学习内容之间的相互关系进而表示这些信息，具有很高的实用价值，成为亟待解决的问题。

（3）提升模型动态结构输出质量：让任务处理系统能够更有效地获得输入数据与当前的输出数据之间的有用信息，从而提高输出的质量。在实际使用中可能要求在不同时刻关注不同信息，可通过引入现有注意力机制及注意力机制的变种动态获取需要关注的部分信息，进而生成更合理的输出。

注意力机制（attention mechanism）的出现，使得传统Encoder-Decoder框架中存在的问题得以缓解。注意力机制通过对模型中不同关注部分赋予不同的权重，并从中抽取出更加重要和关键的信息，从而优化模型并做出更为准确的判断。Google DeepMind团队于2014年率先在RNN模型上引入注意力机制来实现图像的分类[4]，完成了图像中多个物体对象的高效准确识别[5]，使其在MNIST分类任务中错误率下降4%，验证了注意力机制在图像处理领域的有效性，同时也使得结合注意力机制的神经网络成为了研究热点。随后，Bahdanau等[6]在文献[7]工作的基础上，第一个将注意力机制引入到机器翻译领域，这也是注意力机制在NLP中的首次应用。随着研究的推进，W Yin在其本人先前工作[8]的基础上引入注意力机制，并将新模型应用于句子建模的任务中[9]。在此基础上，Google机器翻译团队摒弃了依赖于复杂RNN或CNN来处理序列变换的Encoder-Decoder架构，转而采用更为简单的基于注意力机制的序列转换器网络架构（Transformer）进行序列变换[10]，并在WMT2014 English-to-German翻译任务中，将BLEU值提升至28.4，超当时的最佳模型2分。

可见，如何更好地将注意力机制与神经网络相结合，已成为当前深度学习各领域共同关注的热点问题[6,9-10]。本文主要针对注意力机制与深度学习网络模型结合过程中相关的技术挑战以及相应的最新研究工作进展进行分析与综述。

1 注意力机制定义与原理

注意力机制是由Treisman和Gelade提出的一种模拟人脑注意力机制的模型[4]，它可以看成是一个组合函数，通过计算注意力的概率分布，来突出某个关键输入对输出的影响。一般大多数注意力机制均基于Encoder-Decoder框架，特别是在文本处理领域中常用的Encoder-Decoder抽象框架，如图1所示。

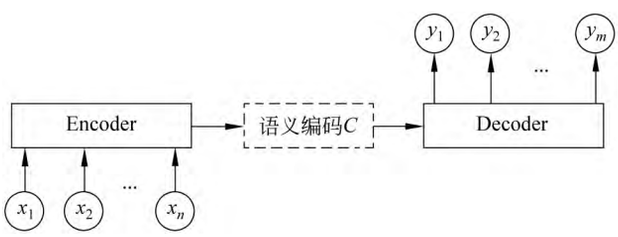

图1 抽象的Encoder-Decoder框架

该模型将一个变长的输入$X=(x_1,x_2,\cdots,x_n)$，映射到一个变长输出$Y=(y_1,y_2,\cdots,y_m)$。其中，Encoder（编码器）把一个变长的输入序列X，通过非线性变换转化为一个中间的语义表示C：$C=f(x_1,x_2,\cdots,x_n)$；Decoder（解码器）的任务是根据输入序列X的中间语义表示C和先前已经生成的$y_1,y_2,\cdots,y_{i-1}$来预测并生成i时刻的输出$y_i=g(y_1,y_2,\cdots,y_{i-1},C)$，其中，$f()$和$g()$均为非线性转化函

数。由于传统的 Encoder-Decoder 框架对输入序列 X 缺乏区分度，因此 Bahdanau 等[6]引入了注意力机制来解决这个问题，他们提出的模型结构如图 2 所示。

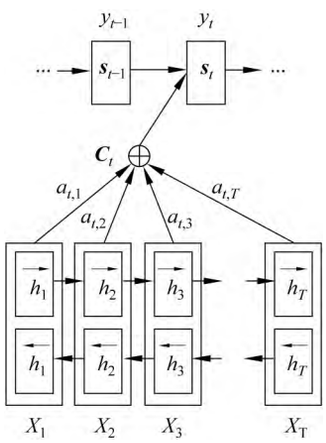

图 2　注意力机制的结构示意图[6]

其中，s_{t-1} 是 Decoder 端在 $t-1$ 时刻的隐状态，y_t 是目标词，C_t 是上下文向量，则 t 时刻的隐状态，如式(1)所示。

$$s_t = f(s_{t-1}, y_{t-1}, C_t) \tag{1}$$

C_t 依赖于编码端输入序列的隐藏层表示，通过加权处理后可表示如式(2)所示。

$$C_t = \sum_{j=1}^{T} \alpha_{t,j} h_j \tag{2}$$

其中，h_j 表示 Encoder 端第 j 个词的隐向量，它包含整个输入序列的信息，但重点关注第 j 个词周围的部分。T 是输入端长度，$\alpha_{t,j}$ 表示 Encoder 端第 j 个词对 Decoder 端第 t 个词的注意力分配系数，且 $\alpha_{t,j}$ 概率值之和为 1。$\alpha_{t,j}$ 的计算公式，如式(3)所示。

$$\alpha_{t,j} = \frac{\exp(a_{t,j})}{\sum_{j=1}^{T} \exp(a_{t,j})}$$
$$a_{t,j} = a(s_{t-1}, h_j) \tag{3}$$

其中，$a_{t,j}$ 表示一个对齐模型，用于衡量 Encoder 端位置 j 的词相对于 Decoder 端位置 t 的词的对齐程度/影响程度。通常将对齐模型 a 参数化作为前馈神经网络与系统中其余部分共同训练。常见的对齐方式有如下 4 种。

1) 加性注意力(additive attention)

$$a(s_{t-1}, h_j) = v_a^{\mathrm{T}} \tanh(W_a s_{t-1} + U_a h_j) \tag{4}$$

其中，$W_a \in \mathrm{R}^{n \times n}$，$U_a \in \mathrm{R}^{n \times 2u}$ 和 $v_a \in \mathrm{R}^n$ 表示权重矩阵，u 为单向隐藏层单元数。加性注意力是最经典的注意力机制[6]，它通过使用隐含层的前馈神经网络来计算注意力的权重，由于 $U_a h_j$ 不依赖于 t，可提前计算以最小化计算量。

2) 乘法(点积)注意力(multiplicative attention)

$$a(s_{t-1}, h_j) = s_{t-1}^{\mathrm{T}} W_a h_j \tag{5}$$

其中，$W_a \in \mathrm{R}^{n \times 2u}$ 表示权重矩阵。乘法注意力[11]和加性注意力在复杂度上是相似的，但由于乘法注意力可以使用矩阵操作，使其在实践中计算速度更快，且存储性能更高。在低维度解码器状态中两者性能相似，但在高维情况上，加性注意力的性能更优[11]。

3) 自注意力(self-attention)

$$A = \mathrm{softmax}(v_a \tanh(W_a H^{\mathrm{T}}))$$
$$H = (h_1, h_2, \cdots, h_T) \tag{6}$$

其中，$H \in \mathrm{R}^{T \times 2u}$ 表示输入序列的隐向量。$W_a \in \mathrm{R}^{d_a \times 2u}$ 是一个权重矩阵，$v_a \in \mathrm{R}^{r \times d_a}$ 是一个参数向量，其中 d_a 为一个自定义的超参数，r 为需要从输入序列中抽取的信息个数，A 即为最终得到的注意力矩阵。由此可见自注意力机制通常不需要其他额外信息，它能够关注自身进而从中抽取相关信息[12]。

4) 关键值注意力(key-value attention)

关键值注意力是 Daniluk 最近提出的注意力机制的变体[13]，它将形式和函数分开，从而为注意力计算保持分离的向量。具体而言，关键值注意力将每一个隐藏向量 h_j 分离为一个键 k_j 用于计算注意力分布 α_j 和一个值 v_j 用于编码下一个词的分布和上下文表示，如式(7)所示。

$$\begin{bmatrix} k_j \\ v_j \end{bmatrix} = h_j$$
$$a_{t,j} = \tanh(W_1[k_{j-L}; \cdots; k_{j-1}] + (W_2 k_j)\mathbf{1}^{\mathrm{T}})$$
$$\alpha_{t,j} = \mathrm{softmax}(v_a^{\mathrm{T}} a_{t,j})$$
$$c_{t,j} = [v_{j-L}; \cdots; v_{j-1}]\alpha_{t,j}^{\mathrm{T}} \tag{7}$$

其中，$W_1, W_2 \in \mathrm{R}^{n \times n}$ 和 $v_a \in \mathrm{R}^n$ 是权重矩阵，L 为注意力窗体的长度，$\mathbf{1}$ 为所有单元为 1 的向量。

2　注意力机制分类

2.1　基本注意力机制结构

2.1.1　软注意力机制和硬注意力机制

Xu K 根据每一时间步所关注的区域是一个区域还是所有的区域，将注意力机制分为如下两类[14]：

1) 软注意力机制(soft attention)

软注意力机制考虑所有的输入，但并不是给每个输入相同的权重，而是更关注某些特定的输入。如图2所示，软注意力机制会为每一个特征分配一个注意力权值，即一个概率分布。其特定区域信息的上下文向量 $C_{t,j}$ 可直接通过比重加权求和得到，如式(8)所示。

$$\mathrm{E}_{p(s_t|a)}[\boldsymbol{C}_{t,j}]=\sum_{j=1}^{T}\alpha_{t,j}\boldsymbol{h}_j \tag{8}$$

软注意力机制是参数化的，光滑且可微，可以被嵌入到模型中直接训练，且梯度可以通过注意力机制模块反向传播到模型的其他部分。

2) 硬注意力机制(hard attention)

硬注意力机制是一个随机过程，在某一时刻只关注一个位置信息，注意力相对集中，常采用 One-Hot 形式。位置信息的多元伯努利分布如式(9)所示。

$$\begin{aligned}&p(s_{t,j}=1\mid s_{i<t},\boldsymbol{H})=\alpha_{t,j}\\&\boldsymbol{c}_{t,j}=\sum_i s_{t,j}\boldsymbol{h}_j\end{aligned} \tag{9}$$

注意力权重 $\alpha_{t,j}$ 在此所起的作用是表明该位置是否被选中，只有0，1两个选项。$s_{t,j}$ 是一个 one-hot 指示器，值为1表示第 j 个位置被选中，否则为0。为了实现梯度的反向传播，需要采用蒙特卡罗采样[11]方法来估计模块的梯度。

两种注意力机制都有各自的优点，软注意力机制相对发散，而硬注意力机制会专注于某一特定区域。Minh-Thang Luong[11]在此基础上进一步提出针对上述两种注意力机制的改进版本，即全局注意力(global attention)和局部注意力(local attention)。全局注意力机制关注全部位置的信息，因此计算开销较大，为提升模型效率，遂提出局部注意力机制，该机制每次仅需关注源语言编码中一个较小的上下文窗口，其计算复杂度要低于全局注意力机制和软注意力机制，且与硬注意力机制不同的是，它几乎处处可微，易于训练。因此，常认为局部注意力机制是软注意力机制和硬注意力机制优势上的混合体。

2.1.2 位置注意力机制

一般地，在文本中与目标词距离较近的上下文词汇比距离较远的词汇更重要。因此，Duyu Tang 将位置信息(location attention)编码到注意力模型中，并归纳出如下四种编码策略[15]。

1) $\boldsymbol{m}_j=\boldsymbol{e}_j\odot\boldsymbol{l}_j$

$l_j^k=(1-L_j/T)-(k/d)(1-2\times L_j/T)$ (10)

其中，$\odot$ 表示逐元素相乘，$\boldsymbol{e}_j\in\mathrm{R}^{d\times 1}$ 和 $\boldsymbol{l}_j\in\mathrm{R}^{d\times 1}$ 分别表示词 x_j 的词向量和位置向量，$\boldsymbol{m}_i$ 即记忆向量表示。位置向量可通过式(10)计算得出，其中，T、k 和 d 分别表示句子长度、跳数及维度，L_j 是词 x_j 的位置。

2) $\boldsymbol{l}_j=1-L_j/T$ (11)

这是模型1)的简化版，其在不同的跳中使用相同的位置向量 $\boldsymbol{l}_j$。当 L_j 距离越大时，$\boldsymbol{l}_j$ 的重要性就越低。

3) $\boldsymbol{m}_j=\boldsymbol{e}_j+\boldsymbol{l}_j$ (12)

将位置向量 $\boldsymbol{l}_j$ 视为模型的一个参数，使用向量相加得到记忆向量。位置向量随机初始化，并通过梯度下降学习得到。

4) $\boldsymbol{m}_j=\boldsymbol{e}_j\odot\sigma(\boldsymbol{l}_j)$ (13)

与模型3)不同的是，位置表示被认为是控制有多少单词语义被写入记忆的神经门。对位置向量进行 sigmoid 函数 σ 处理，并使用逐元素相乘计算记忆向量 $\boldsymbol{m}_j$。

在上述4个模型中，模型1)和模型2)位置向量的值是固定的，并以启发式的方式进行计算；模型3)和模型4)位置向量作为参数与其他参数共同训练。模型2)比较直观、计算成本更低且不损失精度。模型4)对神经门的选择非常敏感。

2.1.3 输入序列注意力机制

虽然基于位置的编码取得了一定的进步，但是它们不足以完全捕获特定词汇与目标实体的关系以及它们可能对目标关系的影响。因此，Linlin Wang 提出将注意力机制用于输入序列(input attention mechanism)，并设计了基于多层注意力机制的卷积神经网络模型，用来自动识别与关系分类相关的输入句的部分[16]。图3为将注意力机制应用在输入序列的整体结构示意图。

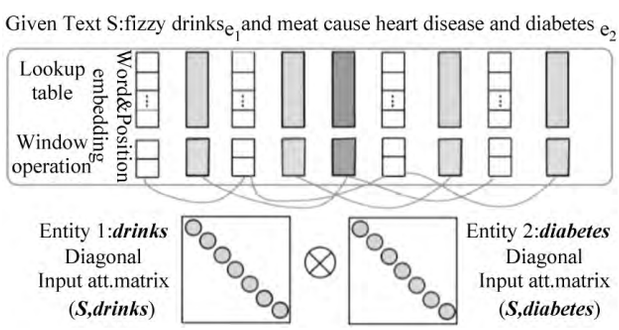

图3 面向输入序列的注意力机制示意图[15]

通过学习输入语句中各部分对核心实体的注意力机制，训练出实体上下文相关的对角矩阵 $\boldsymbol{A}^j$，该矩阵中各元素 $A_{i,i}^j=f(n_j,x_i)$ 反映出词语 x_i 与给定

实体 n_j 的上下文相关强度，即分配在该词上对于实体的注意力。打分函数 f 表示二者的内积，被参数化到网络中并在训练过程中更新。为了衡量输入序列中第 i 个词与第 j 个实体间的关联程度，可定义如下因子，如式(14)所示。

$$\alpha_i^j = \frac{\exp(A_{i,i}^j)}{\sum_{i'=1}^{T}\exp(A_{i',i'}^j)} \tag{14}$$

当输入序列中存在两个实体时，第 i 个词对该两个实体的相关因子分别为 α_i^1 和 α_i^2，可通过如下三种处理方式识别该词与两个实体之间的关系 r_i。

1）直接平均

$$\boldsymbol{r}_i = \boldsymbol{c}_i \frac{\alpha_i^1 + \alpha_i^2}{2} \tag{15}$$

其中，$\boldsymbol{c}_i \in \mathrm{R}^{(d_w+2d_p)k}$ 为第 i 个词的上下文信息，d_w 和 d_p 是超参数，k 为滑窗大小。

2）串联

$$\boldsymbol{r}_i = [(\boldsymbol{c}_i\alpha_i^1)^{\mathrm{T}}, (\boldsymbol{c}_i\alpha_i^2)^{\mathrm{T}}]^{\mathrm{T}} \tag{16}$$

通过将词向量串联获得第 i 个词输入注意力部分更多的信息，它包含了两个实体的相关关系。

3）距离

$$\boldsymbol{r}_i = \boldsymbol{c}_i \frac{\alpha_i^1 - \alpha_i^2}{2} \tag{17}$$

该处理方式将关系理解为两个实体之间的映射，且结合了两个实体特定的权重以获取它们之间的关系。

输入注意力部分的最终输出是矩阵 $\boldsymbol{R}=[\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_T]$。

2.1.4　自注意力机制

自注意力[10,17-18]（Self-attention，SAN）又称内部注意力，它通常仅关注自身并从中抽取相关信息，而不使用其它额外信息。传统的注意力机制中，源端和目标端的内容是不一样的，得到的结果即源端的每个词与目标端每个词之间的依赖关系。而在自注意力机制中，注意力发生在源端内部元素之间或者目标端内部元素之间。

2.2　组合注意力机制结构

2.2.1　协同注意力机制(co-attention)

协同注意力是注意力机制的一种变体，是一种双向注意力，它不再只关注单独的数据源 $\boldsymbol{P}\in\mathrm{R}^{d\times N}$，而是共同关注数据源 $\boldsymbol{P}$ 和数据源 $\boldsymbol{Q}\in\boldsymbol{R}^{d\times T}$。协同注意力需要使用数据源 $\boldsymbol{Q}$ 引导生成数据源 $\boldsymbol{P}$ 的注意力，而且还需要使用数据源 $\boldsymbol{P}$ 引导生成数据源 $\boldsymbol{Q}$ 的注意力权重。如图 4 所示，它多用于问答系统[19-20]，不仅需要给阅读的资源（文档或图片）生成一个注意力权重，而且还需要给问句生成一个注意力权重，该方法适用于多模态问题。

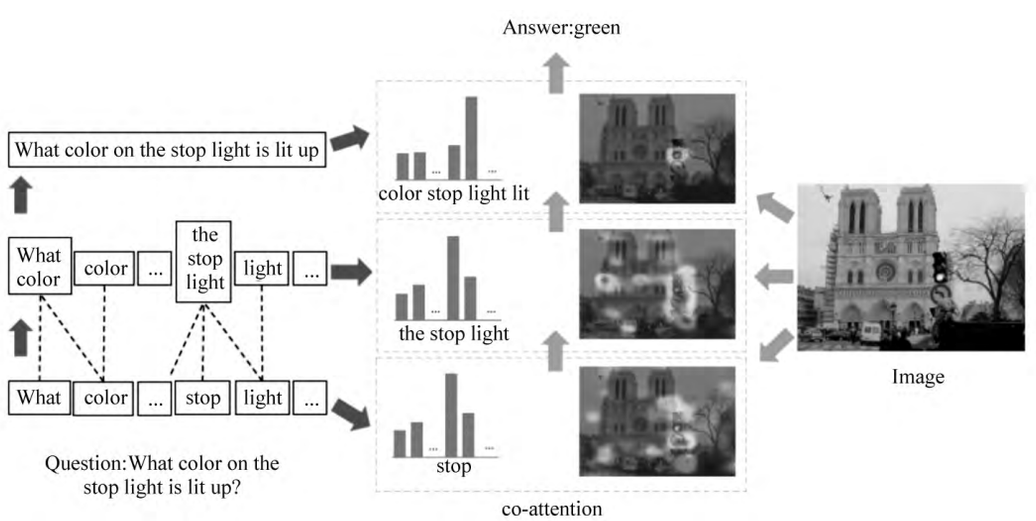

图 4　层次协同注意力模型[19]

一般地，根据生成数据源注意力顺序的不同，可将协同注意力分为两种方式[19]：

1. 平行协同注意力(parallel co-attention)

平行协同注意力同时关注数据源 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$。通过计算数据源 $\boldsymbol{P}$ 和数据源 $\boldsymbol{Q}$ 特征的相似性来将其结合，并形成一个关联矩阵 $\boldsymbol{M}\in\mathrm{R}^{T\times N}$，如式(18)所示。

$$\boldsymbol{M} = \tanh(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{W}_b\boldsymbol{P}) \tag{18}$$

其中，$\boldsymbol{W}_b\in\mathrm{R}^{d\times d}$ 是权重矩阵。关联矩阵 $\boldsymbol{M}$ 横向代表数据源 $\boldsymbol{Q}$，纵向代表数据源 $\boldsymbol{P}$，数据源 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$ 的注意力计算可采用如下两种方法实现。

(1) 关联矩阵简单的最大化输出

$$a^p[n] = \max_i(M_{i,n})$$
$$a^q[t] = \max_j(M_{t,j}) \tag{19}$$

通过简单的最大化输出与其他形态位置的关联性,来分别计算数据源 P 和 Q 的注意力。

(2) 将关联矩阵作为一个特征

$$\boldsymbol{H}^p = \tanh(\boldsymbol{W}_p\boldsymbol{P} + \boldsymbol{M}(\boldsymbol{W}_q\boldsymbol{Q}))$$
$$a^p = \mathrm{softmax}(\boldsymbol{w}_{hp}^{\mathrm{T}}\boldsymbol{H}^p)$$
$$\boldsymbol{H}^q = \tanh(\boldsymbol{W}_q\boldsymbol{Q} + \boldsymbol{M}(\boldsymbol{W}_p\boldsymbol{P}))$$
$$a^q = \mathrm{softmax}(\boldsymbol{w}_{hq}^{\mathrm{T}}\boldsymbol{H}^q) \tag{20}$$

其中,$\boldsymbol{W}_p, \boldsymbol{W}_q \in \mathrm{R}^{k\times d}$,$\boldsymbol{w}_{hp}, \boldsymbol{w}_{hq} \in \mathrm{R}^k$ 是权重矩阵。$a^p \in \mathrm{R}^N$ 和 $a^q \in \mathrm{R}^T$ 是注意力概率分布。

2. 交替协同注意力(alternation co-attention)

在交替协同注意力机制中,顺序地交替生成数据源 P 和数据源 Q 的注意力。其执行过程包括如下三个方面的步骤:

① 将数据源 Q 转化为单个向量 q;

② 将 q 加入到数据源 P 中,生成数据源 P 的注意力权重;

③ 将②的输出加入到数据源 Q 中,生成数据源 Q 的注意力权重。

该方法类似于交替使用两次传统注意力机制。

2.2.2 层叠式注意力机制(attention-over-attention,AoA)

层叠式注意力机制是一种交互式注意力,率先出现在阅读理解式问答系统任务中。一般地,阅读理解式问答系统的研究将问题看作一个整体,或者只考虑问题对文档的影响,并没有考虑文档对问题的影响,而模型实际上可以利用更多的文档—问题之间的交互信息[21]。

通常,问答系统在得到成对匹配矩阵 $\boldsymbol{M} \in \mathrm{R}^{|D|*|Q|}$ 后,按列计算文档 D 中每个单词对问题 Q 中某个单词的重要程度(即注意力),最终形成一个文档级别的注意力分布 $\alpha(t)$,如式(21)所示。

$$\alpha(t) = \mathrm{softmax}(M(1,t),\cdots,M(|D|,t))$$
$$\alpha = [\alpha(1),\alpha(2),\cdots,\alpha(|Q|)] \tag{21}$$

其中,矩阵 $\boldsymbol{M}$ 横向表示问题,纵向表示文档,它是文档级问题词向量的匹配矩阵,$M(i,j)$ 代表文档中第 i 个单词的上下文嵌入与问题中第 j 个单词的上下文嵌入的点积之和。

Yiming Cui 等设计了一种新的层叠式注意力模型,对问题进行了更细致的拆解[21]。在上述文档级注意力的基础上叠加了一层[22],按行计算问题中每个单词对文档中某个单词的重要程度,形成一个问题级别的注意力分布 $\beta(t)$,然后对这些分布进行累加求平均得到 β,如式(22)所示。

$$\beta(t) = \mathrm{softmax}(M(t,1),\cdots,M(t,|Q|))$$
$$\beta = \frac{1}{n}\sum_{t=1}^{|D|}\beta(t) \tag{22}$$

最后将文档级的注意力分布 α 与问题级别的注意力分布 β 进行点积计算即求得增加了文档级别的注意力 α_D,如式(23)所示。

$$\alpha_D = \alpha^T\beta \tag{23}$$

在文档级注意力的基础上叠加了问题级注意力,使得这个注意力更有侧重。相比于先前使用启发式融合函数[22]或设置多个超参数,Yiming Cui 等[23]提出的模型结构相对简单,且不需要设置额外的手工超参数,该模型还可在多个文档级注意力之上自动生成一个集中注意力,并进行双向的查找。相比于最先进的 EpiReader,层叠式注意力阅读器在 CBTest NE 和 CBTest CN 数据集上将准确率分别提升了 2.3%和 2.0%。近期,层叠式注意力被应用于属性级情感任务中,在 Restaurant 和 Laptop 两个数据集上平均准确率分别达到 81.2%和 74.5%[23]。

2.2.3 多头注意力机制(multi-head attention)

Google 提出的简单网络架构 Transformer,是注意力机制的完善。并在此基础上提出多头注意力模型,如图 5 所示[10]。

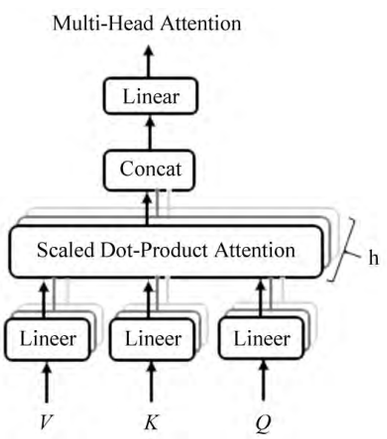

图 5 多头注意力模型[10]

其中,Q、K、V 分别是 query、key 和 value 的简写,K 和 V 是一一对应的,即为 key-value 的关系。相对于单一注意力只对 K、Q、V 进行注意力计算,研究发现通过使用模型学习得到多种映射器,分别对 K、Q、V 的各个维度进行多次线性映射,效果更佳[10],如式(24)所示。

$$\mathrm{MultiHead}(Q,K,V)=\mathrm{Concat}(\mathrm{head}_1,\cdots,\mathrm{head}_h)W^O \tag{24}$$

其中，$Where\ head_i=Attention(QW_i^Q,KW_i^K,VW_i^V)$，$W_i^Q\in R^{d_k\times d_k}$，$W_i^K\in R^{d_k\times d_k}$，$W_i^V\in R^{d_k\times d_v}$，$W^O\in R^{hd_v\times d_k}$ 是权重矩阵。

将 Q、K、V 通过矩阵映射后，重复执行 h 次 Attention 操作（参数不共享），最后将结果拼接，即为“多头”，本模型中的注意力都为 SAN。通过多头自注意力机制，能够使用不同序列位置的不同子空间的表征信息来进行序列数据处理。而在单一注意力机制中，这些不同位置不同子空间的表征信息由于取均值操作的存在，而将被模型丢弃。

3　注意力机制研究进展

3.1　针对长程记忆能力有限的问题

在神经网络中增加外部记忆，解码时与之交互，可以扩展神经网络的表达能力[24]。外部记忆可以将当前时刻重要的中间信息存储起来，用于后续时刻，在一定程度上弥补了注意力机制的不足，能够更好地扩展模型的表达能力及增强长距离依赖效果。如 Duyu Tang 等提出将基于上下文的注意力机制（content attention）运用到属性级的情感分类中[15]。系统的输入由外部记忆和属性向量两部分组成，输出向量为每一块外部记忆的加权和。该模型在 Laptop 和 Restaurant 数据集上分别在叠加 7 层和 9 层时达到最佳性能，准确率分别为 72.37% 和 80.95%。

Han Zhang[25] 提出自注意生成式对抗网络（SAGAN），能够为图像生成任务实现注意力驱动的、长范围的依存关系建模。传统的卷积 GAN 只根据低分辨特征图中的空间局部点生成高分辨率细节，在 SAGAN 中，可使用所有特征点的线索来生成高分辨率细节，且鉴别器能检查图片相距较远部分的细节特征是否彼此一致。该模型在 ImageNet 数据集中将最好的 Inception 分数纪录从 36.8 提高至 52.52，并将 Frechet Inception 距离从 27.62 减少到 18.65[25]。

3.2　序列转化过程中的相互关系

注意力机制对源端和目标端对应关系建模，是无监督的模型。近期，在多个领域引入了注意力机制的变种，使其能更好地利用源端和目标端的信息。如：Yiming Cui 等[22]提出的层叠式注意力机制，对问题进行了更细致的拆解，而不是简单地将其看成一个整体[22]，相比于最先进的 EpiReader，层叠式注意力阅读器在 CBTest NE 和 CBTest CN 数据集上将准确率分别提升了 2.3% 和 2.0%。Caiming Xiong 等在 Yiming Cui 等[21]层叠注意力的基础上借鉴交替协同注意力机制的思想，将协同注意力应用于机器阅读理解式问答任务中，在 SQuAD 数据集上，整体 F_1 值达到 80.4%[20]。Jiasen Lu 等[16]提出使用协同注意力机制处理 VQA（visual question answering）任务，将注意力机制分别用于图像和问句中，并采用平行和交替两种协同注意力机制策略。

3.3　模型动态结构输出质量的提升

注意力机制通过改进源语言表达方式，在解码中动态选择源语言相关信息，从而极大地提升了模型动态结构输出质量。如 Linlin Wang 等提出将注意力机制应用在池化层，用于学习目标类别的注意力。该方法通过对窗口大小中各输入向量的对应维度进行卷积，以此来抽取窗口中有意义的 N-gram 短语，在 SemEval-2010 数据集上，将 F_1 值提升至 88%[16]。

虽然注意力机制可以改善传统编码器—解码器中存在的一些问题，但引入注意力机制后，由于在获得注意力分配权重时，需要计算源语言句子中所有词语的权重，该过程耗费计算资源，且会导致这些模型的训练速度和推断速度变慢[17]，因此会对模型的时间和空间性能产生一定的影响。

（1）注意力机制对计算量的影响：模型引入注意力机制后可能存在计算量较大的问题，可适当使用 Minh-Thang Luong 等提出的局部注意力代替全局注意力[11]，来达到减少计算量的目的。局部注意力在生成上下文向量时只关注源语言小部分区域，把无关信息过滤掉，可以显著减少计算量。在 WMT 2014 英语到德语翻译上，局部注意力比全局注意力提高了 0.9 个 BLEU 值。另外在亚琛工业大学（RWTHAachen）英德词对齐语料上，局部注意力词对齐错误率为 34%，比全局注意力词对齐错误率减低了 5%。

（2）注意力机制对存储空间的影响：SAN 虽然可以通过高度并行计算获取输入序列中每对元素应用注意力机制生成的上下文表示，且相较于 RNN 和 CNN 而言，SAN 在对远距离和局部相关性两方面都比较灵活。但是 SAN 需要很大的存储空间存

储所有元素对的对齐分数，对存储空间的需求随序列长度呈二次方增长。为解决上述问题，Shen T等提出了一种双向分块自注意力机制(Bi-directional block self-attention，Bi-BloSA)，实现更快且节省空间的上下文融合，然后基于Bi-BloSA提出了不使用RNN/CNN的序列编码模型，称为双向分块自注意力网络(Bi-directional block self-Attention network，Bi-BloSAN)，它使用注意力机制将Bi-BloSA的输出压缩为一个向量表示[26]。这种模型具有高度的并行运算性，同时对局部和远距离相关性进行了良好的建模，相较于DiSAN(双向自注意力网络)的空间消耗随序列长度增加呈现指数级的增长，经过改进的Bi-BloSAN，其空间消耗明显降低。

(3) 注意力机制对模型训练速度和推断速度的影响：为解决模型训练速度和推断速度较慢的问题，Adams Wei Yu等在问答领域提出了一个名为QANet的新型问答系统框架，它不再需要循环网络，其编码器仅由卷积和自注意力机制构成，卷积可以对局部相互作用建模，而自注意力机制可以对全局相互作用建模[17]。在SQuAD数据集上，QANet模型的训练速度提升到对应的RNN模型的3～13倍、推断速度提升到4～9倍，并且取得了和循环模型同等的准确率。速度的提升使得能够使用更多的数据来训练模型。因此，Adams将QANet模型和使用神经机器翻译模型回译得到的数据结合了起来。在SQuAD数据集上，使用增强的数据训练的模型在测试集上获得了84.6的F_1值，这远远优于目前公开的最佳模型81.8的F_1值。C Zhou等使用自注意力机制替代卷积神经网络或长短期记忆网络(long short-term memory，LSTM)，在Amazon购买行为的公开数据集上进行单行为预测实验时，训练速度较CNN/LSTM提升近4倍[27]。

4 注意力模型的主要应用

基于注意力机制的神经网络被广泛应用于图像识别、语音识别及自然语言处理等各种不同类型的深度学习任务中，是近两年深度学习技术中值得关注与深入了解的核心技术之一。

4.1 图像识别

2014年Google Deep Mind团队率先在RNN模型上使用了注意力机制来进行图像分类[4]，该团队提出了一种基于注意力机制的任务驱动的视觉处理框架。随后，他们又提出一种基于注意力机制的用于图像中多个物体识别的模型[5]，该模型利用深度学习来训练Deep RNN，其目的是找到输入图像中最相关的区域。Google Deep Mind团队的这两项工作的实验都是基于变换的MNIST，他们将一个新的思路引入该领域，并在公开数据集上将图像分类错误率降低了4%[5]，这也为将其应用于计算机视觉中大规模对象识别及分类任务提供了一个新的方向。

4.2 语音识别

Jan Chorowski等提出了一种基于注意力机制的新模型，即“attention-based recurrent sequence generator with Convolutional Features”[28]，将注意力机制应用到语音识别领域。该模型对长输入具有很强的鲁棒性，且在单一话语及10倍长(重复)的话语中分别达到了18% PER(phoneme error rate)和20%PER。

语音识别中的经典模型(connectionist temporal classification，CTC)，在基于注意力机制的Encoder-Decoder框架中由于注意力机制建立了语音与单词之间的对应关系，因此取得了较好的结果。Dzmitry Bahdanau等[29]在文献[28]工作的基础上，将注意力机制应用到LVCSR中，主要是对注意力的计算范围进行了2w的加窗，加快训练和解码速度，提升了模型的性能。但其仍存在一些不足：适合短语识别，对长句子的识别较差；当存在噪声数据时训练不稳定。随后，Suyoun Kim等提出了将注意力与CTC结合对语音声学建模的方法，该模型共用一个Encoder，将另一端Decoder分为两部分，一部分是注意力，另一部分是CTC，并通过权重λ给定不同的权重比[30]。

4.3 自然语言处理

4.3.1 机器翻译

Bahdanau首次使用单层注意力机制解决机器翻译中不同长度的源语言对齐问题，将翻译和对齐同时进行，显著提高了神经机器翻译模型的翻译性能[6]。

为了解决定长源语言句子向量难以捕获长距离依赖的问题，Junczys Dowmunt等[31]引入了注意力机制动态计算源语言端上下文，在联合国平行语料库30个语言对上，与传统的统计机器翻译相比，除在两个语言对上神经机器翻译略逊色于基于短语的

统计机器翻译外，在其他翻译对上神经机器翻译都取得了压倒性的优势[31]。尽管如此，它仍存在一些问题：由于缺乏对注意力调整的约束，可能会导致最终翻译结果出现“过译”和“漏译”现象。

为了缓解上述问题，Z Tu 等使用一个覆盖率向量来记录注意力历史[32]，覆盖率向量作为模型的输入用于调整后续的注意力权重，它能够让神经机器翻译系统考虑更多的未翻译词。相比基于标准注意力机制的神经翻译系统，改进后的模型提高了 0.2%的翻译质量和 4.17%的对齐质量。

Cheng Yong 等发现源语言到目标语言翻译模型和目标语言到源语言翻译模型在计算注意力时均存在不足但可以相互弥补，因而通过在训练目标中加入一致性约束，鼓励两个模型相互帮助，将两个翻译方向的性能分别提高了 0.76%和 1.52%[33]。

刘洋提出，尽管将人类的先验知识和数据驱动的神经网络方法相结合在神经机器翻译工作中取得了一定的进展，但目前只能加入有限的先验知识，尚缺乏一个通用的框架来支持向神经机器翻译中加入任意的先验知识[34]。

4.3.2　文本摘要

Radev 等将摘要定义为：从一个或多个文本中提取出来的一段文字，它能够表达原始文本中重要信息，且其长度不超过或远少于原文本的一半[35]。文本摘要旨在通过机器自动输出简洁、流畅、保留关键信息的摘要。文本摘要从技术上通常可分为三类：抽取式摘要(extractive)、理解式摘要(abstractive)和压缩式摘要(compressive)。

庞超等提出了一种基于分类的理解式摘要模型，该模型将注意力机制引入到基于递归神经网络的 Encoder-Decoder 框架中，从而使模型能够更加精确地获取原文的中心内容，模型还与分类器相结合，并在大量的语料下同时训练这两部分内容。在文本摘要任务中相较于 MOSES＋模型，该模型在 ROUGE-1、ROUGE-2 和 ROUGE-L 分别提升 2.78%、0.23%和 1.04%[36]。

使用 Encoder-Decoder 框架的文本摘要是在机器翻译的基础上逐步发展的，但是两者的任务仍存在本质区别：机器翻译是要尽可能地保证信息的完整性，且翻译过程中的输入输出序列长度大致相同；而文本摘要要求尽可能使用凝练的语句来表达整体信息，且摘要的长度不超过或远少于原文本的一半。

4.3.3　问答系统

问答系统(question answering，QA)用于回答人们以自然语言形式提出的问题，周博通等[37]针对大规模知识库问答的特点，构建了一个包含 3 个主要步骤的问答系统：问句中的命名实体识别、问句与属性的映射和答案选择，使用结合注意力机制的双向 LSTM 进行属性映射。该系统在 NLPCC-ICCPOL 2016 KBQA 任务提供的测试数据集上的平均 F_1 值为 0.809 7，接近已发表的最好水平。

David Golub 等提出一种 character-level 的 Encoder-Decoder 方法，引入注意力机制，改进了之前基于 word-level 的 Encoder-Decoder 中存在的 OOV 问题及训练参数较多的问题[18]。David Golub 分别对比了 word-level 和 character-level 模型在 entity 和 predicate 上的准确率。结果显示，在 predicate 上的预测准确率相差不大，但是在 entity 上的预测准确率相差较大，word-level 的准确率大约为 45.0%，和 character-level 准确率的 96.6%相差甚远。结果表明引入注意力机制的 character-level 方法，有效地改善了问答系统中 OOV 问题，它使用较少的训练集，并且减少了训练参数，将准确率提高了 7 个百分点。

5　未来研究方向

目前，结合注意力机制的 Encoder-Decoder 框架取得巨大成功，新的研究成果不断涌现出来，注意力机制如何更好地和神经网络结合以及其自身的不断改进仍然是当下和未来的研究热点。

(1) 多模态注意力机制：多模态注意力机制利用的资源不限于文本，目前研究主要结合使用图像或语音信息。通常采用两个编码器，一个对文本信息进行编码，另一个对图像或者语音信息编码。在解码时，通过注意力机制将不同模态信息进行融合。目前工作利用的多模态信息较为单一，未来我们可将其运用于富媒体，如视频等领域。

(2) 构建通用的注意力的评价机制：虽然结合注意力机制的深度学习网络均取得了一定成功，但现在尚未存在统一的评价机制。如何结合注意力机制的优势构建通用的评测方法，将会成为未来研究的热点。

(3) 增强模型的可解释性：神经网络相当于一个“黑匣子”，无法检查系统中存在的偏差，无法对运行良好的系统提供具体的解释。可视化是最常见的事后解释类型，目前广泛应用于机器翻译等领域[6-7]，未来可借鉴 Za Chary Lipton[38] 提出的事后

解释和透明度两种解释方式增强模型的可解释性。

(4) 注意力机制与新模型的融合：神经网络的新模型不断出现，如何设计注意力机制使其能更好地与新模型相融合也是一个需要关注的热点问题。如将注意力机制融入 Hinton 的胶囊网络[39]中。

6 总结

本文从多个角度对深度学习中注意力机制进行了介绍，包括注意力模型的定义与原理、多种不同的分类方式和主要应用，并针对注意力机制研究过程中所存在的长程记忆能力有限、序列转化过程中的相互关系、模型动态结构输出质量和引入注意力机制后对模型时间和空间性能的影响等关键性技术问题与挑战进行了分析与综述。使用注意力机制的一个主要优势是它能更好地解释并可视化整个模型，即便于理解在模型输出过程中输入序列中的信息是如何影响最后生成序列的，因此融合注意力的模型在深度学习的各个领域均得到广泛的应用。

注意力机制扩展了神经网络的能力。它可以专注于输入的特定部分，使自然语言基准测试的性能得到改进，以及赋予图像字幕、记忆网络和神经程序全新的能力。结合注意力机制的深度神经网络在部分领域呈现出全面超越传统神经网络的趋势。虽然目前该机制在多模态、评价机制、可解释性及与新模型的融合等方面尚存在不足之处，但必将成为未来深度学习的发展方向。

参考文献

[1] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv: 1408.5882, 2014.

[2] Sutskever I, Martens J, Hinton G E. Generating text with recurrent neural Networks[C]//Proceedings of International Conference on Machine Learning. Bellevue, Washington:DBLP, 2016:1017-1024.

[3] Ma X Z, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf[J]. arXiv preprint arXiv: 1603.01354, 2016.

[4] Mnih, Volodymyr, Heess, et al. Recurrent models of visual attention[J]. arXiv preprint arXiv: 1406-6247, 2014.

[5] Ba J, Mnih V, Kavukcuoglu K. Multiple object recognition with visual attention[J]. arXiv preprint arXiv: 1412.7755, 2014.

[6] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv peprint arXiv: 1409,0473, 2014.

[7] Cho K, Van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN Encoder-Decoder for statistical machine translation[J]. arXiv: 1406.1078v3. 2014, 2(11): 23-37.

[8] Yin W, Schütze H. Convolutional neural network for paraphrase identification[C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics. North American: Human Language Technologies, 2015:901-911.

[9] Yin W, Schütze H, Xiang B, et al. ABCNN: Attention-based convolutional neural network for modeling sentence pairs[J]. arXiv preprint arXiv:1512.05193, 2015.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.

[11] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal:ACL, 2015:1412-1421.

[12] Lin Z, Feng M, Santos C N D, et al. A structured self-attentive sentence embedding[J]. arXiv: 1703.03130,2017.

[13] Daniluk M, Rocktäschel T, Welbl J, et al. Frustratingly short attention spans in neural language modeling[J]. arXiv preprint arXiv:1702.04521,2017.

[14] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention [J]. arXiv: 1502.03044v1. 2015:2048-2057.

[15] Tang D Y, Qin B, Liu T. Aspect level sentiment classification with deep memory network[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing . Austin, Texas: ACL, 2016:214-224

[16] Wang L L, Cao Z, Melo G D, et al. Relation classification via multi-level attention CNNs[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: ACL,2016:1298-1307.

[17] Yu A W, Dohan D, Luong M T, et al. QANet: Combining local convolution with global self-attention for reading comprehension[J]. arXiv preprint arXiv: 1804.09541, 2018.

[18] Golub D, He X. Character-level question answering with attention[J]. arXiv preprint arXiv:1604.00727, 2017.

[19] Lu J, Yang J, Batra D, et al. Hierarchical question-

image co-attention for visual question answering[J]. arXiv preprint arXiv:1606.00061,2016.

[20] Xiong C M, Zhong V, Socher R. Dynamic coattention networks for question answering[J]. arXiv preprint arXiv:1611.01604,2016.

[21] Cui Y, Chen Z, Wei S, et al. Attention-over-attention neural networks for reading comprehension[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada:ACL, 2017:593-602

[22] Cui Y, Liu T, Chen Z, et al. Consensus attention-based neural networks for Chinese reading comprehension[C]//Proceedings of COLING 2016. Osaka, Japan: The COLING 2016 Organizing Commitee, 2016:1777-1786

[23] Huang B, Ou Y, Carley K M. Aspect level sentiment classification with attention-over-attention neural networks[J]. arXiv:1804.06536,2018.

[24] Graves A, Wayne G, Reynolds M, et al. Hybrid computing using a neural network with dynamic external memory[J]. Nature, 2016, 538(7626):471-476.

[25] Zhang H, Goodfellow I, Metaxas D, et al. Self-attention generative adversarial networks[J]. arXiv: 1805.08318, 2018.

[26] Shen T, Zhou T, Long G, et al. Bi-directional block self-attention for fast and memory-efficient sequence modeling[C]//Proceedings of International Conference on Learning Representations, 2018.

[27] Zhou C, Bai J, Song J, et al. ATRank: An attention-based user behavior modeling framework for recommendation[J]. arXiv:1711.06632,2017.

[28] Chorowski J, Bahdanau D, Serdyuk D, et al. Attention-based models for speech recognition[J]. Computer Science, 2015, 10(4):429-439.

[29] Bahdanau D, Chorowski J, Serdyuk D, et al. End-to-end attention-based large vocabulary speech recognition[J]. Computer Science, 2015:4945-4949.

[30] Kim S, Hori T, Watanabe S. Joint CTC-attention based end-to-end speech recognition using multi-task learning[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2017: 4835-4839.

[31] Junczys-Dowmunt M, Dwojak T, Hoang H. Is neural machine translation ready for deployment? A case study on 30 translation directions[J]. arXiv preprint arXiv:1610.01108,2016.

[32] Tu Z, Lu Z, Liu Y, et al. Modeling coverage for neural machine translation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: ACL, 2016:76-85.

[33] Cheng Y, Wu H, Wu H, et al. Agreement-based joint training for bidirectional attention-based neural machine translation[C]//Proceedings of IJCAI, New York, USA, 2016.

[34] **刘洋. 神经机器翻译前沿进展**[J]. **计算机研究与发展**, 2017, 54(6):1144-1149.

[35] Radev, Dragomir R, Hovy, et al. Introduction to the special issue on summarization[J]. Computational Linguistics, 2002, 28(28):399-408.

[36] **庞超，尹传环. 基于分类的中文文本摘要方法**[J]. **计算机科学**, 2018, 45(1):144-147.

[37] **周博通，孙承杰，林磊，等**. InsunKBQA:**一个基于知识库的问答系统**[J]. **智能计算机与应用**, 2017, 7(5):150-154.

[38] Lipton Z C. The mythos of model interpretability[J]. arXiv preprint arXiv:1606.03490,2016.

[39] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules[J]. arXiv preprint arXiv: 1710.09829,2017.

朱张莉(1993－)，**硕士研究生，主要研究领域为自然语言处理、智能问答与深度学习。**
E-mail: 532878474@qq.com

吴渊(1990－)，**硕士研究生，主要研究领域为自然语言处理、智能问答与深度学习。**
E-mail: clock24@xjtu.edu.cn

饶元(1973－)，**通信作者，博士生导师，主要研究领域为社会智能与复杂数据处理。**
E-mail: yuanrao@163.com