工程科学学报, 第 43 卷, 第 11 期: 1499−1511, 2021 年 11 月
Chinese Journal of Engineering, Vol. 43, No. 11: 1499−1511, November 2021
https://doi.org/10.13374/j.issn2095-9389.2021.01.30.005; http://cje.ustb.edu.cn

深度学习中注意力机制研究进展

刘建伟✉, 刘俊文, 罗雄麟

中国石油大学 (北京) 自动化系, 北京 102249
✉通信作者, E-mail: liujw@cup.edu.cn

摘　要　对注意力机制的主流模型进行了全面系统的概述. 注意力机制模拟人类视觉选择性的机制, 其核心的目的是从冗杂的信息中选择出对当前任务目标关联性更大、更关键的信息而过滤噪声, 也就是高效率信息选择和关注机制. 首先简要介绍和定义了注意力机制的原型, 接着按照多个层面对各种注意力机制结构进行分类, 然后对注意力机制的可解释性进行了阐述同时总结了在各种领域的应用, 最后指出了注意力机制未来的发展方向以及会面临的挑战.

关键词　注意力机制; 全局/局部注意力机制; 硬/软注意力机制; 自注意力机制; 可解释性
分类号　TP181

Research progress in attention mechanism in deep learning

LIU Jian-wei✉, *LIU Jun-wen*, *LUO Xiong-lin*

Department of Automation, China University of Petroleum, Beijing 102249, China
✉ Corresponding author, E-mail: liujw@cup.edu.cn

ABSTRACT　There are two challenges with the traditional encoder–decoder framework. First, the encoder needs to compress all the necessary information of a source sentence into a fixed-length vector. Second, it is unable to model the alignment between the source and the target sentences, which is an essential aspect of structured output tasks, such as machine translation. To address these issues, the attention mechanism is introduced to the encoder–decoder model. This mechanism allows the model to align and translate by jointly learning a neural machine translation task. The whose core idea of this mechanism is to induce attention weights over the source sentences to prioritize the set of positions where relevant information is present for generating the next output token. Nowadays, this mechanism has become essential in neural networks, which have been researched for diverse applications. The present survey provides a systematic and comprehensive overview of the developments in attention modeling. The intuition behind attention modeling can be best explained by the simulation mechanism of human visual selectivity, which aims to select more relevant and critical information from tedious information for the current target task while ignoring other irrelevant information in a manner that assists in developing perception. In addition, attention mechanism is an efficient information selection and widely used in deep learning fields in recent years and played a pivotal role in natural language processing, speech recognition, and computer vision. This survey first briefly introduces the origin of the attention mechanism and defines a standard parametric and uniform model for encoder–decoder neural machine translation. Next, various techniques are grouped into coherent categories using types of alignment scores and number of sequences, abstraction levels, positions, and representations. A visual explanation of attention mechanism is then provided to a certain extent, and roles of attention mechanism in multiple application areas is summarized. Finally, this survey identified the future direction and challenges of the attention mechanism.

KEY WORDS　attention mechanism; global/local attention; hard/soft attention; self-attention; interpretability

收稿日期: 2021−01−30
基金项目: 中国石油大学(北京)科研基金资助项目(2462020YXZZ023)

随着深度学习领域的发展, 注意力机制在计算机视觉和自然语言处理等领域取得了长足发展. 注意力机制的广泛应用始于机器翻译领域, 目前已成为神经网络中的一个重要概念, 不仅仅是从属概念, 已然发展成独立的注意力网络[1].

神经网络中注意力机制的快速发展具有如下三点优势:

(1)有效克服循环神经网络(Recurrent neural network, RNN)的一些挑战, 例如随着输入长度的增加, 预测性能下降和输入顺序处理导致的计算效率低下; 在机器翻译中源语言和目标语言之间对齐以及大范围长期依赖学习问题.

(2)可广泛用于提高神经网络的可解释性, 而神经网络又被视为黑盒模型. 这是一个显著的好处, 主要是因为人们对影响人类生活的应用中机器学习模型的公平性、问责制和透明度有越来越多的渴求, 而注意力机制在一定程度上可以提供可视化解释.

(3)很明显的优势就是直接提高了模型性能, 使得这些模型的预测推理结果最先进的, 不管是用于机器翻译、回答问题、情绪分析、对话系统, 还是图像视觉等多项任务[2], 这也是注意力机制广泛得到应用的根本推动力.

鉴于注意力机制的理论意义、所蕴含的应用价值以及可观的发展潜力, 本文对注意力机制的研究进展进行了系统性的综述, 为进一步深入研究注意力机制、开发注意力机制应用潜力确立良好的基础. 文中首先在第一节对注意力机制进行了概述以及问题的数学定义, 并在第二节着重对注意力机制进行分类及归纳, 从五个方面给出了注意力机制的不同描述. 第三节阐述了目前注意力机制对神经网络的可解释性的讨论, 第四节介绍了注意力机制的应用场景, 第五节给出了注意力机制未来发展方向, 最后一节则对注意力机制进行了总结.

1　注意力机制数学表述

为了方便, 采用 Bahdanau 等[3] 神经机器翻译(Neural machine translation, NMT)中的解码器-编码器结构来描述注意力机制. 传统的编码器框架有两个众所周知的挑战. 首先, 编码器必须将所有输入信息压缩成一个固定长度的向量, 然后将其传递给解码器. 使用一个固定长度的向量压缩输入序列可能会导致信息丢失[4]. 其次, 它无法对输入和输出序列之间的对齐关系进行建模, 这是翻译或摘要等结构化输出任务的一个重要方面. 为了解决这个问题, 在编码器–解码器体系结构引入了注意力机制, 如图 1 所示.

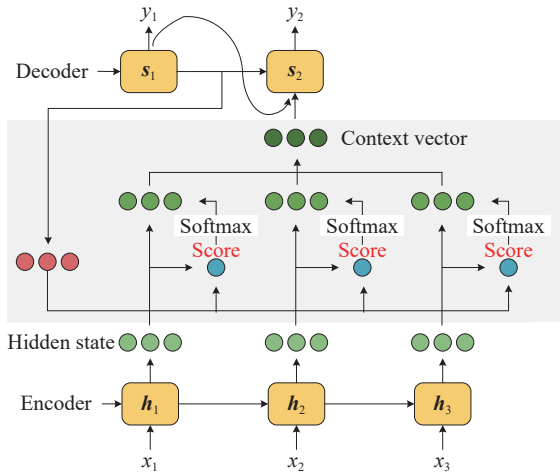

图 1　带有注意力机制的 Seq2Seq 结构经典编码器–解码器网络[3]

Fig.1　Seq2Seq structure of a classic encoder–decoder network with an attention mechanism[3]

假设源序列 $\boldsymbol{x}=[x_1,x_2,\cdots,x_n]$与目标序列 $\boldsymbol{y}=[y_1,y_2,\cdots,y_n]$, 源序列经过双向循环神经网络后输出两个不同方向的隐状态向量: 前向隐状态$\boldsymbol{h}_i'$和后向隐状态$\boldsymbol{h}_i''$, 然后将两者进行拼接来表示编码器的隐状态$\boldsymbol{h}_i=[\boldsymbol{h}_i';\boldsymbol{h}_i'']$. 在解码器经过位置$t$时, 通过$\boldsymbol{s}_t=g(\boldsymbol{s}_{t-1},y_{t-1},\boldsymbol{c}_t)$计算得出每个单词的隐状态向量, 其中$g(\cdot)$为计算隐状态向量的函数、上下文向量$\boldsymbol{c}_t$是输入序列的隐状态$h_i$之加权, 其中权重由对齐函数确定:

$$\alpha_{t,i}=\text{align}(y_t,x_i)=\frac{\exp(\text{score}(\boldsymbol{s}_{t-1},\boldsymbol{h}_i))}{\sum_{j=1}^{n}\exp\left(\text{score}\left(\boldsymbol{s}_{t-1},\boldsymbol{h}_j\right)\right)} \tag{1}$$

这里的对齐函数实际上为每个位置i的输入单词和位置t的输出单词(y_t,x_i)赋予一个分数, 衡量它们之间的匹配度.

2　注意力机制分类

2.1　软注意力机制与硬注意力机制

2.1.1　共同框架

2015 年, Xu 等[5] 受机器翻译和对象检测工作的启发引入了一种基于注意力机制的模型, 它自动学习描述图像内容的文字. 文中使用了两种不同的模型: 硬随机注意力和软确定性注意力. 首先都使用卷积神经网络来提取一组称之为注释向量的特征向量$\boldsymbol{f}=\{f_1,f_2,\cdots,f_L\}$, 分别对应于图像的部分区域, 这里, L 为图像区域划分的个数, 然后定义一个机制ϕ从注释向量中计算出上下文向量ζ_t,

对于每个位置，该机制都能产生一个权重α_i. 这里ϕ函数的定义就决定了如何将位置信息和权重信息结合.

2.1.2 硬注意力机制

在硬注意力机制中，权重$\alpha_{t,i}$所扮演的角色是图像区域a_i在时刻t被选中作为输入编码器信息的概率，有且仅有一个区域会被选中. 为此，引入位置变量$s_{t,i}$，当区域i被选中时取值为 1，否则为 0，即$p(s_{t,i}=1|s_{j<t},a_i)=\alpha_{t,i}$，然后计算上下文向量$\zeta_t=\sum_i s_{t,i}f_i$.

整个硬注意力机制是一个随机模型，会采样输入的隐状态，而不是整个编码端的隐状态，算出单词出现在某个位置的条件后验概率. 为了实现梯度的反向传播，需要采用蒙特卡洛采样的方法来逼近目标函数的梯度.

2.1.3 软注意力机制

相比之下，权重$\alpha_{t,i}$所扮演的角色是图像区域a_i在时刻t的输入编码器的信息中的所占的比例. 软注意力机制可以通过计算一个加权注释向量，直接得到上下文向量ζ_t的数学期望，从而构造一个确定性注意力机制模型，即$E_{p(s_t|a)}[\zeta_t]=\sum_{i=1}^{L}\alpha_{t,i}f_i$.

这相当于在系统中加入了加权上下文向量. 整个模型在确定性软注意力机制下是光滑的、可微的，因此使用标准的反向传播过程可以实现端到端的学习. 在此之前，大部分的传统注意力机制都属于软注意力机制. 软注意力机制是可以直接求梯度的，能直接代入到模型中去，整体进行训练. 所求的梯度可以经过注意力机制模块，反向传播到模型其它部分. 两种注意力机制模型都有好有坏，但目前主流的研究和应用还是更倾向于使用软注意力机制，因为其可以直接求导，进行反向传播.

2.2 全局和局部注意力机制

2.2.1 共同框架

2015 年，Luong 等[6]提出了全局和局部注意力两种简单有效的注意机制，其中全局注意力机制能顾及到输入语言的所有源语言单词，局部注意力机制则只能一次查看源语言单词的一个子集，如图 2 所示. 二者的区别在于注意力被放在所有的源语言位置上还是仅放在部分源语言位置上. 这两个模型的共同点是，在解码过程中，每个时刻都是先将源语言输入到堆叠长短时记忆网络（Long-short term memory, LSTM），计算源语言的各个隐状态对应当前目标语言隐状态对应的上下文向量，得到目标语言隐状态. 这样做的目的是为了得到上下文向量，进而用源语言句子的信息来帮助预测当前目标语言单词.

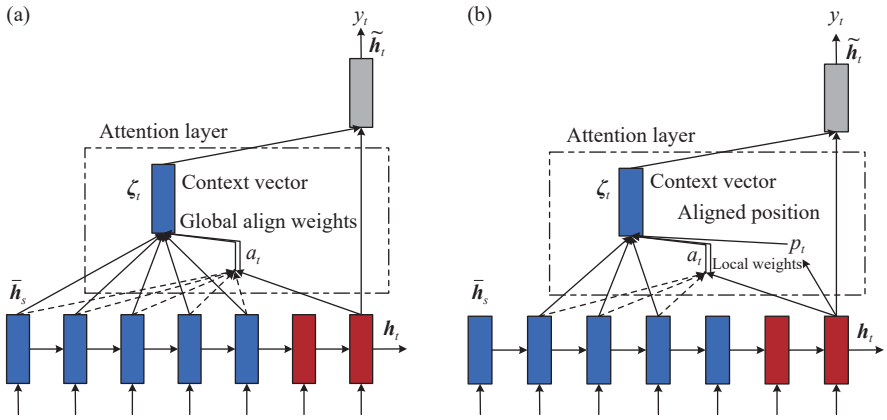

图 2 两种简单有效的注意机制[6]. (a)全局注意力，对每一步隐状态都计算了注意力值；(b)局部注意力，只对部分范围的隐状态进行注意值的计算

Fig.2 Two simple and effective classes of attention mechanism[6]: (a) a global approach that always attends to all source words; (b) a local approach that only looks at a subset of source words at a time

2.2.2 全局注意力机制

全局注意力机制在生成上下文向量时考虑编码器的所有隐状态. 在这个模型中，通过将当前目标隐状态$\boldsymbol{h}_t$与每个源隐状态$\bar{\boldsymbol{h}}_s$进行比较，得到一个可变长度的对齐向量$\boldsymbol{\beta}_t$，其大小等于源语言端输入句子的长度. 把对齐向量作为权重，通过对源语言隐状态的加权平均得到上下文向量. 在每个时间节点，模型根据当前目标语言隐状态$\boldsymbol{h}_t$和所有的源语言隐状态$\boldsymbol{h}_s$得出一个变长对齐权重向量. 然后对所有源状态的加权平均计算出全局上下文向量. 图 2 中，$\tilde{\boldsymbol{h}}_t$为最终计算得到的经过注意力加权后的全局上下文向量.

与 Bahdanau 模型相比，Luong 等提出的全局注意力模型在本质上相似，但是也有几个重要的

不同点, 此模型中在编码和解码器中都只用了 LSTM 顶层的隐状态, 而前者在双向编码器中用了前向和反向源语言隐状态的级联, 在非堆叠单向解码器中使用了目标隐状态.

2.2.3 局部注意力机制

全局注意力机制有一个缺点, 其对于每一个目标单词都要考虑源语言句子中的所有单词, 此过程算法复杂性太大, 并且不太可能翻译长序列. 而局部注意可以克服这种问题, 针对每个目标单词, 其只关注小部分的源语言子句子.

在时刻t, 模型首先针对每个目标单词生成一个对齐位置υ_t. 针对对齐位置υ_t如何确定, 此模型有两种变体: 单调对齐(local-m)和预测对齐(local-p). 前者简单地设$\upsilon_t = t$, 假设源语言序列和目标语言序列大体上单调对齐; 而后者不假设源语言序列和目标语言序列单调对齐, 模型按照以下方式预测对齐位置: $\upsilon_t = L \cdot \mathrm{sigmoid}(\boldsymbol{v}_\upsilon^{\mathrm{T}} \tanh(\boldsymbol{W}_\upsilon \boldsymbol{h}_t))$. 这里 L 是源语句长度, $\boldsymbol{W}_\upsilon$和v_υ是将要被学习用来预测位置的模型参数. 为了更偏向于υ_t附近的对齐点, 设置了一个以υ_t为中心的高斯分布模拟对齐程度. 高斯分布重新定义的对齐权重如下: $\beta_t(\upsilon_t) = \mathrm{align}(\boldsymbol{h}_t, \overline{\boldsymbol{h}}_s)\exp\left(-\dfrac{(\mu-\upsilon_t)^2}{2\sigma^2}\right)$, 其中标准差为$\sigma = D/2$, D 是凭经验选取的一个常数, υ_t是一个实数, 而μ是一个在以υ_t为中心的窗口内的整数.

与 Bahdanau 等相比, 其使用了与ζ_t相似的上下文向量来构造后续隐状态, 虽然也能达到“覆盖”效果, 但其没有分析这种连接是否有效. 此处的模型更具有通用性, 模型可应用于常规堆栈循环结构, 包括非注意力模型.

2.3 分层注意力机制

2.3.1 层次注意力机制

Yang 等[7] 最早把注意力分层的思想用于文档分类, 而且引入层次注意力(Hierarchical attention), 除了提高模型的精确度之外还可以进行单词与单词之间、句子与句子之间重要性的分析和可视化. 正如其名, 层次注意力机制构造了两个层次的注意力机制结构. 第一个层次是对句子中每个单词的注意力机制, 并非所有的单词对句子含义的表示, 都有同样的贡献. 因此, 引入注意机制来提取这些关键词, 这对于单词在句子中起的作用来说, 是很重要的选择和判断标准, 而且还汇总了这些表示形成句子向量的各种有价值的信息词汇. 第二个层次是针对文档中每个句子的注意力机制, 与单词级别类似.

层次注意力机制主要思想是: 首先从文档的分层结构出发, 单词组成句子, 句子组成文档, 所以自然而然建模时也分这两个层次进行. 其次, 不同的单词对句子理解和不同的句子对于文本理解和分类, 具有不同的信息量和关注度, 不能单纯均匀对待, 所以引入分层注意力机制, 分层注意力机制让我们对文本分类的内部机制有一定的白箱理解.

2.3.2 自顶向下注意力机制

Zhang 等[8] 在卷积神经网络中提出了基于自顶而下神经注意力(Top-down neural attention), 使神经网络在学习过程中的注意力更加有针对性, 其实就是层次化注意力结构变体形式, 而这种实现也十分贴近我们真正的生物视觉机制, 具有十分重要的生物神经学理论依据. 为了实现这种自顶而下神经注意力机制, 采用了一种泛化的确定性赢者通吃(Winner-Take-All)的方法, 从而可以选择出与这个自顶而下信号最相关的神经元.

同时还提出了一个基于概率性的赢者通吃公式来建立自顶而下的层次化神经注意力机制的卷积神经网络(Convolutional neural networks, CNN)分类器模型, 将确定性方法泛化到了概率性版本, 使得学出来的注意力映射不再是二值结构. 这种注意力映射其实也可以叫做软注意力映射, 它的好处也很明显, 就是可以去捕捉更加细微的一些特征和变化等等. 基于赢者通吃假设还提出了一种改进的传播方法, 可以有效地计算注意力上下文向量, 得出每个神经元赢得可能性的边缘概率, 并且通过网络中的反向传播误差对比自上而下的信号的重要性.

2.3.3 多步注意力机制

2017 年 5 月, Gehring 等[9] 在机器翻译任务中提出了完全基于 CNN 构造序列到序列模型, 文中提出的多步注意力机制(Multi-step attention)通过该注意力结构来获取编码器和解码器中输入句子之间的关系. ConvS2S 模型在翻译任务上不仅仅效果显著, 而且所需训练时间也很短. 多步注意力机制实际上也是一种分层注意力机制, 它在解码器的每一层, 都单独使用了注意力机制.

该模型通过堆叠多层注意力机制来获取输入句子中单词与单词之间的依赖关系, 特别是当句子非常长的时候, 实验证明层叠的层数往往达到 10 层以上才能取得比较理想的结果. 针对每一个卷积步骤都对编码器的隐状态和解码器的隐状态进行点积得到注意力矩阵, 并且

基于最终的注意力矩阵去指导解码器的解码操作.

2.3.4 多头注意力机制

递归神经网络，特别是 LSTM 和门控循环神经网络是解决语言建模和机器翻译这种序列建模和转换问题的先进方法. Vaswani 等[10]提出了一种新的框架，被称作 Transformer，其与以往的模式不同，并没有用任何 CNN 或者 RNN 的结构，而是完全依赖注意机制来表示输入和输出之间的全局依赖关系.

模型中的注意力可以描述为将一个查询向量（Queries）$\boldsymbol{Q}$ 和一组键−值对（Key-Value）$\boldsymbol{K}$ 和 $\boldsymbol{V}$，映射为一个输出. 输出是由值的加权和得到的，每个值的权重是根据查询向量和相应的键通过一个对齐函数计算出来的，计算公式为：$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d_k})\boldsymbol{V}$，$d_k$为查询向量的维数.

多头注意力（Multi-head）则是用不同的、需要学习的线性映射，对查询向量，键及值进行多次变换，然后分别对每一个映射之后得到的查询向量、键及值，再进行上述多个单头注意力的并行运算，进而生成多个输出值，然后拼接起来成为高维向量，并再次映射，进而得到最终值. 计算公式为：$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(\text{head}_1,\cdots,\text{head}_h)\boldsymbol{W}_{\text{head}}$，其中 $\text{head}_i = \text{Attention}(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i)$，$\boldsymbol{W}_{\text{head}}$为多头注意力的权重，下标$h$为多头注意力的个数.

2.4 多维自注意力机制

2.4.1 通用结构

Shen 等[11]在 additive 注意力在每个 token 的特征层的做出推广，叫做多维注意力，图 3 刻画了经典注意力与其区别，图中 d_e 表示神经网络的输入层神经元个数. 多维注意力没有对每个 token 嵌入向量$\boldsymbol{x}$的分量x_i计算一个标量得分值，而是对 x_i 中每个分量计算了一个向量得分值，即利用权重矩阵$\boldsymbol{W}$，查询向量 $\boldsymbol{q}$ 与偏置b：$\kappa(x_i,\boldsymbol{q}) = \boldsymbol{W}^{\mathrm{T}}\sigma(\boldsymbol{W}^{(1)}x_i + \boldsymbol{W}^{(2)}\boldsymbol{q} + b^{(1)}) + b$. 在此基础上，作者定义了两种形式的自注意力机制：source2token 型和 token2token 型. 前者用于计算每个x_i与整个句子的相关性，将整句压缩为一个向量，Lin 等[12]率先把这种自注意力机制引入自然语言处理某些任务中的句子嵌入表示，在数学表达上即去掉对齐函数$\kappa(x_i,\boldsymbol{q})$中 $\boldsymbol{q}$ 有关的项；而后者将对齐函数$\kappa(x_i,\boldsymbol{q})$中的 $\boldsymbol{q}$ 换为x_j，下面介绍后者的几种变体形式.

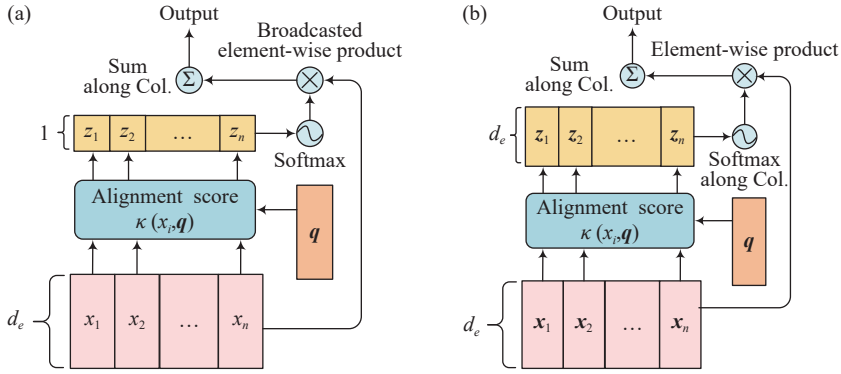

图 3 经典注意力机制（a）和多维注意力机制（b）[11]. $z_i \in \{z_1, z_2, \cdots, z_n\}$为计算对齐函数$\kappa(x_i,\boldsymbol{q})$得到的对应值，图（a）中$z_i$为标量，图（b）中其值$z_i$为向量

Fig.3 Traditional (additive/multiplicative) attention (a) and multi-dimensional attention (b)[11]. z_i denotes the alignment score$\kappa(x_i,\boldsymbol{q})$, in figure (a) z_i is a scalar and in figure (b) z_i is a vector

2.4.2 方向型自注意力机制

Shen 等[11]提出了基于掩码的 token2token 型多维自注意力，称为方向型自注意力（Directional self-attention, DiSA）. 方向型自注意力机制考虑了单词之间的依赖和时序关系，并融合了自注意力模块的输入和输出. 主要做了以下两个修改：把权值矩阵$\boldsymbol{W}$成了常数 c，把 sigmoid 激活函数换成了 tanh 激活函数；使用了位置掩码矩阵，使得两元素之间的注意力矩阵是不对称的. 使用掩码很容易对结构丰富的先验知识编码，比如时序关系和稀疏依赖关系编码.

2.4.3 双向分块自注意力机制

传统自注意力主要缺点在于需要很大的存储空间存储所有元素对的对齐，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 等[13]又提出了一种双向分块自注意力机制（Bidirectional block self-attention, Bi-BloSA），自下而上可分为三个主要部分：分块内的注意力机制、分块间的自注意力机制、上下文融合，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，这种模型具有高度的可并行运算性，同时对局部和远距离相

关性进行了良好的建模, 在多种自然语言处理任务下达到最优效果.

2.4.4 强化学习自注意力机制

软注意力机制在建模句子的局部或全局依赖关系的时候, 有前景, 但是其计算效率低; 而硬注意力机制虽然直接有效, 但是其不可微分. 所以 Shen 等[14] 又将硬注意力机制和软注意力机制在强化学习的方法下进行巧妙融合, 提出了一种强化型自注意力(Reinforcement self-attention, ReSA)模块, 这样就能使用硬注意力机制处理长句子依赖问题, 用策略梯度进行学习, 同时也能为软注意力机制选择子集进行计算; 而软注意力机制的前馈信号反过来用来对硬注意力机制提供奖励信号, 进行更精细化的操作, 同时为硬注意力机制提供指导.

2.5 结构化自注意力机制

Kim 等[15] 利用图模型将经典的注意力机制进行拓展, 考虑了深度神经网络结构上的依赖, 使得注意力机制从普通的软选择变成了既能内部结构建模信息又不破坏端到端训练的新机制. 在结构注意力模型中, 没有对单输入式的选择建模, 而是对连续, 子序列多输入式的选择建模. 在这种情况下, 注意力机制需要引入n个离散的二元隐变量$\boldsymbol{o}=(o_1,o_2,\cdots,o_n)$, o_n表示给定输入元素是否包括在对应的子序列中. 另外, 注意力分布$p(\boldsymbol{o}|x,\boldsymbol{q})$用线性链式条件随机场(Linear chain CRF)来刻画每个o_n之间的依赖关系, 以此确定结构信息变量$\boldsymbol{o}$的依赖结构. 然后使用图模型推断前向传递期望值和上下文向量.

分片注意力层中可以选择源句子中的子序列, 而不再是经典注意力机制中 token 为单位. 其次, 在语法树结构的建模中引入合成注意力层. 这两种新的注意力层, 在多个任务上都取得了比经典注意力更好的结果. 虽然结构化关注背后的基本思想很有洞见, 但作者坦诚在实践中可能难以使用. 一个问题是, 从计算的角度来看, 在使用图形模型时数值稳定性往往差, 因此计算在对数空间中执行更好, 这会给模型代码增加相当大的复杂性. 另一个问题是, 简单应用现有的自动微分工具往往效率低下, 为了使结构化注意易于处理大问题, 通常需要手工编写梯度计算.

2.6 注意力总结

2.6.1 概述

以上对常用的注意力机制做出了详细的分类, 注意机制基本上可以分为两大类: (互)注意力机制和自注意力机制. 简单来说, 互注意力机制就是模拟源序列和目标序列中不同位置之间的关系, 而位置个数的选择产生了各种丰富的注意力机制; 而自注意力机制就是模拟相同输入的不同位置之间的关系, 也就是把互注意力机制中的目标序列替换成源序列即可, 在 Transformer 模型中使用了大量的自注意力机制. 为了使得研究者更好地使用注意力机制, 下面首先对注意力机制从对齐函数方面进行分类(表 1).

除了上述方式, Chaudhari 等[16] 提出了从四个不同角度对注意力机制进行分类, 分别是序列个数, 抽象层个数, 位置个数, 以及表示个数(见 2.6.6 节). 作者强调上面这些类别并不相互排斥, 可以交叉重叠, 某种注意力机制可以归属多个不同的分类, 可以作为多个类别的组合来混合使用.

表 1 几种常用的注意力机制及其对齐函数的数学形式

Table 1 Summary of several attention mechanisms and corresponding alignment score functions

Name of attention mechanism	Alignment score functions	References
Content-base attention	$\text{score}(\boldsymbol{s}_t,\boldsymbol{h}_t)=\text{cosine}[\boldsymbol{s}_t,\boldsymbol{h}_i]$	[17]
Additive attention	$\text{score}(\boldsymbol{s}_t,\boldsymbol{h}_t)=\boldsymbol{v}_a^{\mathrm{T}}\tanh(\boldsymbol{W}_a[\boldsymbol{s}_t;\boldsymbol{h}_i])$	[3]
Location-base attention	$\alpha_{t,i}=\text{softmax}(\boldsymbol{W}_a\boldsymbol{s}_t)$	[6]
Bi-linear attention	$\text{score}(\boldsymbol{s}_t,\boldsymbol{h}_t)=\boldsymbol{s}_t^{\mathrm{T}}\boldsymbol{W}_a\boldsymbol{h}_i$	[6]
Dot-product attention	$\text{score}(\boldsymbol{s}_t,\boldsymbol{h}_t)=\boldsymbol{s}_t^{\mathrm{T}}\boldsymbol{h}_i$	[6]
Scaled dot-product attention	$\text{score}(\boldsymbol{s}_t,\boldsymbol{h}_t)=\boldsymbol{s}_t^{\mathrm{T}}\boldsymbol{h}_i/\sqrt{n}$	[10]

2.6.2 对齐函数

对齐函数是衡量输入和输出匹配度的函数, 在不同的注意力类型中, 对齐函数采用了不同的数学形式, 效果也不尽相同, 在设计具体注意力网络时可以提供不同的选择. 表 1 总结了几种常用的注意力机制及其对齐函数的数学形式. 主要有基于内容的(Content-base), 基于加和的(Addition), 基于位置信息的(Location-base), 基于双线性的(Bilinear), 基于点积的(Dot product), 基于比例点积(Scaled dot-product)的注意力机制类型. 表 1 中$\boldsymbol{W}_a$是神经网络权值矩阵, $\boldsymbol{v}_a^{\mathrm{T}}$为反向传播过程训练的参数.

2.6.3 序列个数

序列个数指的是输入序列的个数和输出序列的个数. 如果输入序列和对应输出序列都只有一个, 称之为一对一(Distinctive)型注意力机制. 用于翻译的大多数注意力模型[3]、图像描述[5] 和语音识别[18] 这些都属于一对一类型的注意力机制. 如

果模型有多个输入序列，并且学习不同输入序列之间的权重矩阵，以捕获这些输入序列之间的关系，把这类称之为协同注意力机制（Co-attention）. 协同注意力机制典型的应用场景有：阅读理解，对输入（问题回答，文本）之间的注意力建模，找出对回答问题最相关的问题关键词，协同注意力机制非常有助于同时检测问题中的关键词和答案相关文章的段落. 还有一类是自注意力机制（Self-attention），一般输入是序列，输出不是，比如分类和推荐任务，在此场景中，可以使用注意力来学习输入序列中对应于相同输入序列中的每个标记的相关标记.

2.6.4 抽象层个数

在最一般的情况下，只为原始输入序列计算注意力权重，这种注意力称为单级的（Single-level）. 另一方面，注意力可以按顺序应用于输入序列的多个抽象层次. 较低抽象级别的输出（即上下文向量）成为较高抽象级别的查询状态，这种类型称之为多级（Multi-level）.

上面提到的层次注意力就是典型的多层次抽象，在两个不同的抽象层次（即单词级别和句子级别）上使用了注意力模型来完成文档分类任务，因为它捕获了文档的自然层次结构，即文档由句子组成，句子由单词组成.

2.6.5 位置个数

这里的位置个数指的是参与计算上下文向量的隐状态向量个数. Bahdanau 等[3] 介绍的注意力机制也被称为软（soft）注意力. 顾名思义，它使用输入序列所有隐状态的加权平均值来构建上下文向量. 软加权虽然使得神经网络易于通过反向传播进行有效的学习，但是也增加了计算成本. 如果这种加权的权值变成只有一个 1，其余全是 0，也就是此时的上下文向量是随机采样的某个隐状态，这种称之为硬（Hard）注意力，大大减少了计算量，但是训练过程不可微分，难以优化.

Luong 等[6] 在机器翻译中提出了局部（Local）和全局（Global）注意力. 全局注意力类似于软注意力. 另一方面，局部注意力介于软注意和硬注意之间. 关键思想是首先检测输入序列中的一个注意点或位置，然后在该位置周围选择一个窗口，创建一个局部软注意力，此时的隐状态的个数就是窗口的大小. 输入序列中的位置可以设置单调对齐或通过预测对齐学习. 因此，局部注意力的优点是在软注意和硬注意、计算效率和窗口内的可微性之间提供参数权衡，因此，为了克服这一局限性，提出了变分学习方法和强化学习策略梯度方法.

2.6.6 表示个数

大多数情况下，神经网络都是使用输入序列的单一特征表示，但是在某些场景，使用输入的一个特征表示可能不足以满足下游任务，需要注意力来为这些不同的表示分配权重，这些表示可以确定最相关的方面，而忽略输入中的噪声和冗余信息. 典型地，在自然语言场景中，Kiela 等[19] 学习了同一输入句子的不同单词嵌入表示的注意力权重，以改善句子表示，同时通过权重的可解释性，确定哪些单词嵌入对句子的贡献度的大小. 类似地，Maharjan 等[20] 使用注意力机制来动态给书籍的不同特征表示赋权，捕捉词汇、句法、视觉和类型等不同层面的信息.

还有一种情况，就是引入权重来确定输入嵌入向量的各个维度分量的相关性，计算向量的每个特征的分数可以选择在任何给定上下文中最能描述标记特定含义的特征. 这对于自然语言应用程序来说尤其有用，因为在自然语言应用中，传统的单词嵌入表示会受到一词多义问题的影响. Lin 等[12] 和 Shen 等[11] 针对语言理解问题给出了这种方法的例子，以获得更有效的句子嵌入表示. 在此项类型的分类中把以上两种多元特征表示的形式统一称为多表示（Multi-representational）. 表 2 总结了近几年注意力机制的应用文献.

3 注意力机制的可解释性

可解释性是指人类能够理解决策结果的原因的程度，模型可解释性指对模型内部机制的理解以及对模型结果的理解. 近年来受到模型的性能以及透明度和公平性的推动，人工智能模型的可解释性引起了人们的极大兴趣. 然而，神经网络虽然在大部分任务表现良好，但是因为是黑盒模型，缺乏可解释性，大大削弱了工业应用上对模型所做的决定或预测的理解，而注意力机制的引入可以直觉地窥探神经网络内部的运行机制：对一个给定的输出，可以通过检查注意力机制权重，得知模型分配了较大注意力权重的输入是哪一个.

假设注意力权重的重要性与序列中每个位置的输出的预测值和输入对象的特定区域的相关程度高度相关，那么可以通过可视化一组输入和输出对的注意权重来增强对模型结果的理解，这种理解是否符合人类的思考逻辑值得商榷. 在自然语言处理中，研究者普遍认为，注意力机制为神经

表 2 重要的注意力机制模型从四个不同方面的总结

Table 2 Summary of key papers for technical approaches within each category

References	Number of sequences	Number of abstraction levels	Number of representations	Number of positions	Scenario of applications
[3]	Distinctive	Single-level	Single-representational	Soft	Machine translation
[5]	Distinctive	Single-level	Single-representational	Hard	Image captioning
[6]	Distinctive	Single-level	Single-representational	Local	Machine translation
[7]	Self-attention	Single-level	Single-representational	Soft	Document classification
[18]	Distinctive	Multi-level	Single-representational	Soft	Speech recognition
[21]	Distinctive	Single-level	Single-representational	Soft	Visual question answering
[22]	Co-attention	Multi-level	Single-representational	Soft	Sentiment classification
[23]	Self-attention	Multi-level	Single-representational	Soft	Recommender systems
[11]	Self-attention	Single-level	Multi-representational	Soft	Language understanding
[19]	Self-attention	Single-level	Multi-representational	Soft	Text representation

模型的工作方式提供了一种重要的解释方式.

注意力机制实现可解释性, 已经广泛应用在各种学习场景. Bahdanau 等[3] 在机器翻译领域引入注意力机制, 解决了大范围序列依赖建模问题, 同时也对源语言英语和目标语言法语之间的自动对齐问题, 给出了可视化解释, 即使对于两种不同结构和文法的语言, 不同语言间相互对应的单词之间的注意力权重明显更大. 在机器阅读理解中, 注意力机制模型已经成为网络结构中必不可少的一环, 机器阅读理解通过结合文本和问句两者的信息, 生成一个关于文本段落各个部分的注意力权重, 对文本信息进行加权, 试图通过其去捕捉问题和文本篇章之间的匹配关系. 而后来提出的协同注意力机制是一个双向的注意力机制, 不仅要给阅读的文本段落生成一个注意力权重, 还要给问句也生成一个注意力权重. 最后, Xu 等[5] 对图像文字描述任务中, 生成的文本与相关图像区域关联关系, 进行可视化.

实际上, 除了上述应用场景的可解释性研究, 还有很多工业场景下, 对于注意力机制的可解释性研究文献. De-Arteaga 等[24] 研究了社会职业分类中性别偏见, 并分析了这种偏见主要与哪些词汇相关, 被注意的词汇如何解释性性别偏见. 作为注意力机制的另一个有趣的应用, Lee 等[25] 和 Liu 等[26] 发布开源工具, 用于可视化深度神经网络的注意力权重, 通过注意力权重注入扰动信号, 以便模拟特定假设情景, 并交互式观察深度神经网络预测值的变化, 侦测注意力权重是否存在某种相关解释性.

注意力机制已被应用在各种各样的学习场景中, 几乎普遍存在, 不免让很多研究人员对注意力机制能否解释模型预测提出了疑问. Jain 等[27] 认为注意力机制并不能提高模型的可解释性. 如果注意力机制能提供解释, 那么必须满足以下两个性质: (a) 服从特定概率分布的注意力权值, 权值的大小, 必须与特征重要性度量值相关; (b) 如果训练的注意力概率分布发生改变或变换, 那么预测结果也应该发生相应的 (comparable) 变化. 并给出了两组实验对比来验证自己的观点, 首先假设计算得到的对象的注意力权重和对象的特征重要性度量值之间不总是一致的, 也就是注意力机制只能为模型的预测提供微弱的解释; 接下来提出了一种替代性对抗注意力概率分布, 它可以最小程度地改变模型预测结果. 为此, 其控制训练好的模型的注意力权值所服从的概率分布, 来判别是否存在替代性分布使得模型输出接近原始预测值, 但是预测结果依然相同, 即使是注意到了不同的输入特征, 甚至随机置换注意力权重, 是否通常只会导致输出的微小变化. 其结果综合表明: 注意力权重基本上无法提高模型可解释性.

而 Wiegreffe 和 Pinter [28] 对以上结果提出了质疑, 认为 Jain 等所得到的结论[27] 依赖模型解释性的定义, 且对于模型的测试是否正确, 需要考虑模型的所有元素, 使用更加严谨的实验设计过程. 认为违反事实的注意力权重实验, 无法 Jain 等自身的论点, 首先其所提取的注意力权值所服从的概率分布不是原始的 (Primitive), 是分离了模型各部分而获得的注意力权重, 与模型整体的依赖度会降低; 还有就是注意力重要性分数可以提供可解释性, 但不是唯一的可解释性, 取决于每个人对模型解释性所作的定义.

4 注意力机制的应用

4.1 计算机视觉方面的应用

在图像分类方面，Mnih 等[17] 为了解决在高分辨率图片上使用卷积神经网络时，计算复杂性高的问题，在传统的 RNN 上加入了注意力机制进行图像分类，即在高分辨图片或者视频帧上自适应地提取一系列的区域框，然后从被选区域提取图片或视频信息. Jetley 等[29] 提出了一种用于图像分类的 CNN 架构的端到端可训练注意力模块. 该模块将二维特征矢量图作为输入，其形成 CNN 流水线中不同阶段的输入图像的中间表示，并输出每个特征图的得分矩阵. 通过结合该模块来修改标准 CNN 架构，并且在约束下训练中间 2 维特征向量的凸组合单独用于分类. Sharma 等[30] 针对视频中的动作识别任务提出了一种基于软注意力的多层递归神经网络，在网络中加入关注区域的移动、缩放机制，连续部分信息的序列化输入，将目标作进一步精细化，让模型可以捕获更精细的特征，通过将特征分成更小的块，注意力机制将筛选出更有利于描述特征的那部分图像块.

图片生成任务通常使用深度神经网络来提取图片高层次特征，通过图片特征重构图像，然而从包含丰富内容的图片生成图片是很棘手的事情. 为了克服这一困难，Kataoka 等[31] 提出了一个基于注意力机制的生成网络，生成网络被训练用来关注图像的局部细节并逐步分阶段生成图像. 这使得网络能够处理图像的一部分和整个图像的粗略结构的细节，验证了通过注意力机制和生成对抗网络生成图像的有效性. Gregor 等[32] 用生成对抗网络和深度递归注意力写入器（Deep recurrent attentive writer, DRAW），实现图像的迭代构造过程，以便产生更逼真的图像. Parmar 等[33] 受卷积神经网络启发的 Transformer 变种提出了 Image Transformer，重点是局部注意范围，即将接受域限制为局部领域. 不过，这种模型有一个限制条件，即要以失去全局接受域为代价，以降低存储和计算成本.

在与图像有关的多模态领域，Huang 等[34] 提出了一种图像和文本双模态的神经网络翻译模型，探索了将文本和图像多模态信息集成到基于注意力机制的编码器–解码器结构中的方法. 在学习图像描述子的背景下也探讨了注意力机制的有效性. Zhang 等[35] 在自注意力机制层加入生成对抗网络，使得生成器和判别器更好地对空间区域之间关系进行建模. 模型结构采用了非局部神经网络，利用注意力机制进行计算，赋予感兴趣的区域更大的权重.

4.2 自然语言处理方面的应用

注意力机制在自然语言处理中有着巨大的应用潜力，特别是神经机器翻译等任务. 神经机器翻译任务中大多使用编码器–解码器的网络结构，这种结构有一个潜在的问题是，神经网络需要将源语句所有的信息压缩成固定长度的向量. 这可能使得神经网络难以处理长句子，尤其是那些比训练语料库中句子更长的句子，随着输入句子长度的增加，原始编码器–解码器的性能会迅速下降.

为了解决这一问题，Bahdanau 等[3] 引入了一种基于注意力机制的编码器–解码器扩展模型. 每当生成的模型在翻译中生成一个单词的时候，它会（软性地）搜索源句中最相关信息集的位置. 然后，该模型根据与源语句位置相关的上下文向量和之前产生的所有目标语言单词来预测下一个目标单词，改善源语言和目标语言的对齐问题. 之后，Luong 等[6] 提出了全局注意力机制和局部注意力机制两种注意机制. 全局注意机制在生成上下文向量时考虑编码器的所有隐状态，但是全局注意机制有一个缺陷，其针对每一个目标语言单词都要考虑源语言语句中所有单词，此过程计算复杂性很高. 局部注意机制可以克服这种问题，它对每个目标单词，只关注源语言句子中的小部分单词.

2016 年，Cohn 等[36] 扩展了注意力机制的神经机器翻译模型，包括基于对齐的文字结构偏差模型，直接将这些对齐误差信息引入注意力机制模型. Feng 等[37] 在注意力机制模型中应用了传统统计机器翻译的扭曲度（Distortion）和繁衍度概念（Fertility），认为当对齐不正确时，基于注意力机制的“编码器–解码器”模型的翻译质量严重下降，文中直接将前一时刻的上下文向量信息输入注意力模型，以帮助注意力模型更好地预测目标语言句子的词语顺序.

Eriguchi 等[38] 提出了一种新的端到端的句法 NMT 模型，利用源语言端的短语结构构造了一个序列到序列的翻译模型. 句法 NMT 模型利用句法解析树建立基于句法解析树的编码器，基于句法解析树的编码器是顺序编码器模型的自然扩展，编码器中句法解析树的路径，可以与其对应的顺序编码器一起工作. 此外，句法 NMT 模型引入了注意力机制，允许基于句法解析树的编码器不仅实现输入句子单词级的对齐，而且实现输入句子

短语结构与输出翻译句子短语结构的对齐.

Sankaran 等[39] 发现基于注意力机制的 NMT 模型, 会出现只记忆当前输入的注意力缺陷问题, 提出了一种称为时序注意力机制的模型, 该模型会记忆翻译过程中每一时刻目标语言单词和源语言单词之间的对齐信息, 并根据历史对齐信息对当前注意力机制进行调整. 由于自然语言之间复杂的结构差异, 单向注意力机制模型可能只抓住注意自然语言中存在的的部分规律. 于是, Cheng 等[40] 为了使注意力机制模型更加全面准确反映自然语言中存在的规律, 提出了联合训练的双向注意力机制模型, 该模型的中心思想是对于相同的机器翻译训练数据, 使源语言到目标语言和目标语言到源语言的两个翻译模型的对齐矩阵保持一致. 而不是独立地对源语言和目标语言进行源语言翻译模型的训练. 之后, Liu 等[41] 从重新排序的角度对注意力机制的对齐准确度问题进行分析, 提出将传统机器翻译中目标词语与源词语的对齐信息, 作为监督信号引入神经机器翻译训练过程, 利用该监督信号自动地引导注意力机制模型, 对注意力机制进行调整的方法.

之后, Vaswani 等[10] 提出一种著名的网络架构 Transformer, 在第二节做出了详细的介绍, 其模型架构避免了循环并完全依赖于注意力机制来绘制输入和输出之间的全局依赖关系, 并完全避免循环和卷积. 其结构允许进行更多的并行化, 在翻译质量上更加优越、并行性更好并且需要的训练时间显著减少. Britz 等[42] 和 Tang 等[43] 也同样在 NMT 使用注意力在性能方面做出了开放式的实验和评价.

还有, Yin 等[44] 为了探索注意力机制在整合句子之间的相互关系, 提出了两种基于注意力的卷积神经网络来建模句子对. Zhuang 等[45] 提出在文本分类引入分层注意力, 利用包含在句子中的重要单词学习句子表示, 然后利用文本中的重要句子学习文本上下文表示. Zhou 等[46] 在双向长短时记忆网络基础上加入注意力后, 可以实现对分类起决定性作用的词语的自动对焦, 提出的模型不使用 NLP 系统的任何特性. Wang 等[47] 和 Ma 等[48] 提出的基于刻面的情感分类方法将与刻面相关的概念的额外知识纳入模型, 并利用关注度来适当权衡概念与内容本身的区别, 学习输入文本的情绪标识.

4.3 语音识别方面的应用

注意力机制逐渐在语音识别[49−50] 领域中得到应用, 且效果极好. Chorowski 等[51] 为注意力机制提供了两种新的思路: 一种更好的标准化方法, 产生更平滑的对齐方式和一种通用的提取和使用先前路径特征的原则, 这两种方法都可以被应用于语音识别之外的应用场景.

Bahdanau 等[52] 把注意力机制模型用在了大规模词汇连续语音识别中, 在注意力机制的基础之上, 研究了一种更直接的方法, 用一个递归神经网络代替隐马尔可夫模型, 直接在字符级执行序列预测. 对于每个要预测的字符, 注意机制扫描输入序列并选择相关帧. 同时也指出注意力机制模型虽然有优点, 但是还是有自身的问题, 其适合短语识别, 对长句子识别比较差; 数据包含噪音的时候训练不稳定.

Shen 和 Lee[53] 发现语音序列太长的时候, 序列会包含噪音或者很多无关的信息. 通过强调序列中的重点部分, 注意力机制可以解决这个问题: 用 LSTM 去读取输入序列, 注意力机制选择序列中的重点部分, 通过突出重点语音子序列去预测序列的类标签, 该方法在关键词提取, 对话行为侦测两项任务中取得了很好的表现; Liu 和 Lane[54] 引入了编码器–解码器框架和注意力机制, 联合建模说话者意图识别和空缺值填充两个子问题, 将注意力机制加入到基于 RNN 的校准模型中去, 使编码器–解码器结构学会同时校准和解码, 能够在没有给定对齐信息时, 对不同长度的序列进行映射关系学习, 分辨说话者意图, 从自然语言对话重抓取语义成分, 简化了口语理解技术的建模.

5 注意力机制的未来方向

本文对注意力机制进行了全面的总结, 本文从神经学、心理学方面引出了注意力机制, 然后重点介绍了注意力机制的结构分类以及一些其它的结构. 最后简单地归纳了计算机视觉注意力机制, 详尽地总结了注意力机制在各种领域中发挥的作用. 在加入了注意力机制之后, 大部分深度学习算法的准确性都得到了提高, 模型的泛化能力得到改善. 尽管迄今报告的成果令人鼓舞, 但仍然存在一些不足和缺陷. 我们尝试提出注意力机制的未来研究方向如下:

(1)注意力机制作为一种特征选择机制, 实现了输出对输入的多个变量加权选择的过程, 寻找与当前输出最相关的特征, 按比例加权得到上下文向量作为输入去预测输出, 但是, 同一个数据作为输入和作为输出时, 起的作用应该是不一样的,

如何实现双向的注意力选择机制是一个值得研究的课题.

(2)注意力机制目前实现的上下文向量求解，还没有完全考虑输入向量各分量的结构关系，实际上，实际的输入数据上总存在一些结构关系，可以加以利用，使得特征选择具有结构关系，也就是注意力机制的上下文向量应该具有结构关系，比如稀疏结构，树结构[55−56]等，如何考虑这些因素，更好地利用先验信息是一个值得尝试的方向.

(3)注意力机制目前主要应用在深度神经网络上，应该拓展注意力机制的使用范围，比如推广到逻辑斯蒂回归和分类，支持向量机回归和分类模型中去.

(4)注意力机制目前主要应用在监督学习上，应该考虑无监督学习和强化学习场景下的注意力机制，探索无监督聚类，降维，主成分分析，矩阵因式分解上的注意力实现机制，探索动态规划，赌博机在线学习，蒙特卡洛抽样学习，暂态差学习，行动者−批评家强化学习场景下的注意力实现机制.

(5)目前的注意力机制，只实现输入特征上的注意力选择，应该探讨实现输入特征组，输入特征矩阵，流数据输入特征上的注意力实现机制.

(6)最近出现的胶囊网络[57]与注意力机制有着密切关系，研究胶囊网络和注意力机制的区别和联系，可能会得到更大的启示.

(7)研究离散动态模型和连续动态模型混合的多智能体混杂模型中，多智能体混杂模型的混合过程由事件驱动，在不同的事件场景中，注意力机制的切换机制.

6 结论

鉴于深度学习中的注意力机制的理论意义和实际应用价值，本文对深度学习中的注意力机制进行了系统的综述. 本文首先介绍了注意力机制的认知心理学起源与技术动机，然后着重对注意力机制的分类进行浓墨重彩的描述和提纲挈领式的总结归纳，主要对软注意力机制和硬注意力机制，全局注意力机制和局部注意力机制层次注意力机制以及多种自注意力机制等典型注意力机制进行了详细的分析，并且从五个方面给出了注意力机制的不同层面阐述和洞见. 接着，对目前注意力机制在神经网络中的性能表现和可解释性进行讨论，并给出了三个典型应用场景下注意力机制的应用：计算机视觉中，神经机器翻译中以及语音识别中的注意力机制应用，在最后，本文提出了注意力实现机制当前存在的问题和面临的挑战，并尝试对注意力实现机制未来的研究方向进行了展望，对以后注意力机制在不同方向和领域的研究进行了一定的指导.

参 考 文 献

[1] Carrasco M. Visual attention: The past 25 years. *Vision Res*, 2011, 51(13): 1484

[2] Walther D, Rutishauser U, Koch C, et al. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Comput Vis Image Underst*, 2005, 100(1-2): 41

[3] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate // *Proceedings of the* 3*rd International Conference on Learning Representations*. San Diego, 2015: 1

[4] Cho K, van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J/OL]. *arXiv preprint* (2014-6-3) [2020-12-16]. https://arxiv.org/abs/1406.1078

[5] Xu K, Ba J L, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention // *Proceedings of the* 32*nd International Conference on Machine Learning*. Lille, 2015: 2048

[6] Luong T, Pham H, Manning C D. Effective approaches to Attention-based Neural Machine Translation // *Proceedings of the* 2015 *Conference on Empirical Methods in Natural Language Processing*. Lisbon, 2015: 1412

[7] Yang Z C, Yang D Y, Dyer C, et al. Hierarchical Attention Networks for Document Classification // *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, 2016: 1480

[8] Zhang J M, Bargal S A, Lin Z, et al. Top-down neural attention by excitation backprop. *Int J Comput Vis*, 2018, 126(10): 1084

[9] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning // *Proceedings of the* 34*th International Conference on Machine Learning*. Sydney, 2017: 1243

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need // *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*. Long Beach, 2017: 6000

[11] Shen T, Zhou T Y, Long G D, et al. DiSAN: directional self-attention network for RNN/CNN-free language understanding // *Proceedings of the AAAI Conference on Artificial Intelligence*. Louisiana, 2018: 32

[12] Lin Z H, Feng M W, Santos C N, et al. A structured self-attentive sentence embedding [J/OL]. *arXiv preprint* (2017-3-9) [2020-12-16].https://arxiv.org/abs/1703.03130

[13] Shen T, Zhou T Y, Long G D, et al. Bi-directional block self-attention for fast and memory-efficient sequence modeling // *Proceedings of the* 6*th International Conference on Learning Representations*. Vancouver, 2018: 1

[14] Shen T, Zhou T Y, Long G D, et al. Reinforced Self-Attention Network: a Hybrid of Hard and Soft Attention for Sequence Modeling // *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. Stockholm, 2018: 4345

[15] Kim Y, Denton C, Hoang L, et al. Structured attention networks [J/OL]. *arXiv preprint* (2017-2-16) [2020-12-16].https://arxiv.org/abs/1702.00887

[16] Chaudhari S, Mithal V, Polatkan G, et al. An attentive survey of attention models [J/OL]. *arXiv preprint* (2019-4-5) [2020-12-16]. https://arxiv.org/abs/1904.02874

[17] Mnih V, Heess N, Graves A. Recurrent models of visual attention // *Proceedings of the* 27*th International Conference on Neural Information Processing Systems*. Cambridge, 2014: 2204

[18] Chan W, Jaitly N, Le Q, et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition // 2016 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). Shanghai, 2016: 4960

[19] Kiela D, Wang C H, Cho K. Dynamic Meta-Embeddings for Improved Sentence Representations // *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*. Brussels, 2018: 1466

[20] Maharjan S, Montes M, González F A, et al. A genre-aware attention model to improve the likability prediction of books // *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*. Brussels, 2018: 3381

[21] Lu J S, Yang J W, Batra D, et al. Hierarchical question-image co-attention for visual question answering. *Adv Neural Infor Processing Syst*, 2016, 29: 289

[22] Wang W, Pan S J, Dahlmeier D, et al. Coupled multi-layer attentions for co-extraction of aspect and opinion terms // *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. San Francisco, 2017: 3316

[23] Ying H C, Zhuang F Z, Zhang F Z, et al. Sequential recommender system based on hierarchical attention networks // *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. Stockholm, 2018: 3926

[24] De-Arteaga M, Romanov A, Wallach H, et al. Bias in bios: a case study of semantic representation bias in a high-stakes setting // *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Atlanta, 2019: 120

[25] Lee J, Shin J H, Kim J S. Interactive visualization and manipulation of attention-based neural machine translation // *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Copenhagen, 2017: 121

[26] Liu S S, Li T, Li Z M, et al. Visual Interrogation of Attention-Based Models for Natural Language Inference and Machine Comprehension // *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, 2018: 36

[27] Jain S, Wallace B C. Attention is not explanation [J/OL]. *arXiv preprint*(2019-2-26)[2020-12-16].https://arxiv.org/abs/1902.10186

[28] Wiegreffe S, Pinter Y. Attention is not not explanation // *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*). Hong Kong, 2019: 11

[29] Jetley S, Lord N A, Lee N, et al. Learn to pay attention [J/OL]. *arXiv preprint* (2018-4-6) [2020-12-16].https://arxiv.org/abs/1804.02391

[30] Sharma S, Kiros R, Salakhutdinov R. Action recognition using visual attention [J/OL]. *arXiv preprint* (2015-12-12) [2020-12-16]. https://arxiv.org/abs/1511.04119

[31] Kataoka Y, Matsubara T, Uehara K. Image generation using generative adversarial networks and attention mechanism // 2016 *IEEE/ACIS* 15*th International Conference on Computer and Information Science* (*ICIS*). Okayama, 2016: 1

[32] Gregor K, Danihelka I, Graves A, et al. Draw: A recurrent neural network for image generation // *Proceedings of the* 32*nd International Conference on Machine Learning*. Lille, 2015: 1462

[33] Parmar N, Vaswani A, Uszkoreit J, et al. Image transformer // *Proceedings of the* 35*th International Conference on Machine Learning*. Stockholm, 2018: 4052

[34] Huang P Y, Liu F, Shiang S R, et al. Attention-based Multimodal Neural Machine Translation // *Proceedings of the First Conference on Machine Translation*. Berlin, 2016: 639

[35] Zhang H, Goodfellow I, Metaxas D, et al. Self-attention generative adversarial networks // *Proceedings of the* 36*nd International Conference on Machine Learning.* Long Beach, 2019: 7354

[36] Cohn T, Hoang C D V, Vymolova E, et al. Incorporating structural alignment biases into an attentional neural translation model // *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, 2016: 876

[37] Feng S, Liu S, Yang N, et al. Improving attention modeling with implicit distortion and fertility for machine translation // *Proceedings of the* 26*th International Conference on Computational Linguistic*. Osaka, 2016: 3082

[38] Eriguchi A, Hashimoto K, Tsuruoka Y. Tree-to-sequence attentional neural machine translation // *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics*. Berlin, 2016: 823

[39] Sankaran B, Mi H, Al-Onaizan Y, et al. Temporal attention model for neural machine translation [J/OL]. *arXiv preprint* (2016-8-9) [2020-12-16].https://arxiv.org/abs/1608.02927

[40] Cheng Y, Shen S Q, He Z J, et al. Agreement-based joint training for bidirectional attention-based neural machine translation // *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. New York, 2016: 2761

[41] Liu L, Utiyama M, Finch A, et al. Neural machine translation with supervised attention [J/OL]. *arXiv preprint* (2016-9-14) [2020-12-16].https://arxiv.org/abs/1609.04186

[42] Britz D, Goldie A, Luong M T, et al. Massive exploration of neural machine translation architectures // *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*. Copenhagen, , 2017: 1442

[43] Tang G B, Müller M, Rios A, et al. Why Self-attention? A targeted evaluation of neural machine translation architectures // *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*. Brussels, 2018: 4263

[44] Yin W P, Schütze H, Xiang B, et al. ABCNN: attention-based convolutional neural network for modeling sentence pairs. *Trans Assoc Comput Linguist*, 2016, 4: 259

[45] Zhuang P Q, Wang Y L, Qiao Y. Learning attentive pairwise interaction for fine-grained classification. *Proc AAAI Conf Artif Intell*, 2020, 34(7): 13130

[46] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification // *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics*. Berlin, 2016: 207

[47] Wang Y Q, Huang M L, zhu X Y, et al. Attention-based LSTM for aspect-level sentiment classification // *Proceedings of the* 2016 *Conference on Empirical Methods in Natural Language Processing*. Austin, 2016: 606

[48] Ma Y, Peng H, Cambria E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM // *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. New Orleans, 2018: 5876

[49] Zhang S C, Loweimi E, Bell P, et al. On the usefulness of self-attention for automatic speech recognition with transformers // *Proceedings of* 2021 *IEEE Spoken Language Technology Workshop* (*SLT*). Shenzhen, 2021: 89

[50] Sarı L, Moritz N, Hori T, et al. Unsupervised speaker adaptation using attention-based speaker memory for end-to-end ASR // *Proceedings of the* 2020 *IEEE International Conference on Acoustics, Speech and Signal Processing*. Barcelona, 2020: 7384

[51] Chorowski J, Bahdanau D, Serdyuk D, et al. An online attention-based model for speech recognition [J/OL]. *arXiv preprint* (2015-06-24) [2020-12-16].https://arxiv.org/abs/1506.07503

[52] Bahdanau D, Chorowski J, Serdyuk D, et al. End-to-end attention-based large vocabulary speech recognition // *Proceedings of the* 2016 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). Shanghai, 2016: 4945

[53] Shen S, Lee H. Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection [J/OL]. *arXiv preprint* (2016-3-31) [2020-12-16].https://arxiv.org/abs/1604.00077

[54] Liu B, Lane I. Attention-based recurrent neural network models for joint intent detection and slot filling [J/OL]. *arXiv preprint* (2016-9-6) [2020-12-16].https://arxiv.org/abs/1609.01454

[55] Shen Y, Tan S, Sordoni A, et al. Ordered neurons: Integrating tree structures into recurrent neural networks [J/OL]. *arXiv preprint* (2018-10-22) [2020-12-16].https://arxiv.org/abs/1810.09536

[56] Nguyen X P, Joty S, Hoi S C H, et al. Tree-structured attention with hierarchical accumulation [J/OL]. *arXiv preprint* (2020-2-19) [2020-12-16].https://arxiv.org/abs/2002.08046

[57] Tsai Y H H, Srivastava N, Goh H, et al. Capsules with inverted dot-product attention routing [J/OL]. *arXiv preprint* (2020-2-19) [2020-12-16].https://arxiv.org/abs/2002.04764