第八章　记忆与注意力机制

> 当一个人在吵闹的鸡尾酒会上和朋友聊天时，尽管周围噪音干扰很多，他还是可以听到朋友的谈话内容，而忽略其他人的声音。同时，如果未注意到的背景声中有重要的词（比如他的名字），他会马上注意到。
>
> — 鸡尾酒会效应

在人工神经网络模型中，深度学习可以简单地理解为表示学习和浅层分类器两部分构成。表示学习可以通过前馈神经网络、卷积神经网络或者循环神经网络来将输入信息转换到向量表示，可以看作是编码过程。一般来讲，这个过程是有信息损失的。在简单的一些分类任务中，只需要提取对分类有帮助的、有限的关键信息，因此可以容忍编码向量丢失部分信息。但是在一些复杂的任务中，需要尽可能保留多的输入信息，任何丢失的信息都可能会降低任务的准确率。以自然语言处理中的文本建模为例，用循环神经网络来将一段文本转换为向量表示（即最终的隐藏状态）。当要处理的文本很长时，循环神经网络很难将这段文本的所有语义信息都编码到一个向量中。在文本分类这类比较简单的任务中，不需要编码所有信息，因此这样直接对文本进行建模的方法通常比较有效。但是在机器翻译这类比较复杂的任务中，用一个向量来存储文本的语义信息就会丢失很多关键的信息。

在循环神经网络中，丢失信息的另外一个因素是**远距离依赖问题**。

一般来讲，向量存储信息的容量和向量维度以及编码网络的复杂度成正比。如果要存储的信息越多，向量维度就要越大或者编码网络要越复杂。这都会导致编码网络的参数成倍地增加。这个问题称为网络容量（network capacity）问题。

在增加少量参数的前提下，可以通过**注意力机制**和**外部记忆**来提高网络存储信息的容量。

首先，借鉴计算机的体系结构，通过引入外部记忆（External Memory）来增强神经网络。将难以编码的信息存储于外部记忆中，需要时在读取到网络中并参与计算。以LSTM模型为例，其内部记忆单元可以类比于计算机的寄存器，外部记忆可以类比于计算机的内存。

其次，对于外部记忆中存储的信息，可以通过注意力机制（attention mechanism）来进行寻址，并读取对应的信息。

8.1 记忆

在生物神经网络中，记忆是外界信息在人脑中存储机制。生理学家发现信息是作为某种整体效应（collective effect）存储在大脑组织中。当大脑皮层的不同部位损伤时，其导致的不同行为表现似乎取决于损伤的程度而不是损伤的确切位置 [Kohonen, 2012]。大脑组织的每个部分似乎都携带一些导致相似行为的信息。也就是说，记忆在大脑皮层是分布式存储的，而不是存储于某个局部区域 [Thompson, 1975]。

大脑记忆毫无疑问是通过生物神经网络实现的，其机理目前还无法解释。直观上，记忆机制和神经网络的连接以及神经元的活动相关。大脑记忆的一个主要特点是联想记忆。联想记忆（associative memory）是指一种学习和记住不同对象之间关系的能力，比如记住一个人的名字，或食物的味道等。联想记忆是人脑的重要能力，可以归结为人脑中信息的存储、关联及检索的神经活动机制，因此对人工智能的研究都有着极其重要的指导意义。

现代计算机的存储方式是根据地址来进行存储的，称为随机访问存储（random access memory，RAM）。和RAM不同，联想记忆是指一种可以通过内容匹配的方法进行寻址的信息存储方式，也称为基于内容寻址存储（content-addressable memory，CAM）。

联想记忆是一个人工智能、计算机科学和认知科学等多个交叉领域的热点研究问题，不同学科中的内涵也不太相同。

在人工智能领域，联想记忆是指使用神经网络来实现的基于内容寻址的信息存储模型。记忆一般分为长期记忆和短期记忆。长期记忆（long-term memory），也称为知识（knowledge），体现为神经元之间的连接关系和连接权重，其更新速度比较慢；短期记忆（short-term memory）体现为神经元的活动，更新较

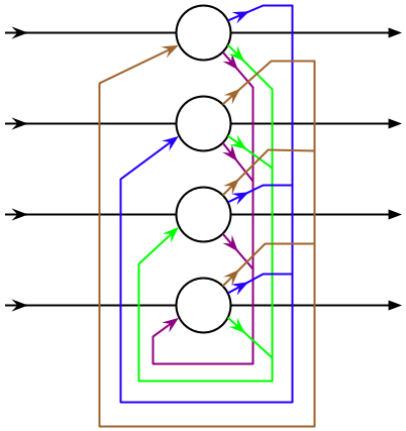


图 8.1: Hopfield 网络。图片来源于 https://en.wikipedia.org/wiki/Hopfield_-network。

快。短期记忆可以不断强化从而形成长期记忆。短期记忆、长期记忆的动态更新过程称为演化（evolution）过程。

联想记忆大体上可以分为两个阶段：学习和检索。将一个模型存储在记忆中的过程是一种学习过程。当学习完成后，联想记忆可以根据一个不完整或带噪声的输入，通过网络迭代来回忆出记忆中存储的正确模式。联想记忆模型主要是通过神经网络的动态演化来进行联想，有两种应用场景：1）输入的模式和输出的模式在同一空间，这种模型叫做自联想记忆模型（auto-associative model）。自联想模型可以通过前馈神经网络或者循环神经网络来实现，也经常称为自编码器（auto-encoder）。2）输入的模式和输出的模式不在同一空间，这种模型叫做异联想记忆模型（hetero-associative model）。从广义上讲，大部分模式识别问题都可以看作是异联想，因此异联想记忆模型可以作为分类器使用。

8.1.1 Hopfield 网络

Hopfield 网络（Hopfield network）是一种循环神经网络模型。

8.1.2 外部记忆

在目前循环神经网络中，比如LSTM和GRU，网络容量问题也是限制其能力的主要因素。为了克服这个限制，一些研究者引入了一种基于内容寻址的外部记忆来提高网络容量，比如神经图灵机[Graves,2014]、记忆网络[Sukhbaatar,2015]等。这些外部记忆被保存在数组、栈或队列等结构中。给定一个线索向量，主网络使用注意力机制来从外部记忆中选择相关记忆并参与主网络的计算。外部记忆的引入更多是受现代计算机架构的启发，将计算和存储功能进行分离，这些外部记忆的结构也缺乏生物学的解释性。为了具有更好的生物学解释性，Danihelka[2016]将一个联想记忆模型作为部件引入LSTM网络中，而从在不引入额外参数的情况下增加网络容量。Ba[2016]将循环神经网络中的部分连接权重作为短期记忆，并通过一个联想记忆模型进行更新，而从提高网络性能。Trabelsi[2017]进一步将模块化的深度网络（比如残差网络）也作为联想记忆，并引入更加有效的复数表示来提高网络性能。在上述的网络中，联想记忆都是作为一个更大网络的部件，用来增加短期记忆的容量。联想记忆部件的参数作为整个网络参数的一部分来进行学习。

8.2 注意力机制

注意力（Attention）是在计算能力有限情况下的一种资源分配方案，将计算资源分配给更重要的任务。

心理学 注意力是一种人类不可或缺的复杂认知功能，指人可以在关注一些信息的同时忽略另一些信息的选择能力。在日常生活中，我们通过视觉、听觉、触觉等方式接收大量的感觉输入。但是我们的人脑可以在这些外界的信息轰炸中还能有条不紊地工作，是因为人脑可以有意或无意地从这些大量输入信息中选择小部分的有用信息来重点处理，并忽略其他信息。这种能力就叫做**注意力**。注意力可以体现在外部的刺激（听觉、视觉、味觉等），也可以体现在内部的意识（思考、回忆等）。不管这些注意力是有意还是无意，大部分的人脑活动都需要依赖注意力，比如记忆信息，阅读或思考等。有意识的注意力是指有预定目的、需要主动有意识地聚焦于某一对象的注意力，而无意识的注意力是由外界刺激驱动的注意，主观感觉不到的注意力。

注意力一般会随着环境或情景的不同而选择不同的信息。比如当要从人群

中寻找某个人时，我们会将专注于每个人的脸部；而当要统计人群的人数时，我们只需要专注于每个人的轮廓。

当用神经网络来处理大量的输入信息时，也可以借鉴人脑的注意力机制，只选择一些关键的信息输入进行处理，来提高神经网络的效率。以阅读理解任务为例，给定一篇很长的文本段落，然后就此文本段落的内容进行提问。提出的问题只和段落中的一两个句子相关，其余部分都是无关的。我们仅仅需要把相关的片段挑选出来就足够了。

阅读理解任务参见第??节，第??页。

我们用$m_{1:N} = [m_1, \cdots, m_N]$表示$N$个输入信息。在特定的上下文$q$时，为了节省计算资源，不需要将所有的$N$个输入信息都输入到神经网络进行计算，只需要从$m_1, \cdots, m_N$中选择一个相关的信息输入给神经网络。我们用来表示注意力变量$z \in [1, N]$来表示被选择信息的索引，即$z = i$表示选择了第i个输入信息。为了方便计算，我们采用一种“软”的信息选择机制，首先计算在给定q和$m_{1:N}$下，选择第i个信息的概率α_i，

α_i称为注意力分布（Attention Distribution）。

$$\begin{aligned}\alpha_i &= p(z = i|m_{1:N}, q) \\ &= \text{softmax}\Big(s(m_i, q)\Big) \\ &= \frac{\exp\Big(s(m_i, q)\Big)}{\sum_{j=1}^{N}\exp\Big(s(m_j, q)\Big)},\end{aligned} \tag{8.1}$$

其中，$s(m_i, q)$为注意力打分函数，可以是一个加性模型

$$s(m_i, q) = \mathbf{v}^T \tanh(W\mathbf{m}_i + U\mathbf{q}), \tag{8.2}$$

也可以是点积模型

$$s(m_i, q) = \mathbf{m}_i^\top \mathbf{q}, \tag{8.3}$$

$$s(m_i, q) = \mathbf{m}_i^\top W\mathbf{q}, \tag{8.4}$$

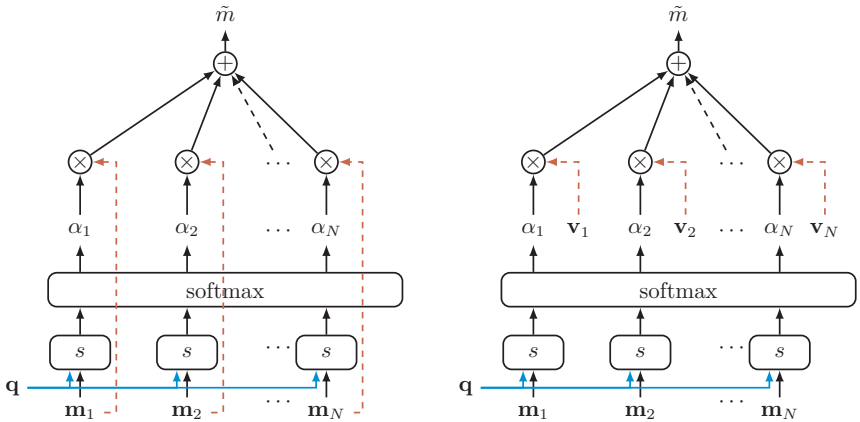
其中，$\mathbf{m}_i$和$\mathbf{q}$为m_i和q的向量表示，其余的$W, U, \mathbf{v}$为网络参数。

注意力分布α_i可以解释为在上下文查询q时，第i个信息受关注的程度。这样，我们采用一种“软”的信息选择机制，基于上下文对输入信息进行编码为

$$\textbf{attention}(m_{1:N}, q) = \sum_{i=1}^{N}\alpha_i\mathbf{m}_i, \tag{8.5}$$

公式(8.5)称为注意力机制（Attention Mechanism）。

注意力机制也可称为**注意力模型**。



(a) 普通模式 (b) 键值对模式

图 8.2: 注意力机制

8.2.1 注意力机制的变体

除了上面介绍的基本模式外，注意力机制还存在一些变化的模型。

键值对注意力 更一般地，我们可以用键值对（Key-Value Pair）格式来表示输入信息，其中“键”用来计算注意力分布 α_i，“值”用来生成选择的信息。我们用 $(k, v)_{1:N} = [(k_1, v_1), \cdots, (k_N, v_N)]$ 表示 N 个输入信息。在特定的上下文 q 时，注意力函数为

$$\mathbf{attention}((k, v)_{1:N}, q) = \sum_{i=1}^{N} \alpha_i \mathbf{v}_i, \tag{8.6}$$

$$= \sum_{i=1}^{N} \frac{\exp\Big(s(\mathbf{k}_i, \mathbf{q})\Big)}{\sum_j \exp\Big(s(\mathbf{k}_j, \mathbf{q})\Big)} \mathbf{v}_i \tag{8.7}$$

其中，$\mathbf{k}_i, \mathbf{v}_i$ 为输入键值对 k_i, v_i 的向量表示，$s(\mathbf{k}_i, \mathbf{q})$ 可以为加性或点积模型。

图8.2给出两种注意力机制的示例。如果在键值对模式中 $\forall i, \mathbf{k}_i = \mathbf{v}_i$，则就等价于普通模式。

多头注意力 多头注意力（Multi-Head Attention）是利用多个查询$Q = \{q_1, \cdots, q_M\}$，来平行地计算从输入信息中选取多个信息。每个注意力关注输入信息的不同部分。

硬注意力 之前提到的注意力是软注意力，即所有输入信息上的概率分布。还有一种注意力是只关注到一个位置上，叫做硬注意力（Hard Attention）。硬注意力可以通过在注意力分布式上随机采样的方式实现。在训练时，我们一般使用软注意力，因为软注意力可以使用反向传播算法。

8.2.2 应用

神经机器翻译 注意力机制最成功的应用是机器翻译 [Bahdanau et al., 2014]。基于神经网络的机器翻译模型也叫做神经机器翻译（Neural Machine Translation，NMT）。一般的神经机器翻译模型采用“编码-解码”的方式进行序列到序列的转换。这种方式有两个问题：一是编码向量的容量瓶颈问题，即源语言所有的信息都需要保存在编码向量中，才能进行有效地解码；二是长距离依赖问题，即编码和解码过程中在长距离信息传递中的信息丢失问题。

神经机器翻译参见第??节，第??页。

通过引入注意力机制，我们将源语言中每个位置的信息都保存下来。在解码过程中生成每一个目标语言的单词时，我们都通过注意力机制直接从源语言的信息中选择相关的信息作为辅助。这样的方式就可以有效地解决上面的两个问题。一是无需让所有的源语言信息都通过编码向量进行传递，在解码的每一步都可以直接访问源语言的所有位置上的信息；二是源语言的信息可以直接传递到解码过程中的每一步，缩短了信息传递的距离。

图像描述生成 图像描述生成是输入一幅图像，输出这幅图像对应的描述。图像描述生成也是采用“编码-解码”的方式进行。编码器为一个卷积网络，提取图像的高层特征，表示为一个编码向量；解码器为一个循环神经网络语言模型，初始输入为编码向量，生成图像的描述文本。在图像描述生成的任务中，同样存在编码容量瓶颈以及长距离依赖这两个问题，因此也可以利用注意力机制来有效地选择信息。在生成描述的每一个单词时，循环神经网络的输入除了前一个词的信息，还有利用注意力机制来选择一些来自于图像的相关信息 [Xu et al., 2015]。

图 8.3: 基于注意力的图像描述生成。图像来源：[Xu et al., 2015]

机器阅读理解

8.3 典型的记忆网络

通过注意力机制，我们可以从大量的输入信息（或历史信息）中选择出对当前决策有帮助的信息。注意力机制可以分为两个步骤：计算注意力分布和信息加权平均。如果类比于计算机的存储器读取，将输入信息看做是存储于计算机存储器中的数据，则计算注意力分布的过程相当于是计算机的“寻址”过程，信息加权平均的过程相当于计算机的“内容读取”过程。

因此，注意力机制可以看做是一个*接口*，将信息的存储与计算分离。这样，神经网络可以分成两个部件：*控制器*（Controller）和*外部记忆*（External Memory）。控制器是一个神经网络，负责与外界的交互，接受外界的输入信息并产生输出到外界。控制器还同时负责对外部记忆的读写操作。大部分信息存储于外部记忆中，不需要全时参与计算。这类神经网络称为*记忆增强的神经网络*（Memory-Enhanced Neural Networks）。

和之前介绍的 LSTM 中的*记忆单元*相比，外部记忆可以存储更多的信息，并且不直接参与计算，通过读写接口来进行操作。而 LSTM 模型中的记忆单元包含了信息存储和计算两种功能，不能存储太多的信息。因此，LSTM 中的记忆单元可以类比于计算机中寄存器，而外部记忆可以类比于计算机中的存储器：随机访问内存（Random Access Memory，RAM）、磁带或硬盘等。

外部记忆从记忆结构、读写方式等方面可以演变出很多模型。在本节中，我们介绍两种比较有代表性的模型。

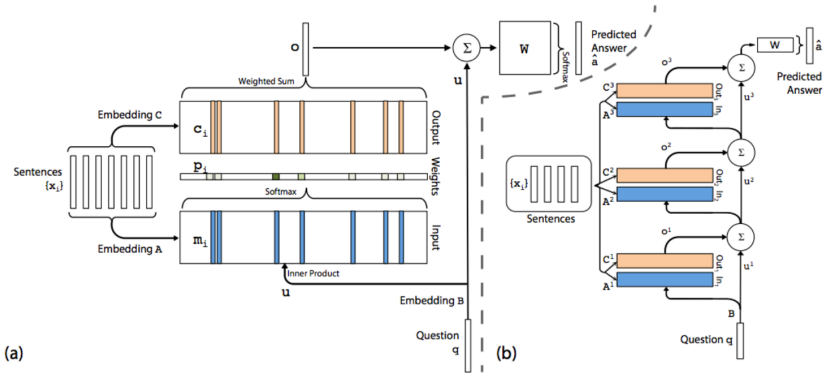


图 8.4: 端到端记忆网络。图像来源：[Sukhbaatar et al., 2015]

8.3.1 指针网络

Vinyals et al. [2015]

8.3.2 端到端记忆网络

我们用 $(k, v)_{1:N} = [(k_1, v_1), \cdots, (k_N, v_N)]$ 表示 N 个输入信息。在特定的上下文 q 时，注意力函数为

$$\mathbf{attention}((k, v)_{1:N}, q) = \sum_{i=1}^{N} \alpha_i \mathbf{v}_i, \tag{8.8}$$

$$= \sum_{i=1}^{N} \frac{\exp\Big(s(\mathbf{k}_i, \mathbf{q})\Big)}{\sum_j \exp\Big(s(\mathbf{k}_j, \mathbf{q})\Big)} \mathbf{v}_i \tag{8.9}$$

端到端记忆网络（End-To-End Memory Network, MemN2N）[Sukhbaatar et al., 2015] 采用一种循环网络的结构，可以多次从外部记忆中读取信息，是一种只读的外部记忆

8.3.3 可读写的外部记忆

神经图灵机（Neural Turing Machine，NTM）[Graves et al., 2014] 主要由两个部件构成：控制器和外部记忆。

控制器为一个神经网络，接受输入并产生输出，同时还可以对外部记忆进行读和写操作。

外部记忆定义为矩阵 $\mathbf{m} \in \mathbb{R}^{K \times M}$，这里 K 是记忆片段的数量，M 是每个记忆片段的大小。K 和 M 为超参数。

读操作 在时刻 t, 外部记忆的内容记为 $\mathbf{m}_t$，读操作为从外部记忆 $\mathbf{m}_t$ 中读取信息 $\mathbf{r}_t \in \mathbb{R}^M$。

$$\mathbf{r}_t = \alpha_t^\top \mathbf{m}_t \tag{8.10}$$

$$\sum_{i=1}^{K} \alpha_t(i)\mathbf{m}_t(i), \tag{8.11}$$

这里，$\mathbf{r}_t$ 定义为读向量（Read Vector），$\alpha_t \in \mathbb{R}^K$ 是权重向量，每一个 $\alpha_t(i)$ 为记忆片段 $\mathbf{m}_t(i)$ 对应的权重，并满足

$$\sum_{i=1}^{K} \alpha_t(i) = 1. \tag{8.12}$$

写操作 外部记忆的写操作可以分解为两个子操作：删除和增加。

我们定义删除向量（erase）$\mathbf{e}_t$ 和增加向量（Add Vector）$\mathbf{a}_t$ 分别为要从外部记忆中删除的信息和要增加的信息。

删除操作可以写为

$$\tilde{\mathbf{m}_t}(i) = \mathbf{m}_{t-1}(i)(\mathbf{1} - \alpha_t(i)\mathbf{e}_t), \tag{8.13}$$

增加操作可以写为

$$\tilde{\mathbf{m}_t}(i) = \tilde{\mathbf{m}_t}(i) + \alpha_t(i)\mathbf{a}_t, \tag{8.14}$$

$$\mathbf{m}_t = \mathbf{f}_{write}(\mathbf{m}_{t-1}, \alpha_t, \mathbf{h}_t). \tag{8.15}$$

$$\alpha_{t,k} = \text{softmax}(g(\mathbf{m}_{t-1,k}, \mathbf{k}_{t-1})) \tag{8.16}$$

https://en.wikipedia.org/wiki/Working_memory

8.3.4 记忆结构

Chandar et al. [2016]

8.4 总结和深入阅读

注意力机制最早在计算机视觉中提出。在神经网络中，Mnih et al. [2014] 在循环神经网络模型上使用了注意力机制来进行图像分类。随后，Bahdanau et al. [2014] 使用注意力机制在机器翻译任务上将翻译和对齐同时进行。Xu et al. [2015] 利用注意力机制来进行看图说话。

比如有代表性的工作为神经图灵机 [Graves et al., 2014]、端到端记忆网络 [Sukhbaatar et al., 2015]、动态记忆网络 [Kumar et al., 2015] 等，这类引入外部记忆的模型都称为记忆增强网络（Memory-Enhanced Networks）。在自然语言处理中的很多任务上，都取得了一定的性能提升。

注意力模型已经是一个非常普遍和强大的技术，并且正变得越来越普遍。在目前的研究中，为了可以端到端进行训练，我们希望注意力是可微的。在实际使用中的注意力模型可以看作是一个注意力分布，即关注所有位置，只是程度不一样。通常，使用基于内容的注意力生成注意力分布。根据一个查询，每一个条目和这个查询计算出一个分数，来描述这个条目与查询匹配程度。这些分数被输入一个 softmax 来生成注意力分布。

但是目前的模型主要存在两个问题：外部记忆的结构，目前的模型中结构还比较简单，需要借鉴神经科学的研究成果。外部记忆和内部记忆的融合方式还需要进一步研究。

习题 8-1 分析 LSTM 模型中，隐藏层神经元数量与参数数量之间的关系。

参考文献

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *ArXiv e-prints*, September 2014.

Sarath Chandar, Sungjin Ahn, Hugo Larochelle, Pascal Vincent, Gerald Tesauro, and Yoshua Bengio. Hierarchical memory networks. *arXiv preprint*

arXiv:1605.07427, 2016.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Teuvo Kohonen. *Self-organization and associative memory*, volume 8. Springer Science & Business Media, 2012.

Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. *arXiv preprint arXiv:1506.07285*, 2015.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, pages 2204–2212, 2014.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, pages 2431–2439, 2015.

Richard F Thompson. *Introduction to physiological psychology*. HarperCollins Publishers, 1975.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700, 2015.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*, pages 2048–2057, 2015.