第八章　注意力与记忆机制

一切都应该尽可能地简单，但不能过于简单。

— 艾伯特●爱因斯坦

根据通用近似定理，前馈网络和循环网络都有很强的能力。但是由于优化算法和计算能力的限制，在实践中很难达到通用近似的能力。特别是在处理复杂任务时，比如需要处理大量的输入信息或者复杂的计算流程时，目前计算机的计算能力依然是限制神经网络发展的瓶颈。

为了减少计算复杂度，通过部分借鉴生物神经网络的一些机制，我们引入了*局部连接*、*权重共享*以及*汇聚*操作来简化神经网络结构。虽然这些机制比较有效的缓解了模型复杂度和表达能力之间的矛盾，但是我们依然还需要进一步在不过多增加模型复杂度（主要是模型参数）的情况下来提高模型的表达能力。以*阅读理解*任务为例，背景文章（background document）一般比较长，如果用循环神经网络来将其转换为向量表示，那么这个编码向量很难反映出背景文章的所有语义。在比较简单的*文本分类*任务中，只需要编码一些对分类有用的信息，因此用一个向量来表示文本语义来行得通。但是在阅读理解任务中，编码时还不知道可能会接收到什么样的问句。这些问句可能会涉及到背景文章的所有信息点，因此就会丢失任何信息都可能导致无法正确回答问题。

阅读理解任务是给机器阅读一篇背景文章，然后询问一些相关的问题，来测试机器是否理解了这篇文章。

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**。

神经网络中可以存储的信息量称为*网络容量*（network capacity）。一般来讲，利用一组神经元来存储信息的容量和神经元的数量以及网络的复杂度成正比。如果要存储越多的信息，神经元数量就要越多或者网络要越复杂，进而导致神经网络的参数成倍地增加。

我们人脑的生物神经网络同样存在*网络容量*问题，人脑中的工作记忆大概

只有几秒钟的时间，类似于循环神经网络中的隐状态。其次，人脑每个时刻接收的外界输入信息也非常多，包括来源于视觉、听觉、触觉的各种各样的信息。但就视觉来说，眼睛每秒钟都会发送千万比特的信息给视觉神经系统。人脑在有限的资源下，并不能同时处理这些过载的输入信息。大脑神经系统有两个重要机制可以解决信息过载问题：*注意力*和*记忆机制*。

同样，我们还可以借鉴人脑解决信息过载的机制，从两方面来提高神经网络处理信息的能力。一方面是信息选择，通过自上而下的信息选择机制（即*注意力*）来过滤掉大量的无关信息；另一方面是优化神经网络的记忆结构，通过引入额外的*外部记忆*来提高神经网络存储信息的容量。

8.1 注意力

在计算能力有限情况下，*注意力机制*（attention mechanism）是解决信息超载问题的主要手段的一种资源分配方案，将计算资源分配给更重要的任务。

认知神经学 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）。

注意力一般分为两种：一种是自上而下的有意识的注意力，称为聚焦式（focus）注意力。聚焦式注意力是指有预定目的、依赖任务的、主动有意识地聚焦于某一对象的注意力；另一种是自下而上的无意识的注意力，称为基于显著性（saliency-based）的注意力。基于显著性的注意力是由外界刺激驱动的注意，不需要主动干预，也和任务无关。如果一个对象的刺激信息不同于其周围信息，一种无意识的“赢者通吃”（winner-take-all）或者门控（gating）机制就可以把注意力转向这个对象。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。

一个和注意力有关的例子是*鸡尾酒会效应*。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（聚焦式注意力）。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（显著性注意力）。

聚焦式注意力一般会随着环境、情景或任务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。

8.1.1 注意力机制

除非特别声明，本节及以后章节中注意力机制是通常指自上而下的会聚式注意力。

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。在目前的神经网络模型中，我们可以将最大汇聚（max pooling）、门控（gating）机制来近似地看作是自下而上的基于显著性的注意力机制。除此之外，自上而下的会聚式注意力也是一种有效的信息选择方式。以阅读理解任务为例，给定一篇很长的文章，然后就此文章的内容进行提问。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。为了减小神经网络的计算负担，只需要把相关的片段挑选出来让后续的神经网络来处理，而不需要把所有文章内容都输入给神经网络。

阅读理解任务参见第??节，第??页。

用$\mathbf{x}_{1:N} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$表示$N$个输入信息，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$m_1, \cdots, m_N$中选择一些和任务相关的信息输入给神经网络。给定一个和任务、情景相关的查询向量$\mathbf{q}$，我们用注意力变量$z \in [1, N]$来表示被选择信息的索引位置，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$\mathbf{q}$和$\mathbf{x}_{1:N}$下，选择第i个输入信息的概率α_i，

α_i称为注意力分布（attention distribution）。

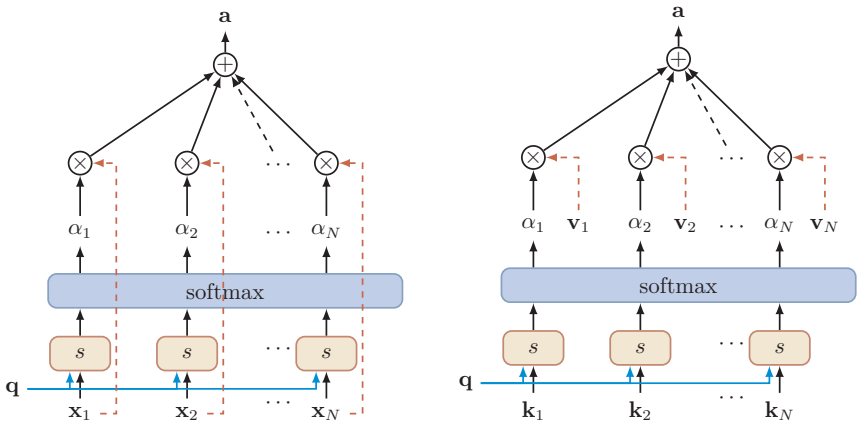
$$\begin{aligned}\alpha_i &= p(z = i|\mathbf{x}_{1:N}, \mathbf{q}) \\ &= \text{softmax}\Big(s(\mathbf{x}_i, \mathbf{q})\Big) \\ &= \frac{\exp\Big(s(\mathbf{x}_i, \mathbf{q})\Big)}{\sum_{j=1}^{N} \exp\Big(s(\mathbf{x}_j, \mathbf{q})\Big)},\end{aligned} \tag{8.1}$$

其中，$s(\mathbf{x}_i, \mathbf{q})$为注意力打分函数，可以是一个加性模型

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

也可以是点积模型

$$s(\mathbf{x}_i, \mathbf{q}) = \mathbf{x}_i^{\mathrm{T}}\mathbf{q}, \tag{8.3}$$



(a) 普通模式 (b) 键值对模式

图 8.1: 注意力机制

$$s(\mathbf{x}_i, \mathbf{q}) = \mathbf{x}_i^{\mathrm{T}} W \mathbf{q}, \tag{8.4}$$

其中 $W, U, \mathbf{v}$ 为可学习的网络参数。

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。我们采用一种“软”的信息选择机制对输入信息进行编码为

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

$$= \mathbb{E}_{z \sim p(z|\mathbf{x}_{1:N}, \mathbf{q})}[\mathbf{x}]. \tag{8.6}$$

注意力机制也可称为注意力模型。

公式 (8.5) 称为软性注意力机制（soft attention mechanism）。

8.1.2 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。

多头注意力 多头注意力（multi-head attention）是利用多个查询 $\mathbf{q}_{1:M} = \{\mathbf{q}_1, \cdots, \mathbf{q}_M\}$，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。

硬性注意力 之前提到的注意力是软性注意力，即基于注意力分布的所有输入信息的期望。还有一种注意力是只关注到一个位置上，叫做*硬性注意力*（hard attention）。

硬性注意力有两种实现方式，一种是选取最高概率的输入信息，即

$$\mathbf{attention}(\mathbf{x}_{1:N}, \mathbf{q}) = \mathbf{x}_j, , \tag{8.7}$$

其中j为概率最大的输入信息的下标，即$j = \arg\max_{i=1}^{N} \alpha_i$。

另一种硬性注意力可以通过在注意力分布式上随机采样的方式实现。

硬性注意力的一个缺点是基于最大采样或随机采样的方式来选择信息。因此最终的损失函数与注意力分布之间的函数关系不可导，因此无法使用在反向传播算法进行训练。为了使用反向传播算法，一般使用软性注意力来代替硬性注意力。

硬性注意力需要通过强化学习来进行训练。

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。用$(\mathbf{k}, \mathbf{v})_{1:N} = [(\mathbf{k}_1, \mathbf{v}_1), \cdots, (\mathbf{k}_N, \mathbf{v}_N)]$表示$N$个输入信息，给定任务相关的查询向量$\mathbf{q}$时，注意力函数为

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

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

其中$s(\mathbf{k}_i, \mathbf{q})$可以为加性或点积模型。

图8.1给出两种注意力机制的示例。如果在键值对模式中$\mathbf{k}_i = \mathbf{v}_i, \forall i$，则就等价于普通模式。

结构化注意力 要从输入信息中选取出和任务相关的信息，主动注意力是在所有输入信息上的多项分布，是一种*扁平*（flat）结构。如果输入信息本身具有*层次*（hierarchical）结构，比如文本可以分为词、句子、段落、篇章等不同粒度的层次，我们可以使用层次化的注意力来进行更好的信息选择[Yang et al., 2016]。此外，还可以假设注意力上下文相关的二项分布，用一种图模型来构建更复杂的结构化注意力分布[Kim et al., 2017]。

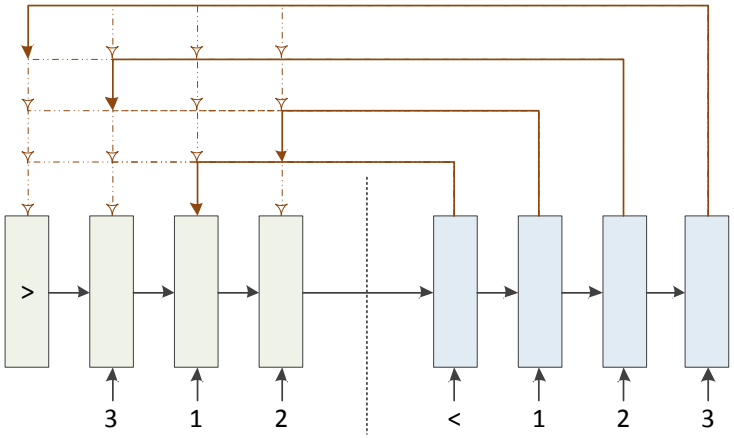


图 8.2: 指针网络

指针网络 上面描述的注意力机制主要是用来做信息筛选，从输入信息中选取相关的信息。事实上，注意力机制可以分为两步：一是计算注意力分布 α，二是根据 α 来计算输入信息的加权平均。我们可以只利用注意力机制中的第一步，将注意力分布作为一个软性的指针（pointer）来指出相关信息的位置。

指针网络（pointer network）[Vinyals et al., 2015] 是一种序列到序列模型，输入是长度为 n 的向量序列 $\mathbf{x}_{1:n} = \mathbf{x}_1, \cdots, \mathbf{x}_n$，输出是下标序列 $c_{1:m} = c_1, c_2, \cdots, c_m$，$c_i \in [1, n], \forall i$。

和一般的序列到序列任务不同，这里的输出序列是输入序列的下标（索引）。比如输入一组乱序的数字，输出为按大小排序的输入数字序列的下标。比如输入为 $3, 1, 2$，输出为 $1, 3, 2$。

条件概率 $p(c_{1:m}|\mathbf{x}_{1:n})$ 可以写为

$$p(c_{1:m}|\mathbf{x}_{1:n}) = \prod_{i=1}^{m} p(c_i|c_{1:i-1}, \mathbf{x}_{1:n}) \tag{8.10}$$

$$\approx \prod_{i=1}^{m} p(c_i|\mathbf{x}_{c_1}, \cdots, \mathbf{x}_{c_{i-1}}, \mathbf{x}_{1:n}), \tag{8.11}$$

其中条件概率 $p(c_i|\mathbf{x}_{c_1}, \cdots, \mathbf{x}_{c_{i-1}}, \mathbf{x}_{1:n})$ 可以通过注意力分布来计算。假设用一个循环神经网络对 $\mathbf{x}_{c_1}, \cdots, \mathbf{x}_{c_{i-1}}, \mathbf{x}_{1:n})$ 进行编码得到向量 $\mathbf{e}_i$，则

$$p(c_i|c_{1:i-1}, \mathbf{x}_{1:n}) = \text{softmax}(s_{i,j}), \tag{8.12}$$

其中 $s_{i,j}$ 为在解码过程的第 i 步时，每个输入向量的未归一化的注意力分布，

$$s_{i,j} = \mathbf{v}^{\mathrm{T}} \tanh(W\mathbf{x}_j + U\mathbf{e}_i), \forall j \in [1, n], \tag{8.13}$$

其中 $\mathbf{v}, W, U$ 为可学习的参数。

8.1.3 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]。

机器阅读理解 待写

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

8.2 记忆

8.2.1 人脑中的记忆

在生物神经网络中，记忆是外界信息在人脑中存储机制。大脑记忆毫无疑问是通过生物神经网络实现的。虽然其机理目前还无法解释，但直观上记忆机制和神经网络的联结形态以及神经元的活动相关。生理学家发现信息是作为一种整体效应（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]。

记忆类型 虽然人脑记忆的存储机制还不清楚，但是在我们已经大概可以确定不同脑区参与了记忆形成的几个阶段。人脑记忆的一个特点记忆一般分为长期记忆和短期记忆。长期记忆（long-term memory），也称为结构记忆或知识（knowledge），体现为神经元之间的联结关形态，其更新速度比较慢。短期记忆（short-term memory）体现为神经元的活动，更新较快，维持时间为几秒至几分钟。短期记忆是神经连接的暂时性强化，通过不断巩固、强化可形成长期记忆。短期记忆、长期记忆的动态更新过程称为演化（evolution）过程。

事实上，人脑记忆周期的划分并没有清晰的界限，也存在其它的划分方法。

此外，在执行某个认知行为（比如记下电话号码，算术运算）时，在人脑中还会存在一个“缓存”，即工作记忆（working memory）。工作记忆是一个记忆的临时存储和处理系统，维持时间通常为几秒钟。从时间上看，工作记忆也是一种短期记忆，但和短期记忆的内涵不同。短期记忆一般指外界的输入信息在人脑中的表示和短期存储，不关心这些记忆如何被使用；而工作记忆是一个和任务相关的“容器”，可以临时存放和某项任务相关的短期记忆和其它相关的内在记忆。工作记忆的容量一般都比较小，一般可以容纳 4 组项目。

作为不严格的类比，现代计算机的存储也可以按照不同的周期分为不同的存储单元，比如寄存器、内存、外存（比如硬盘等）。

联想记忆 大脑记忆的一个主要特点是通过联想来进行检索的。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，比如看见一个人然后想起他的名字，或记住某种食物的味道等。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的存储（content-addressable memory，CAM）。作为对比，现代计算机的存储方式根据地址来进行存储的，称为随机访问存储（random access memory，RAM）。

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。

借鉴人脑中工作记忆，可以在神经网络中引入一个辅助记忆模块来提高网络容量，将和任务相关的短期记忆保存在辅助记忆中，需要时再进行读取。辅助记忆的实现途径有两种：一种是结构化的外部记忆，另一种是联想记忆。

8.2.2 结构化的外部记忆

为了增强网络容量，一种比较简单的方式是引入结构化的记忆模块，将和任务相关的短期记忆保存在辅助记忆中，需要时再进行读取。

这个引入的结构化辅助记忆一般称为外部记忆（external memory），以区别与循环神经网络的内部记忆（即隐状态）。以LSTM模型为例，其内部记忆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。装备外部记忆的神经网络也称为记忆增强神经网络（memory augmented neural network）。

外部记忆有两个特点：

结构性 记忆通过结构化的的方法来存储，一般用一个向量来表示一个记忆片段（memory segment）。用一组向量来表示多个记忆。记忆的组织方式可以是数组、树、栈或队列等。

按内容寻址 要访问外部记忆中存储的信息，需要通过按内容的寻址方式进行定位，然后进行读取或写入操作。通常使用注意力机制来进行按内容寻址。

通过引入外部记忆，可以将神经网络的参数和记忆容量的“分离”，即在不增加网络参数的条件下增加网络容量。

比较典型的结构化外部记忆模型有记忆网络、神经图灵机等。

参见第8.3节，第11页。

8.2.3 联想记忆

还有一种增强网络容量的方式是将联想记忆模型引入到目前的神经网络中。本书中之前介绍的神经网络都是作为一种机器学习模型的输入-输出映射函数，其参数学习方法是通过梯度下降方法来最小化损失函数。除了作为机器学习模型外，神经网络还可以作为一种记忆的存储和检索模型，其学习方式为Hebbian法则等。

在本书中，联想记忆模型指使用神经网络来实现的基于内容寻址的信息存储和检索模型。

联想记忆的特点是基于内容寻址的存储和检索。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。

Hopfield网络

*Hopfield*网络（Hopfield network）是一种循环神经网络模型，由一层相互连接的神经元组成。每个神经元既是输入单元，又是输出单元，没有隐藏神经元。一个神经元和自身没有反馈相连，不同神经元之间连接权重是对称的。

Hopfield网络也可以认为是所有神经元都相互连接的不分层的神经网络。

假设一个Hopfield网络有m个神经元，第i个神经元的更新规则为

$$s_i = \begin{cases} +1 & \text{当} \sum_{j=1}^{m} w_{ij} s_j + b_i \geq 0 \\ -1 & \text{否则} \end{cases}, \tag{8.14}$$

这里我们只介绍离散Hopfield网络，神经元状态为$+1, -1$两种。除此之外，还有连续Hopfield网络，即神经元状态为连续值。

其中w_{ij}为神经元i和j之间的连接权重，b_i为偏置。连接权重w_{ij}有以下性质

$$\begin{aligned} w_{ii} &= 0 \quad \forall i \\ w_{ij} &= w_{ji} \quad \forall i, j. \end{aligned} \tag{8.15}$$

Hopfield网络的更新可以分为异步和同步两种方式。异步更新是每次更新一个神经元。神经元的更新顺序可以是随机或事先固定的。同步更新是指一次

更新所有的神经元，需要有一个时钟来进行同步。第t时刻的神经元状态为$\mathbf{s}_t = [\mathbf{s}_{t,1}, \mathbf{s}_{t,2}, \cdots, \mathbf{s}_{t,m}]^{\mathrm{T}}$，其更新规则为

$$\mathbf{s}_t = f(W\mathbf{s}_{t-1} - \mathbf{b}), \tag{8.16}$$

其中，$\mathbf{s}_0 = \mathbf{x}$，$W = [w_{ij}]_{m\times m}$为连接权重，$\mathbf{b} = [b_i]_m$为偏置，$f(\cdot)$为非线性阶跃函数。

能量函数 在Hopfield网络中，我们给每个不同的网络状态定义一个标量属性，称为“能量”。

$$E = -\frac{1}{2}\sum_{i,j} w_{ij}s_i s_j + \sum_i b_i s_i \tag{8.17}$$

$$= -\frac{1}{2}\mathbf{s}^{\mathrm{T}}W\mathbf{s} + \mathbf{b}^{\mathrm{T}}\mathbf{s}. \tag{8.18}$$

权重对称的要求是一个重要特征，因为它保证了能量方程（称向函数某一点收敛的过程为势能转化为能量）在神经元激活时单调递减，而不对称的权重可能导致周期性的递增或者噪声。然而，Hopfiled网络也证明噪声过程会被局限在很小的范围，并且并不影响网络的最终性能。

吸引子

学习规则 Hopfield网络的学习规则有很多种。一个最主要的学习规则是Hebbian法则。

$$w_{ij} = \frac{1}{N}\sum_{n=1}^{N} x_i^{(n)} x_j^{(n)} \tag{8.19}$$

其中$x_i^{(n)}$是第n个输入模式的第i维特征。

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

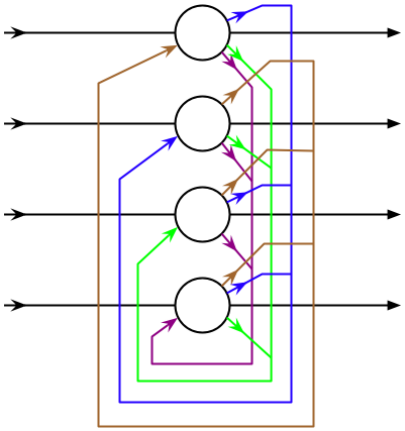


图 8.4: Hopfield 网络。图片来源于 https://en.wikipedia.org/wiki/Hopfield_network。

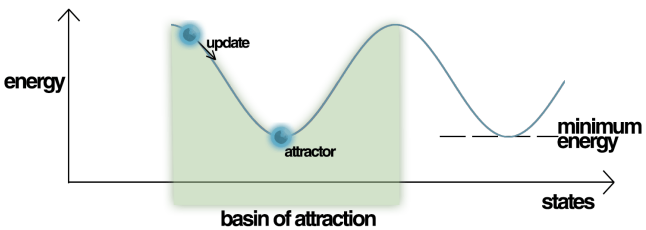


图 8.5: Hopfield 网络。图片来源于 https://en.wikipedia.org/wiki/Hopfield_network。

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，

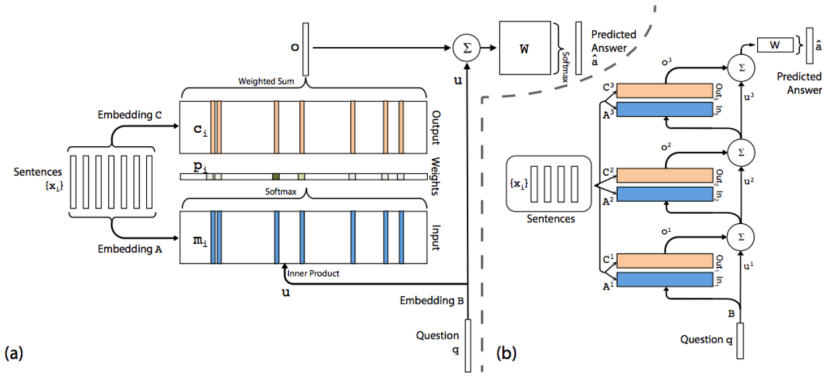


图 8.6: 端到端记忆网络。图像来源：[Sukhbaatar et al., 2015]

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。

8.3.1 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

$$\mathbf{attention}((k, v)_{1:N}, q) = \sum_{i=1}^{N} \alpha_i \mathbf{v}_i, \tag{8.20}$$

$$= \sum_{i=1}^{N} \frac{\exp\Big(s(\mathbf{k}_i, \mathbf{q})\Big)}{\sum_j \exp\Big(s(\mathbf{k}_j, \mathbf{q})\Big)} \mathbf{v}_i \tag{8.21}$$

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

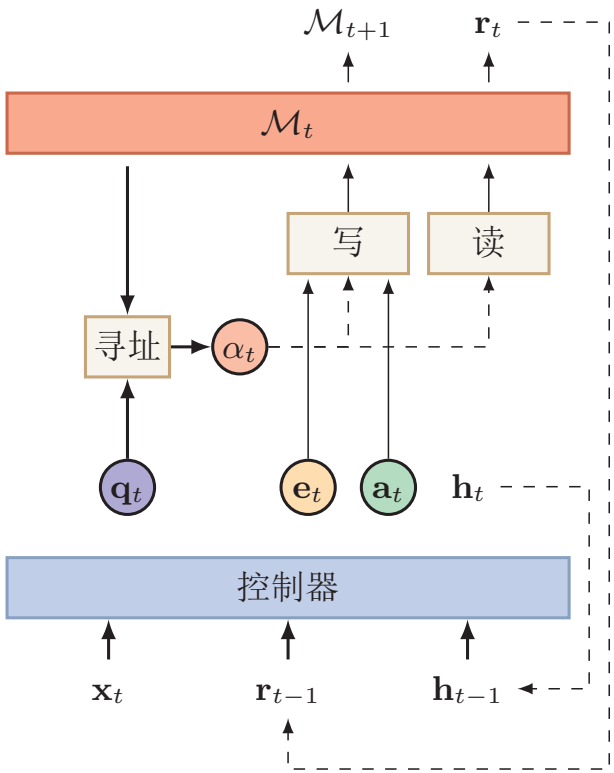


图 8.7: 神经图灵机示例。在每个时刻t，控制器接受输入$\mathbf{x}_t$，上一时刻的输出$\mathbf{h}_{t-1}$和读向量$\mathbf{r}_{t-1}$，并产生输出$\mathbf{h}_t$，同时生成和读写外部记忆相关的三个向量：查询向量$\mathbf{q}_t$，删除向量$\mathbf{e}_t$和增加向量$\mathbf{a}_t$。然后对外部记忆$\mathcal{M}_t$进行读写操作，生成读向量$\mathbf{r}_t$，和新的外部记忆$\mathcal{M}_{t+1}$

。

8.3.2 神经图灵机

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

控制器 (controller) 为一个前馈或循环神经网络。以循环神经网络为例，在每个时刻t，控制器接受当前时刻的输入$\mathbf{x}_t$，上一时刻的输出$\mathbf{h}_{t-1}$和上一时刻从外部记忆中读取的信息$\mathbf{r}_{t-1}$，并产生输出$\mathbf{h}_t$，同时生成和读写外部记忆相关

的三个向量：查询向量$\mathbf{q}_t$，删除向量$\mathbf{e}_t$和增加向量$\mathbf{a}_t$。

神经图灵机中的外部记忆是可以可读写的，其结构如图8.7所示。

读操作 在时刻t, 外部记忆的内容记为$\mathcal{M}_t = \mathbf{m}_{t,1}, \cdots, \mathbf{m}_{t,n}$，读操作为从外部记忆$\mathcal{M}_t$中读取信息$\mathbf{r}_t \in \mathbb{R}^d$。

神经图灵机中还实现了比较复杂的基于位置的寻址方式。这里我们只介绍比较简单的基于内容寻址方式，整个框架不变。

首先通过注意力机制来进行基于内容的寻找，即

$$\alpha_{t,i} = \text{softmax}(s(\mathbf{m}_{t,i}, \mathbf{q}_t)) \tag{8.22}$$

其中$\mathbf{q}_t$为控制器产生的查询向量，用来进行基于内容的寻址。$s(\cdot,\cdot)$为加性或乘性的打分函数。注意力分布$\alpha_{t,i}$是记忆片段$\mathbf{m}_{t,i}$对应的权重，并满足$\sum_{i=1}^{n} \alpha_{t,i} = 1$。

根据注意力分布α_t，可以计算读向量（read vector）$\mathbf{r}_t$

$$\mathbf{r}_t = \sum_{i=1}^{n} \alpha_i \mathbf{m}_{t,i}. \tag{8.23}$$

读向量$\mathbf{r}_t$作为下一个时刻控制器的输入。

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。

首先，控制产生删除向量（erase vector）$\mathbf{e}_t$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。

删除操作是根据注意力分布来按比例地在每个记忆片段中删除$\mathbf{e}_t$，增加操作根据注意力分布来进行按比例地给每个记忆片段加入$\mathbf{a}_t$。

$$\mathbf{m}_{t+1,i} = \mathbf{m}_{t,i}(\mathbf{1} - \alpha_{t,i}\mathbf{e}_t) + \alpha_{t,i}\mathbf{a}_t, \forall i \in [1, n]. \tag{8.24}$$

通过写操作得到下一时刻的外部记忆$\mathcal{M}_{t+1}$。

8.3.3 记忆结构

Chandar et al. [2016]

记忆周期	计算机	人脑	神经网络
短期	寄存器	短期记忆	状态（神经元活性）
中期	内存	工作记忆	外部记忆
长期	外存	长期记忆	可学习参数
存储方式	随机寻址	内容寻址	内容寻址为主

表 8.1: 不同领域中记忆模型的不严格类比。

8.4 总结和深入阅读

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。

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。

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。

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。

习题 8-2 在Hopfield网络中，每次迭代都会导致网络能量的变小或不变。

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.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. Structured attention networks. *arXiv preprint arXiv:1702.00887*, 2017.

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 vi-

sual 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.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J Smola, and Eduard H Hovy. Hierarchical attention networks for document classification. In *HLT-NAACL*, pages 1480–1489, 2016.