arXiv:2304.04556v1 [cs.LG] 7 Apr 2023

注意力：边缘概率就是你所需要的一切吗?

Ryan Singh [1] Christopher L. Buckley [1]

摘要

注意力机制是认知系统的一项核心特性，使其能够灵活地选择性调配认知资源。神经科学领域长期以来一直在研究注意力，并提出了许多试图捕捉其核心属性的现象学模型。近年来，注意力机制已成为机器学习中的主流架构选择，也是 Transformer 的核心创新。其发展所依托的主流直觉与形式体系，借鉴了数据库管理系统中键与查询的理念。在这项工作中，我们提出注意力机制的一种替代性贝叶斯基础，并展示了它如何统一机器学习中的不同注意力架构。这一表述有助于找出不同注意力机器学习架构之间的共性，也为其与神经科学中发展出的架构搭建桥梁。我们希望这项工作能引导人们形成对注意力架构关键属性的更深入直觉，并提出新的属性。

1. 引言

设计具有有利归纳偏置的神经网络架构，是近期深度学习取得诸多成功的基础（Baxter，2000）。尤其是，注意力机制使语言模型具备了此前被认为不可能实现的类人生成能力（Vaswani 等人，2017）。注意力机制作为一种与领域无关的架构取得了成功，促使其被广泛应用于各种任务和领域，尤其是在视觉推理和分割任务中达到了最先进的性能（Dosovitskiy 等人，2021；Wang 等人，2022）。

尽管取得了成功，注意力机制的作用仍未得到充分理解。事实上，尚不清楚其究竟在多大程度上它与启发了这一机制的认知注意力理论的关联程度（Lindsay, 2020年）。在此，我们旨在基于概率推断原理，提供一种简约的描述。这一贝叶斯视角既提供了指定先验信念的原则性方法，也能明确推理注意力变量的作用。此外，理解其基本计算使我们能够统一描述文献中的不同注意力机制。这分为两部分展开。

首先，我们展示了“软”注意力机制（例如自注意力、交叉注意力、图注意力，以下简称 Transformer 注意力）可以从概率角度理解为对可能连接结构取期望，从而在基于 softmax 的注意力与边缘似然之间建立了有趣的联系。

其次，我们将连接关系的不确定性扩展到贝叶斯框架中，从而为迭代注意力机制（槽位注意力、Perceiver 和块槽注意力）（Locatello 等人，2020年；Singh 等人，2022；Jaegle 等人，2021）以及现代连续霍普菲尔德网络（Ramsauer 等人，2021）提供了理论基础。

此外，我们将迭代注意力应用于预测编码网络。这一理论是计算神经科学领域的重要理论，由此在机器学习与认知科学之间架起了一座新的理论桥梁。

$$
\begin{aligned}
Attention(Q,K,V) &= \overbrace{softmax(\frac{QW_QW_K^TK^T}{\sqrt{d_k}})}^{p(E\mid Q,K)}V \\
&= \mathbb{E}_{p(E|Q,K)}[V]
\end{aligned}
$$

一个关键发现是，注意力矩阵可以视为关于邻接结构 E 的后验分布，而整个机制则是在关于键与查询之间可能存在关系的后验信念基础上，计算值函数 $V(X)$ 的期望。

该形式化方法提供了一个替代性的贝叶斯理论框架，用于理解注意力模- 型；这与最初以数据库管理系统和数据检索为基础的框架形成对比，并提供

[1]萨塞克斯大学工程与信息学学院。通信联系人：RyanSingh <rs773@sussex.ac.uk>。

一个用于描述不同注意力架构的统一框架。仅根据其边关系描述它们之间的差异，从而支持对新架构进行更有效的分析和开发。此外，还为理解硬注意力模型与软注意力模型之间的差异提供了有原则的解释。

贡献

- 理解注意力机制的统一概率框架。
- 我们展示了自注意力和交叉注意力可以被视为在可能的网络结构上计算边缘似然。
- 我们展示了，槽位注意力、块槽位注意力和现代连续霍普菲尔德网络都可以被视为坍缩变分推断，其中可能的网络结构构成坍缩变量。
- 从计算神经科学出发，通过预测编码网络的视角，为贝叶斯注意力概念提供桥梁。
- 提供一个用于推理硬注意力以及注意力机制的高效近似方法的框架。

2. 相关工作

注意力作为双层优化 将前馈架构映射为相关能量函数上的一个最小化步骤，这被称为展开优化（Frecon 等人，2022）。采用这一视角，有助于深入理解各类架构所涉及的归纳偏置。已有研究表明，交叉注意力机制可视为在某种形式的霍普菲尔德网络的能量函数上执行的优化步骤（Ramsauer 等人，2021），从而揭示了注意力与联想记忆之间的联系。尽管（Yang 等人，2022）将这一观点扩展到了自注意力，但我们的框架区分了霍普菲尔德注意力与标准注意力机制：前者不允许使用任意值矩阵。尽管两者仍存在紧密的理论联系，但该框架将霍普菲尔德能量视为变分自由能的一种实例，与槽注意力等迭代注意力机制更加一致。

与高斯混合模型的关系 以概率视角研究注意力机制的先前工作指出，它与高斯混合模型中的推断有关（Gabbur 等人，2021年；Nguyen 等人，2022；Ding 等人，2020年）。事实上（Annabi 等人，2022）直接展示了 Hopfield 能量与高斯混合模型变分自由能之间的联系。尽管高斯混合模型是我们在此提出的框架的一种特例，足以解释交叉注意力，但无法涵盖槽位或自注意力。此外，我们的框架能够将结构性归纳偏置扩展到高斯混合模型所能表达的范围之外，并捕捉其与硬注意力之间的关系。

潜在对齐与硬注意力人们已经多次尝试结合软注意力（可微性）与硬注意力的优势。大多数方法通过采样来进行，例如使用 REINFORCE 估计器（Deng etal.，2018）或 $topK$ 近似（Shankaret al.，2018）。与我们的方法最为相似的一种方法在一次前向传递中嵌入了完整的前向-后向算法（Kimet al.，2017），而我们的方法则通过对隐式图模型进行边缘化，以简洁的方式进行描述。

折叠推断 坍缩变分推断最著名的应用是在主题建模中（Teh 等人，2006年）。据我们所知，将折叠推断与深度学习中的注意力联系起来是全新的。

3. Transformer 注意力

3.0.1. 注意力作为期望

我们首先通过说明 Transformer 注意力最好被视为对潜变量的期望来展开。在自注意力和交叉注意力的情况下，这是相对于可能的邻接结构而言的神经网络的期望。

设 $x = (x_1, .., x_n)$ 为观测变量，ϕ 为一组潜变量，y 为我们需要预测的变量。给定潜变量模型 $p(y, x, \phi) = p(y \mid x, \phi)p(x, \phi)$，其中 $p(y \mid x, \phi)$ 由某个函数 $v(y, x, \phi)$ 参数化，例如神经网络。

我们的目标是求得 $p(y \mid x)$，但 ϕ 是未观测的，因此我们计算边缘似然。

$$p(y \mid x) = \sum_{\phi} p(\phi \mid x) v(y, x, \phi)$$

重要的是，softmax 函数是表示后验概率的一种自然方式

$$p(\phi \mid x) = \frac{p(x, \phi)}{\sum_{\phi} p(x, \phi)}$$

$$p(\phi \mid x) = softmax(\ln p(x, \phi))$$

因此，Transformer 注意力可以看作是按后验分布 $p(\phi \mid x)$ 对$v(x, \phi)$加权。

$$\begin{aligned} p(y \mid x) &= \sum_{\phi} softmax(\ln p(x, \phi)) v(y, x, \phi) \\ &= \mathbb{E}_{p(\phi \mid x)}[v(y, x, \phi)] \end{aligned} \tag{1}$$

我们声称，公式（1）正是自注意力和交叉注意力背后的方程。为了更直接地建立联系，我们给出了与它们对应的具体生成模型。潜变量 ϕ 被视为观测变量之间可能存在的关系，或边，其中每个观测变量x （键和查询）。

用于对这些图关系进行建模的一种自然形式化方法是马尔可夫随机场。

3.0.2. 成对马尔可夫随机场

给定一组随机变量 $X = (X_v)_{v\in V}$，其概率分布为 $[p]$，以及一个图 $G = (V, E)$。如果联合密度函数 $P(X = x) = p(x)$ 按以下方式分解，则这些变量相对于 G 构成一个成对马尔可夫随机场（MRF）。

$$p(x) = \frac{1}{Z}\exp\left(\sum_{v\in V}\psi_v + \sum_{e\in E}\psi_e\right)$$

其中 Z 是配分函数 $\psi_v(x_v)$，而 $\psi_e = \psi_{u,v}(x_u, x_v)$ 分别称为节点势函数和边势函数[1]。

除了典型的设定之外，我们还在底层图的邻接结构上添加了结构先验 $p(E)$。

$$\begin{aligned} p(x, E) &= P(x \mid E)P(E) \\ &= \frac{1}{Z}p(E)\exp\left(\sum_{v\in V}\psi_v + \sum_{e\in E}\psi_e\right) \end{aligned}$$

我们简要指出，(1)遵循 $[p]$ 的因子分解，具体而言；如果该分布关于潜变量 $p(x, \phi) = \prod_i f_i(x, \phi_i)$和 $v(x, \phi) = \sum_i v_i(x, \phi_i)$ 那么（应用期望的线性性）我们可以写成

$$\mathbb{E}_{p(\phi|x)}[v(x, \phi)] = \sum_i \mathbb{E}_{p(\phi_i|x)}[v_i] \tag{2}$$

允许对每个因子独立进行边缘化。

对于MRF，这样的因子分解是自然的。如果关于边的分布可分解为局部分布$p(E) = \prod_i p(E_i)$（利用MRF的独立性），我们可以写成 $p(x, E) = \frac{1}{Z}\prod_i f_i(x, E_i)$，其中每个$f_i = P(E_i)\exp\sum_{v\in V}\psi_v\sum_{e\in E_i}\psi_e$本身都是一个未归一化的 MRF。

要恢复交叉注意力和自注意力，这类模-型只需指定一个结构先验和势-函数。

3.0.3. 交叉注意力

- Key nodes $K = (x_1, .., x_n)$

[1]参见（Shah 等人，2021）以了解精确定义。

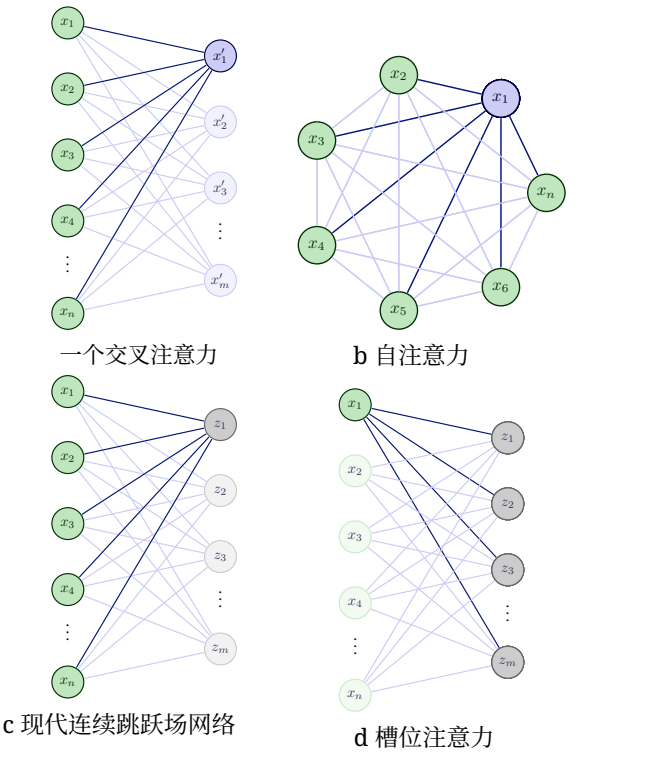


图 1. 文献中不同注意力模块的比较，突出显示的边代表对随机变量进行的边缘化 E_1，在 1a 和 1b中，所有节点均为观测节点，与 1c 和 1d不同，后者存在潜在节点（以灰色表示）。

- 查询节点 $Q = (x'_1, ..., x'_m)$
- 结构先验 $p(E) = \prod_{i=1}^{m} p(E_i)$，其中 $E_i \sim Uniform\{(x_1, x'_i) .., (x_n, x'_i)\}$，使得每个查询节点与每个键节点相连的概率均匀。
- 边势函数 $\psi(x_j, x'_i) = x_i'^T W_Q^T W_K x_j$，实际上用于衡量经过某种变换后的 x_j 与 x'_i 之间的相似度。
- 值函数 $V_i(K, Q, E_i) = W_V x_{s(E_i)}$，即应用于节点的线性变换；$x_{s(E_i)}$，即边的起点 E_i。

对两个因子中定义的每个因子取后验期望（2），即可得到标准的交叉注意力机制

$$\mathbb{E}_{p(E_i|Q,K)}[V_i] = \sum_j softmax_j(x_i'^T W_Q^T W_K x_j) W_V x_j$$

$$\mathbb{E}_{p(E|Q,K)}[V] = softmax(Q^T W_Q^T Q_K K) W_V K$$

3.0.4. 自注意力

- Nodes $K = Q = (x_1, .., x_n)$
- 结构先验 $p(E) = \prod_{i=1}^{n} p(E_i^{\rightarrow})$，其中 $E_i^{\rightarrow} \sim Uniform\{(x_1, x_i) .., (x_n, x_i)\}$，使得每个节点与其他每个节点连接的可能性均等。
- 边势函数 $\psi(k_j, k_i) = x_i^T W_Q^T W_K x_j$，实际上衡量的是 x_j 与 x'_i 在某种变换下的相似度。
- 值函数 $V_i(K, Q, E_i) = W_V x_{s(E_i)}$，应用于节点的线性变换，$x_{s(E_i)}$，边 E_i 的起点。

再次，对二（中定义的每个因子取后验期望2）即可得到标准的自注意力机制

$$\mathbb{E}_{p(E_i|Q,K)}[V_i] = \sum_j softmax_j(x_i^T W_Q^T W_K x_j) W_V x_j$$

$$\mathbb{E}_{p(E|Q,K)}[V] = softmax(K^T W_Q^T W_K K) W_V K$$

4. 迭代注意力

我们继续将注意力扩展到完整的贝叶斯推断。从本质上说，就是将注意力技巧——对注意力变量进行边缘化——应用于变分自由能（即证据下界）。

现代连续霍普菲尔德网络可视为这类系统的一个特例，使我们能够在概率框架下复现（Ramsauer 等人，2021年）中的“霍普菲尔德注意力”更新。在不同的结构先验下，我们可以得到其他迭代注意力模型：槽位注意力（Locatello 等人，2020年）、块槽注意力（Singh 等人，2022）以及 Perciever（Jaegle 等人，2021）。此外，我们还展示了贝叶斯注意力的一项特定优势：硬注意力。

4.0.1. 折叠推断

我们介绍一种坍缩变分推断（Teh 等人，2006年）的版本，展示它如何产生贝叶斯注意力机制。之所以称为注意力机制，是因为变分更新（6）与神经注意力机制（1）在形式上惊人地相似。

我们的设定是潜变量模型 $p(x, z, \phi)$，其中 x是观测变量，而 z、ϕ 是潜变量。通常我们希望给定 x 推断 z。

折叠推断通过边际化消去多余的潜变量 ϕ

$$p(x, z) = \sum_{\phi} p(x, z, \phi) \tag{3}$$

我们定义一个识别密度 $q(z) \sim N(z; \mu)$，并针对该分布的参数 μ 优化变分自由能。

$$\min_{\mu} F(x, \mu) = \mathbb{E}_q[\ln q_\mu(z) - \ln p(x, z)]$$

在典型的拉普拉斯近似下，我们可以将变分自由能写为 $F \approx -\ln p(x, \mu)$ [2]。代入（3）并对变分参数求导，可得，

$$F(x, \mu) = -\ln \sum_{\phi} p(x, \mu, \phi)$$

$$\frac{\partial F}{\partial \mu} = -\frac{1}{\sum_\phi p(x, \mu, \phi)} \sum_{\phi} \frac{\partial}{\partial \mu} p(x, \mu, \phi) \tag{4}$$

这将贝叶斯注意力与标准注意力（1）联系起来。为阐明这一点，我们采用对数导数技巧，代入 $p_\theta = e^{\ln p_\theta}$ 并以两种方式重新表达（4）：

$$\frac{\partial F}{\partial \mu} = -\sum_{\phi} softmax_\phi(\ln p(x, \mu, \phi)) \frac{\partial}{\partial \mu} \ln p(x, \mu, \phi) \tag{5}$$

$$\frac{\partial F}{\partial \mu} = \mathbb{E}_{p(\phi|x,\mu)}[-\frac{\partial}{\partial \mu} \ln p(x, \mu, \phi)] \tag{6}$$

第一种形式揭示了在所有注意力模型中无处不在的 softmax。第二种则暗示了变分无码高清

[2] 更具原则性的推导（考虑了高阶项）见附录

更新应被评估为典型变分梯度（方括号内的项）相对于由随机变量 ϕ 表示的参数后验分布的期望。

换言之，贝叶斯注意力正是迭代应用神经注意力机制，其中值函数是变分自由能梯度。我们先推导一般 MRF 的更新规则，然后通过指定特定分布，再次得到文献中的（迭代式）注意力模型。

4.0.2. 边缘化 MRF 的自由能

回顾因子化 MRF，$p(E) = \prod_i p(E_i)$。$p(x, E) = \frac{1}{Z}\prod_i f_i(x, E_i)$，其中每个 $f_i = P(E_i)\exp\sum_{v\in V}\psi_v\sum_{e\in E_i}\psi_e$。独立性意味着，折叠推断所需的边缘化可以得到简化

$$\sum_E p(x, E) = \frac{1}{Z}\prod_i\sum_{E_i} f_i(x, E_i)$$

在推断场景中，节点被划分为观测节点 x和潜变量节点 z。变分自由能（4）及其导数的相关形式可以表示为

$$F(x, \mu, \theta) = -\sum_i \ln\sum_{E_i} f_i(x, \mu, E_i)$$

$$\frac{\partial F}{\partial \mu_j} = -\sum_i\sum_{E_i} softmax(f_i(x, \mu, E_i))\frac{\partial f_i}{\partial \mu_j}$$

与硬注意力方法类似，随机变量E 是一个显式的对齐变量。然而，与硬注意力不同，我们采用上述的折叠推断方法，避免显式推断 E 。

4.0.3. 二次势函数与凸凹过程

我们沿用（Ramsauer 等人，2021）的方法，使用 CCCP 推导一个固定点方程，这必然会降低自由能。

假设节点势函数是二次的 $\psi(x_i) = -\frac{1}{2}x_i^2$，且边势函数的形式为 $\psi(x_i, x_j) = x_i W x_j$。

$$\mu_j^* = \sum_i\sum_{E_i} softmax(g_i(x, \mu, E_i))\frac{\partial g_i}{\partial \mu_j} \qquad (7)$$

其中 $g_i = \sum_{e\in E_i}\psi_e$。

通过 CCCP（Yuille& Rangarajan，2001），这个不动点方程具有如下性质 $F(x, \mu_j^*, \theta) \leq F(x, \mu_j, \theta)$，当且仅当 μ_j^*是 F的驻点时取等号。

我们沿用第 3 节的做法，通过指定特定的结构先验和势函数来还原不同的迭代注意力机制。

4.0.4. 霍普菲尔德风格交叉注意力

令观测节点 $x = (x_1, .., x_n)$ 和潜在节点 $z = (z_1, .., z_m)$ 具有以下结构先验 $p(E) = \prod_{i=1}^m p(E_i)$，其中 $E_i \sim Uniform\{(x_1, z_i) .., (x_n, z_i)\}$。并定义边势函数 $\psi(x_j, z_i) = z_i Q^T K x_j$，应用 (7)

$$\mu_i^* = \sum_j softmax_j(\mu_i W_Q^T W_K x_j) W_Q^T W_K x_j$$

当 μ_i 被初始化为某个查询 ξ 时，系统（Ramsauer 等人，2021）的固定点更新为$\mu_i^*(\xi) = \mathbb{E}_{p(E_i|x,\xi)}[W_Q^T W_K x_{t(E_i)}]$。当模式 x彼此充分分离时，$\mu_i^*(\xi) \approx W_Q^T W_K x_j$，其中 $W_Q^T W_K x_j$是最接近的向量，因此可以用作联想记忆。

4.0.5. 槽位注意力

槽位注意力（Locatello 等人，2020年）是一种以物体为中心的学习模块，构建于迭代注意力机制之上。这里我们展示，这只是对边集合上的先验信念进行简单调整。

在相同的节点集合和势函数下，将边先验替换为 $p(E) = \prod_{j=1}^n p(E_j)$ $E_j \sim \{(x_1, z_1), .., (x_j, z_m)\}$ $Uniform$

$$\mu_i^* = \sum_j softmax_i(\mu_i Q^T K x_j) Q^T K x_j$$

尽管最初的槽位注意力使用 RNN来辅助此处所示的基本更新，但重要特征在于，softmax 是在“槽位”上进行的，μ。这会迫使槽位之间展开竞争，以解释观测变量，从而形成以对象为中心的表示。例如，如果观测变量 x 是图像块，那么槽位就会被迫将相似的图像块聚类在一起，以提高这些图像块的总体似然。聚类一词非常准确；事实上，这一机制与在高斯混合模型上执行一步 EM 完全等价。

4.0.6. 块槽注意力

(Singh 等人，2022) 建议将联想记忆能力与以对象为中心的槽位式能力相结合，并为此提供一种迭代方案，在槽位注意力更新和霍普菲尔德更新之间交替进行。

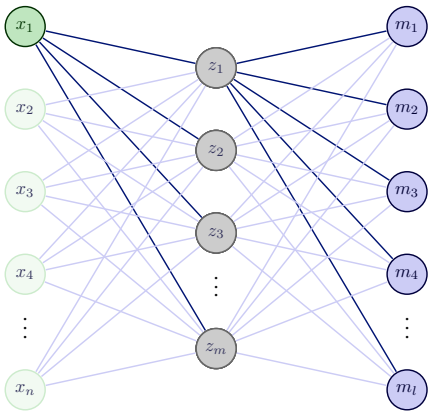


图 2. 块槽注意力

我们的框架允许我们通过不同的潜在图结构灵活组合不同的注意力机制，从而推导出一种由模型提供信息的块槽注意力版本。在这种设定下，我们有三组变量：X观测值、Z 待推断的潜变量以及 M 参数。

定义成对马尔可夫随机场 $X = \{x_1, ..., x_n\}$, $Z = \{z_1, ..., z_m\}$ 和 $M = \{m_1, ..., m_l\}$，其边先验为 $p(E) = \prod_{j=1}^{m} p(E_j) \prod_{k=1}^{l} p(\tilde{E}_k)$, $E_j \sim Uniform\{(x_j, z_1) .., (x_j, z_m)\}$, $E_k \sim Uniform\{(z_1, m_k) .., (z_m, m_k)\}$, 其中 X和 Z之间的边势函数为 $\psi(x_j, z_i) = z_i Q^T K x_j$，而 Z 和 M 之间的边势函数为 $\psi(z_i, m_k) = z_i \cdot m_k$

应用（7）得到

$$\mu_i^* = \sum_j softmax_i(\mu_i Q^T K x_j) Q^T K x_j + \sum_k softmax_k(\mu_i \cdot m_k) m_k$$

在原始的块槽注意力中，每个槽位 z_i 被拆分为多个块，其中每个块都可以访问块特定记忆，即 $z_i^{(b)}$ 可以与记忆节点建立可能的连接 $\{m_k^{(b)}\}_{k \le l}$。这样一来，便可通过槽位表示对象，而槽位又将每个对象的特征解耦到不同的块中。上文展示了单块版本，但很容易看出，将更新扩展到多块版本并应用（7）可得

$$\mu_i^* = \sum_j softmax_i(\mu_i Q^T K x_j) Q^T K x_j + \sum_{k,b} softmax_k(\mu_i^{(b)} \cdot m_k^{(b)}) m_k^{(b)}$$

5. 预测编码网络

预测编码网络（PCN）已成为计算神经科学中一种颇具影响力的理论（Rao & Ballard，1999；Friston & Kiebel，2009；Buckley 等人，2017）。PCN 建立在知觉即推断和贝叶斯脑理论之上，通过最小化变分自由能来执行近似贝叶斯推断，而变分自由能体现为局部预测误差的最小化。单个神经元的连续时间动力学由下式给出

$$\frac{\partial \mathcal{F}}{\partial \mu_i} = -\sum_{\phi^-} k_\phi \epsilon_\phi + \sum_{\phi^+} k_\phi \epsilon_\phi w_\phi$$

其中 ϵ 为预测误差，w 代表突触强度以及 k 为表示生成模型中不确定性的节点特定精度（Millidge 等人，2022）。

一个自然的扩展是对传入和传出连接集合应用折叠推断，即采用局部因子化的先验来描述可能的连接性。用上一节的记法，我们有一个具有层级结构的 MRF $Z = \{Z^{(0)}, ..., Z^{(l)}, ..., Z^{(N)}\}$，其中边上的先验因子化为逐层的 $p(E^{(l)}) = \{(z_i, z_j) : (z_i, z_j) \in Z^{(l-1)} \times Z^{(l)}\}$，势函数为 $\phi(z_i, z_j) = \epsilon_{i,\ j}^2 = k_j(z_j - w_{i,\ j} z_i)^2$。

$$\frac{\partial \mathcal{F}}{\partial \mu_i} = -\sum_{\phi^-} softmax(-{\epsilon_\phi}^2) k_\phi \epsilon_\phi + \sum_{\phi^+} softmax(-{\epsilon_\phi}^2) k_\phi \epsilon_\phi w_\phi$$

由此产生的动力学通过 softmax 函数，对神经元接收到的预测误差进行“归一化”。这与心理学和神经科学中将注意力视为归一化的理论十分契合。相比之下，先前基于预测编码的注意力理论着重关注精度项，k，因为它们能够增强或减弱预测误差的影响（Feldman& Friston，2010）。在这里，我们看到 softmax 项也能实现这种调节，同时还表现出与认知注意力相关的快速赢家通吃动力学。

5.1. 讨论

在本节中，我们将简要讨论将注意力机制视为一个推断问题可以带来哪些收获。

5.1.1. 硬注意力

回想一下（1）神经注意力可以被视为对潜变量求期望 $\mathbb{E}_{p(\phi|x)}[v(x, \phi)]$。这里的

该机制之所以称为“软”，是因为我们对注意力变量 ϕ 的多种可能性进行加权。另一方面，硬注意力则从 $p(\phi \mid x)$ 中抽取单个样本。据称，这种方式更符合生物学特性、更易于解释，且计算复杂度更低。过去，硬注意力表现较差被归因于其难以训练以及具有随机性。然而，我们将软注意力视为精确边缘化的框架提供了另一种解释。与精确边缘化（软注意力）相比，随机近似（硬注意力）总会处于劣势。此外，我们的框架还提供了一种无缝切换硬注意力与软注意力的方法。由于分布 $p(\phi \mid x)$ 是分类分布，因此在任意时刻（训练期间或推断时）都可以通过从 $p(\phi \mid x)$ 中抽取单个样本 ϕ^* 来实现硬注意力，从而得到 $v(x, \phi^*)$。

这种折叠注意力分布的方法存在两个问题。首先，单个样本会消除所有不确定性；其次，为了进行采样而计算 $p(\phi \mid x)$，仍会产生二次惩罚 $O(n^2)$。不过，我们可以运用概率论中的工具来帮助我们分- 析采样的成本，以及注意力分布的线性近似。

5.1.2. 高效Transformer

考虑某个分布 q 试图近似$p(\phi \mid x)$，我们可以用相对熵来量化信息损失

$$\mathcal{L}[p, q] \triangleq D_{KL}[q(\phi) \,||\, p(\phi \mid x)] = H[q] + \mathbb{E}_q[p(\phi \mid x)]$$

在硬注意力近似中，使用从p中抽取的一个样本来近似 $\mathcal{L}[p, q] = -\ln p(\phi^* \mid x)$，这或许符合直觉 $\mathbb{E}[\mathcal{L}] = H[p]$，即当注意力分布的熵较低时，硬注意力是一种很好的近似，而熵可以通过温度参数控制（附录 ??）。

许多注意力的高效替代方案，例如低秩和线性近似，都可以表述为用 $q(\phi \mid x)$近似 $p(\phi \mid x)$，其中计算 q 的成本低于精确边际化。估计 $\mathcal{L}$ 可用于量化采用这些替代方案时相对的信息损失。另一种降低注意力机制计算复杂度的方法是稀疏化注意力矩阵；在我们的框架中，这归结为对边先验的调整（附录 ??）。

5.1.3. 新设计

所呈现的描述与以往概率描述之间的主要区别在于：相对于一个隐含的概率模型，将软注意力视为一种有原则且精确的概率计算，而不是一种贫乏的近似。这带来了以下可能性：

通过改变注意力机制所边缘化的分布来设计新的注意力机制，可以调整结构先验或势函数。我们希望这能帮助人们以有章可循的方式设计新的架构。

参考文献

Annabi,L.、Pitti, A. 和 Quoy, M.《变分推断与自动联想记忆之间的关系》，2022 年 10 月。网址 http://arxiv.org/abs/2210.08013。arXiv:2210。08013 [cs]。

巴克斯特，J. 归纳偏置学习模型。 人工智能研究杂志，12:149–198，2000年3月。ISSN1076-9757。doi: 10.1613/jair.731。网址 https://www.jair.org/index.php/jair/article/view/

Buckley, C. L., Kim, C. S., McGregor, S., and Seth, A. K. 行动与知觉的自由能原理：数学综述。数学心理学杂志，81:55–79，2017年12月。ISSN0022-2496。doi: 10.1016/j.jmp.2017.09.004。URLhttps://www.sciencedirect.com/science/文章/pii

Deng, Y., Kim, Y., Chiu, J., Guo, D., 和 Rush, A.《潜在对齐与变分注意力》。载于神经信息处理系统进展，第31卷。Curran Associates, Inc.，2018年。网址：https://proceedings.neurips.cc/paper/2018/has h/b6

Ding, N., Fan, X., Lan, Z., Schuurmans, D. 和 Soricut, R.《无法消除解释的注意力》，2020年9月。网址：http://arxiv.org/abs/2009.14308。arXiv:2009.14308 [cs, stat]。

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. 和 Houlsby, N.《一张图像胜过16×16个词：用于大规模图像识别的变换器》，2021年6月。网址：http://arxiv.org/abs/2010.11929。arXiv:2010.11929 [cs] 版本：2。

Feldman, H. 和 Friston, K. 注意力、不确定性与自由能。人类神经科学前沿，4，2010 年。ISSN 1662-5161。网址 https://www.frontiersin.org/articles/10.3389/fnhu

Frecon,J.、Gasso, G.、Pontil, M. 和 Salzo, S. Bregman 神经网络。载于 第39届国际机器学习会议论文集，第6779–6792页。PMLR，2022年6月。网址：https://proceedings.mlr.press/v162/frecon22a.html ISSN：2640-3498.

Friston, K. 和 Kiebel, S. 基于自由能原理的预测编码。英国皇家学会哲学汇刊 B辑：生物科学，364(1521):1211–1221，2009年5月。ISSN0962-8436。doi: 10.1098/rstb.2008.0300。URLhttps://www.ncbi.nlm.nih.gov/pmc/artic

Gabbur,P.、Bilkhu, M.和Movellan, J.《交互式分割的概率注意力》，2021年7月。URL http://arxiv.org/abs/2106.15338。arXiv:2106.15338 [cs]。Jaegle,A.、Gimeno, F.、Brock, A.、Vinyals, O.、Zis- serman,A.和 Carreira, J.《Perceiver：通过迭代注意力实现通用感知》。载于 第38届国际机器学习会议论文集，第4651–4664页。PMLR，2021年7月。URLhttps://proceedings.mlr.press/v139/jaeg

ISSN: 2640-3498。Kim, Y.、Denton, C.、Hoang,L.和 Rush, A. M.《结构化注意力网络》，2017年2月。URL http://arxiv.org/abs/1702.00887。arXiv:1702。00887 [cs]。Lindsay, G. W. 心理学、神经科学和机器学习中的注意力。《计算神经科学前沿》，14，2020年。ISSN 1662-5188。URLhttps://www.frontiersin.org/articles/10

Locatello, F.、Weissenborn, D.、Unterthiner, T.、Mahendran, A.、Heigold, G.、Uszkoreit, J.、Doso- vitskiy, A. 和 Kipf, T.《基于槽位注意力的以对象为中心的学习》，2020年10月。网址http://arxiv.org/abs/2006.15055。arXiv:2006。15055 [cs, stat]。

Millidge, B.、Song, Y.、Salvatori, T.、Lukasiewicz, T. 和 Bogacz, R.《预测编码网络中推断与学习的理论框架》，2022年8月。网址 http://arxiv.org/abs/2207.12316。arXiv:2207。12316 [cs]。

Nguyen, T. M.、Nguyen, T. M.、Le, D. D. D.、Nguyen, D. K.、Tran, V.-A.、Baraniuk, R.、Ho, N. 和 Osher, S.《利用概率注意力键改进 Transformer》。收录于 第39届国际机器学习会议论文集，第16595–16621页。PMLR，2022年6月。网址 https://proceedings.mlr.press/v162/nguy

ISSN：2640-3498。Ramsauer, H., Sch¨afl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlovi´c, M., Sandve, G. K., Greiff, V., Kreil, D., Kopp, M., Klambauer, G., Brandstetter, J. 和 Hochreiter, S.《霍普菲尔德网络就是你所需要的一切》，2021年4月。网址 http://arxiv.org/abs/2008.02217.arXiv:2008.02217 [cs, stat]. Rao, R. P. N. 和 Ballard, D. H. 视觉皮层中的预测编码：对某些经典外感受野效应的功能性解释。《自然·神经科学》，2(1):79–87，1999年1月。ISSN 1546-1726。doi：10.1038/4580。网址 http://www.nature.com/articles/nn0199_79。编号：1。出版商：自然出版集团。Shah, A.、Shah, D. 和 Wornell, G. 关于连续成对马尔可夫随机场的学习。载于 第24届国际人工智能与统计会议论文集，第1153–1161页。PMLR，2021年3月。网址 https://proceedings.mlr.press/v130/shah21a.html。ISSN：2640-3498。

Shankar, S.、Garg, S. 和 Sarawagi, S.《令人惊讶地简单的序列到序列学习硬注意力》。载于2018 年自然语言处理实证方法会议论文集，第 640–645 页，比利时布鲁塞尔，2018年10月。计算语言学协会。doi: 10.18653/v1/D18-1065。网址: https://aclanthology.org/D18-1065。

Singh, G.、Kim, Y. 和 Ahn, S.《神经区块槽表示》，2022年11月。网址：http://arxiv.org/abs/2211.01177。arXiv：2211.01177 [cs]。

Teh, Y.、Newman, D. 和 Welling, M.《用于潜在狄利克雷分配的折叠变分贝叶斯推断算法》。载于神经信息处理系统进展，第 19 卷。麻省理工学院出版社，2006年。网址: https://proceedings.neurips.cc/paper_files/paper/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. 注意力就是一切，2017年12月。URL http://arxiv.org/abs/1706.03762。arXiv:1706.03762 [cs]。Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., and Wei, F. 图像作为外语：面向所有视觉和视觉语言任务的BEIT预训练，2022年8月。URL http://arxiv.org/abs/2208.10442。arXiv:2208.10442 [cs]。Yang, Y., Huang, Z., and Wipf, D. 优化视角下的Transformer，2022年5月。URL http://arxiv.org/abs/2205.13891。arXiv:2205.13891 [cs]。

Yuille, A. L. 和 Rangarajan, A. 凹凸过程（CCCP）。载于神经信息处理系统进展，第 14 卷。麻省理工学院出版社，2001。网址：
https://proceedings.neurips.cc/paper/2001/hash/a012869311d64a44b5a0d567cd20de04-Abstract.htm

此图“example_figures.png”提供“png”格式，下载地址：

http://arxiv.org/ps/2304.04556v1

此图“example_graphics.PNG”以“PNG”格式提供，来源：

http://arxiv.org/ps/2304.04556v1