上下文学习的自旋玻璃模型

李宇豪

日期：2024 年 9 月 16 日

1 引言

Transformer 是一种基于自注意力机制的深度学习模型架构，最初由 Google 的研究人员于 2017 年提出。其核心思想是通过自注意力机制来处理序列数据，使得模型在处理输入序列的每个位置时能够直接关注序列中的所有其他位置，从而捕捉长程依赖关系。与传统神经网络相比，Transformer 减少序列长度的依赖，显著提升了训练速度，并且能很好地处理长文本和复杂序列，这种机制推动了 GPT 等大型预训练语言模型的发展。

上下文学习（ICL）是在大语言模型 GPT-3 中提出的一种新的学习方法，其关键思想是从类比中学习：通过一些示例来形成一个演示上下文，将查询的问题和上下文演示连接在一起形成带有提示的输入，然后将其输入到模型中进行预测。与监督学习不同，上下文学习不需要参数更新，直接对预先训练好的语言模型进行预测，这使得大语言模型发展出诸如问答、代码生成等广泛的应用。

目前，对于上下文学习的理论研究主要集中在通过梯度下降动力学、表示能力、贝叶斯推理以及任务多样性等方面，但是几乎没有与物理模型建立联系。在这里，我们将 Transformer 的学习视为一个统计推断问题，然后将推断问题改写为自旋玻璃模型，将 Transformer 的参数映射到连续值自旋，用输入数据充当淬火无序，然后使用自旋玻璃理论的平均场方法来求解这个问题。

2 线性注意力机制

对于特征维度为 D，序列长度为 N 的输入矩阵 $\mathbf{X} \in \mathbb{R}^{D\times N}$，其自注意力矩阵可以表示为

$$\text{Attn} = \mathbf{V} \cdot \text{Softmax}\left(\frac{\mathbf{Q}^\top \mathbf{K}}{\sqrt{D}}\right) \tag{1}$$

其中 $\mathbf{Q} = \mathbf{W}_\text{Q}\mathbf{X}$、$\mathbf{K} = \mathbf{W}_\text{K}\mathbf{X}$ 和 $\mathbf{V} = \mathbf{W}_\text{V}\mathbf{X}$ 分别表示查询（query）、键（key）和值（value）矩阵，$\mathbf{W}_\text{Q}, \mathbf{W}_\text{K}, \mathbf{W}_\text{V} \in \mathbb{R}^{D\times D}$ 是三个可训练的权重矩阵。

如果将自注意力机制中的 $\mathbf{W}_\text{Q}$ 和 $\mathbf{W}_\text{K}$ 矩阵合并表示为 $\mathbf{W} \equiv \mathbf{W}_\text{Q}^\top \mathbf{W}_\text{K} \in \mathbb{R}^{D\times D}$，将 $\mathbf{W}_\text{V}$ 取为单位矩阵 $\mathbf{W}_\text{V} = \mathbf{I}$，并且去掉非线性函数 Softmax$(\cdot)$，我们可以得到一个更简单的线性模型

$$\mathbf{A} = \frac{1}{D\sqrt{N}}\mathbf{X}\mathbf{X}^\top \mathbf{W}\mathbf{X} \tag{2}$$

称为线性注意力机制。

对于高维线性回归任务，我们从高斯分布中采样特征向量和权重向量，即 $\boldsymbol{x}, \boldsymbol{w} \sim \mathcal{N}(0, \mathbf{I}_D)$，然后通过内积生成标签 $y = \boldsymbol{w}^\top \boldsymbol{x}$。我们按照上下文学习的形式，使用 N 个样本 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ 作为提示，使用一个额外的样本 $(\tilde{\boldsymbol{x}}, \tilde{y})$ 作为查询，构成如下形式的输入矩阵：

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_N & \tilde{\boldsymbol{x}} \\ y_1 & y_2 & \cdots & y_N & 0 \end{bmatrix} \in \mathbb{R}^{(D+1)\times(N+1)} \tag{3}$$

其中查询样本的标签被置为零（称为掩码）。

我们将模型输出的注意力矩阵 $\mathbf{A}$ 中对应掩码位置的元素值作为模型对查询样本的预测值，即 $\hat{y} = \mathbf{A}_{D+1,N+1}$，并且使用均方误差作为损失函数，单个样本的损失写为 $\ell = \frac{1}{2}(\hat{y} - \tilde{y})^2$。我们在训练中共使用

P 个样本，并且使用 ℓ_2 正则化，因此整个训练集上的损失函数写为

$$\mathcal{L} = \frac{1}{2P}\sum_{\mu}^{P}(\tilde{y}^{\mu} - \hat{y}^{\mu})^2 + \frac{\lambda}{2}\|\mathbf{W}\|^2 \tag{4}$$

3 映射到自旋玻璃模型

我们定义 $\mathbf{C} \equiv \mathbf{X}\mathbf{X}^{\top}$，并把模型的预测值改写为如下形式

$$\hat{y} = \frac{1}{D\sqrt{N}}\sum_{m,n}\mathbf{C}_{D+1,m}\mathbf{W}_{m,n}\mathbf{X}_{n,N+1} \tag{5}$$

然后我们定义一个将矩阵扁平化为向量的指标映射 $\Gamma:(m,n)\to i$（比如，一种可行的方案是令 $i=(D+1)(m-1)+n$），因而我们可以分别将权重和数据扁平化为 $\sigma_i = \Gamma\mathbf{W}_{m,n}$ 以及 $s_i = \frac{1}{D\sqrt{N}}\Gamma\mathbf{C}_{D+1,m}\mathbf{X}_{n,N+1}$。这样，我们就可以将模型的预测值重新表示为 $\hat{y}=\sum_i s_i\sigma_i$，根据式（1），损失函数也就可以写为

$$\mathcal{L} = \frac{1}{2P}\sum_{\mu}\sum_{i,j}s_i^{\mu}s_j^{\mu}\sigma_i\sigma_j - \frac{1}{P}\sum_{\mu}y^{\mu}\sum_i s_i^{\mu}\sigma_i + \frac{\lambda}{2}\sum_i\sigma_i^2 \tag{6}$$

其中略去了常数项 $\frac{1}{2P}\sum_{\mu}(y^{\mu})^2$。定义 $J_{ij} = -\frac{1}{P}\sum_{\mu}s_i^{\mu}s_j^{\mu}$，$h_i = \frac{1}{P}\sum_{\mu}y^{\mu}s_i^{\mu}$ 以及 $\lambda_i = \lambda - \frac{1}{P}\sum_{\mu}J_{ii}^{\mu}$，我们就可以把损失函数改写为如下形式

$$\mathcal{H}(\boldsymbol{\sigma}) = -\sum_{i<j}J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i + \frac{1}{2}\sum_i\lambda_i\sigma_i^2 \tag{7}$$

这是一个具有连续值的自旋玻璃模型！

在这个映射中，我们将数据的结构和无序性映射到自旋之间的相互作用 $\mathbf{J}$ 以及自旋感受到的外场 $\boldsymbol{h}$，将模型的权重映射到自旋的值，模型的训练过程表现为自旋值的变化，而当自旋系统达到平衡态时，就意味着模型的训练收敛了。

得到 $\mathbf{J}$ 和 $\boldsymbol{h}$ 中元素的解析分布是十分困难的，但是我们可以从数值上直观的观察它们的统计性质。我们首先关注 s_i 未被 Γ 映射展平的原始形式，即 $S_{mn} = \frac{1}{D\sqrt{N}}\mathbf{C}_{D+1,m}\mathbf{X}_{n,N+1}$，其中省略索引 μ，表示对所有输入矩阵 $\{\mathbf{X}^{\mu}\}$ 的平均值。矩阵 $\mathbf{S}$ 的结构如图 1(a) 所示，最后一列是全零向量，另外两个不同的分块分别被标记为 $\mathcal{A}$（$m<D+1, n\neq D+1$）和 $\mathcal{B}$（$m=D+1, n\neq D+1$），它们各自的元素的分布展

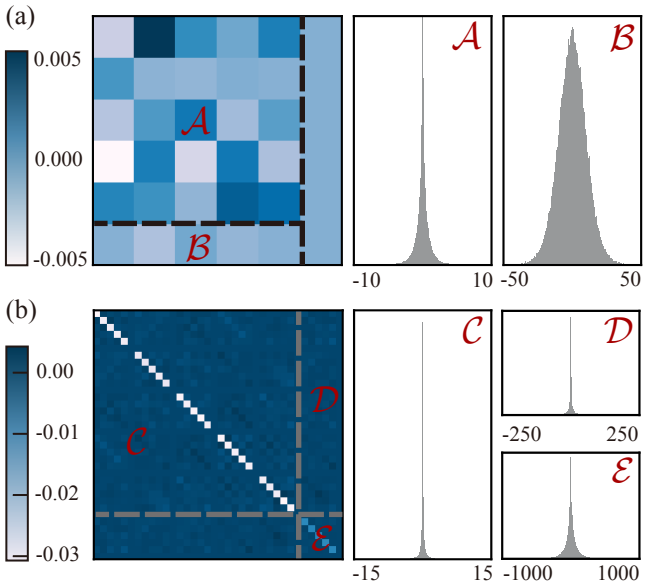

图 1: 自旋玻璃模型的相互作用矩阵 $\mathbf{J}$ 和外场 $\boldsymbol{h}$ 的统计性质。（a）矩阵 $\mathbf{S}$ 的分块结构和对应的元素分布，由于 h_i 和 s_i 只相差一个常数，这完全可以表示外场 $\boldsymbol{h}$ 的统计性质；（b）自旋相互作用矩阵 $\mathbf{J}$ 的分块结构和对应的元素分布。

示在右侧。类似地，$\mathbf{J}$ 矩阵的结构如图 1(b) 所示，其非对角元素可以分为 $\mathcal{C}$、$\mathcal{D}$、$\mathcal{E}$ 三个不同分块。

$\mathbf{J}$ 和 $\boldsymbol{h}$ 的元素分布几乎都有长尾的特征，并且在某些情况中十分严重，这是因为它们可以表示为一系列高斯随机变量的求和以及乘积的复杂组合。而这种长尾的分布事实上也影响了我们后文中使用的 AMP 算法的收敛范围——在低数据量时需要很大的正则化系数才能够使得 AMP 算法收敛。

4 统计力学分析

自旋玻璃理论提供了一种强大的平均场方法——空腔方法来求解这类系统。对于哈密顿量 (7)，可以写出其玻尔兹曼分布 $P(\boldsymbol{\sigma}) = e^{-\beta\mathcal{H}(\boldsymbol{\sigma})}/Z$，其中 Z 是配分函数。在两体相互作用中，我们可以推导出如下的自洽迭代方程

$$\eta_{i\to j}(\sigma_i) = \frac{1}{z_{i\to j}} e^{\beta h_i\sigma_i - \frac{1}{2}\beta\lambda_i\sigma_i^2} \prod_{k\neq i,j}\left[\int \mathrm{d}\sigma_k\, \eta_{k\to i}(\sigma_k)\, e^{\beta J_{ik}\sigma_i\sigma_k}\right] \tag{8}$$

其中 $z_{i\to j}$ 是归一化常数，$\eta_{i\to j}$ 是节点 i 与 j 之间没有相互作用时节点 i 的空腔概率。当所有的空腔概率迭代到它们的不动点后，就可以计算每个节点的边缘概率密度：

$$\eta_i(\sigma_i) = \frac{1}{z_i} \mathrm{e}^{\beta h_i\sigma_i - \frac{1}{2}\beta\lambda_i\sigma_i^2} \prod_{j\neq i}\int \mathrm{d}\sigma_j\, \eta_{j\to i}(\sigma_j)\, \mathrm{e}^{\beta J_{ij}\sigma_i\sigma_j} \tag{9}$$

考虑到自旋之间的相互作用时密集但是微弱的，在大 D 极限下，通过傅立叶变换和泰勒展开，我们可以进一步简化平均场方程 (8)，得到所谓的近似消息传递算法（AMP）。它可以通过 $\eta_i(\sigma_i) \sim \mathcal{N}(m_i, v_i)$ 来计算边缘概率密度 $\eta_i(\sigma_i)$，其中均值 $\{m_i\}$ 和方差 $\{v_i\}$ 分别由以下两条方程的不动点给出：

$$m_i = \frac{\beta h_i + \beta\sum_{j\neq i} J_{ij} m_j}{\beta\lambda_i - \beta^2\sum_{j\neq i} J_{ij}^2 v_j} \tag{10a}$$

$$v_i = \frac{1}{\beta\lambda_i - \beta^2\sum_{j\neq i} J_{ij}^2 v_j} \tag{10b}$$

5 结果与讨论

线性回归任务的结果是可以预期的。我们将模型的权重矩阵按照如下方式分块

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{bmatrix} \tag{11}$$

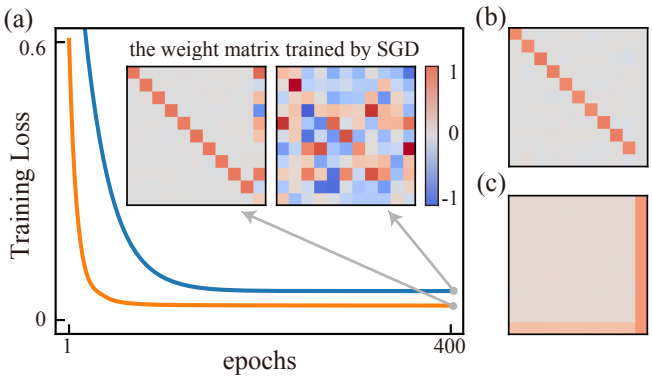

图 2: 优化后的权重矩阵。(a) 使用 SGD 训练的损失曲线和得到的权重矩阵，蓝色线表示 $P = 10$ 的情况，训练得到的权重矩阵仍然是无序的，橙色线表示 $P = 1000$ 的情况，训练得到的矩阵收敛到最优解；(b) AMP 算法得到的权重矩阵的均值，收敛到了同样的最优解；(c) AMP 算法得到的权重矩阵的方差。

其中，$\mathbf{W}_{11} \in \mathbb{R}^{D\times D}$，$\mathbf{W}_{12} \in \mathbb{R}^{D\times 1}$，$\mathbf{W}_{21} \in \mathbb{R}^{1\times D}$，$\mathbf{W}_{22} \in \mathbb{R}$。那么，很容易将模型对于查询样本 $\tilde{\boldsymbol{x}}$ 的预测值写为

$$\hat{y} = \boldsymbol{w}^\top \left(\mathbf{W}_{11} + \boldsymbol{w}\mathbf{W}_{21}\right)\tilde{\boldsymbol{x}} \tag{12}$$

对比该样本的真实标签 $\tilde{y} = \boldsymbol{w}^\top \tilde{\boldsymbol{x}}$，可以预期，一个训练良好的模型应该满足

$$\mathbf{W}_{11} + \boldsymbol{w}\mathbf{W}_{21} = \mathbb{1}_D \tag{13}$$

方程 (13) 是一个线性方程组，当 $P > 1$ 时，方程组有唯一解 $\mathbf{W}_{11} = \mathbb{1}_D$，$\mathbf{W}_{21} = \mathbf{0}$。当然，在真实的神经网络训练过程（如 SGD）中，训练误差不会完全降为零，因此需要更大的数据量 P 使得模型收敛到最优解。

图 2(a) 中展示了 $P = 10$ 和 $P = 1000$ 时使用 SGD 训练的损失曲线以及最终的权重矩阵。可以看到，当数据量足够大时，模型的权重矩阵 $\mathbf{W}$ 收敛到了我们预期的最优解。图 2(b)、(c) 分别是 AMP 算法得到的权重矩阵的均值和方差，我们的自旋玻璃模型给出了同样的结果。由于权重矩阵 $\mathbf{W}$ 的最后一列不参与训练，并且初始化为 $\mathcal{N}(0,1)$，因此 AMP 的结果保留了均值为 0 和方差为 1 的特征。

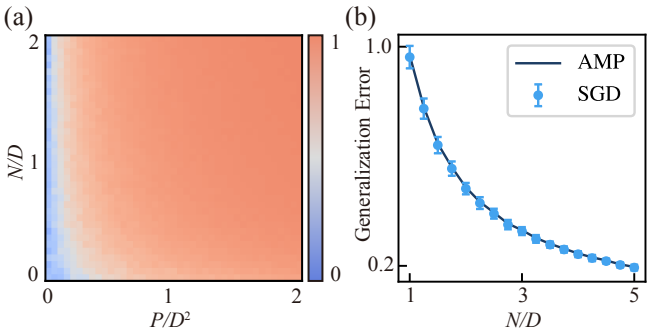

图 3: AMP 算法的结果。（a）$D = 40$，$\lambda = 10$，$\beta = 100$ 时的对比度 $\mathscr{C}$ 热图；（b）AMP 和 SGD 得到的泛化误差随提示长度 N 的变化曲线，$D = 40$，$\lambda = 0.1$，$P = 1000$，$\beta = 100$，所有的结果都是对 100 次实验取平均。

鉴于 $\mathbf{W}$ 的最优解的特点，我们可以定义一个对比度 $\mathscr{C} = (\langle m_{ii}^2\rangle - \langle m_{ij}^2\rangle)/\langle m_{ii}^2\rangle$，$i \neq j$，来描述权重矩阵与最优解之间的距离。$\mathscr{C} = 1$ 意味着矩阵收敛到了最优解，而 $\mathscr{C} = 0$ 意味着模型什么都没学到。图 3(a) 展示了对比度 $\mathscr{C}$ 随着归一化数据量 P/D^2 和归一化提示长度 N/D 的变化。这个热图表明，当任务多样性增加时，会平稳地过渡到完美泛化；同时，保持较大的任务多样性值，增加提示长度会进一步降低泛化误差

最后，我们将 AMP 算法得到的权重矩阵放入模型中，不再进行训练，而是直接计算对测试集的损失函数，作为 AMP 对泛化误差的预测值。结果如图 3(b) 所示，AMP 的结果与 SGD 的结果完全一致，这再次验证了我们的自旋玻璃模型的有效性。