Overcoming Spectral Bias via Cross-Attention

Outline

交叉注意力是什么：

交叉注意力机制是 Transformer 架构的核心组件，核心功能为建模两种不同模态 / 序列域间的语义关联与特征对齐，解决跨域任务中的语义鸿沟问题。

其遵循查询 - 键 - 值（QKV）范式，核心特征是域分离输入：查询向量（Q）源自目标序列，键（K）与值（V）向量源自源序列。通过计算 Q 与 K 的缩放点积相似度得到注意力权重（量化跨域位置关联度），经 Softmax 归一化后加权聚合 V，生成融合目标导向与源域信息的特征表示。

该机制广泛应用于机器翻译、图文对齐等跨模态 / 序列任务，相较于自注意力，通过解耦 Q 与 K V 的来源，可更高效捕捉跨域关键信息，避免单一序列内部冗余关联干扰。

为什么需要融合交叉注意力和随机傅里叶特征？

在神经网络训练中，谱偏置会导致训练动态失衡，具体表现为高频分量的收敛速度可能远慢于低频分量。

为缓解这一问题，一种基于交叉注意力的网络架构被提出，该架构利用可学习缩放因子，对经过尺度变换的多尺度随机傅里叶特征库进行自适应加权。可学习缩放因子负责调整多尺度随机傅里叶特征的幅值，而交叉注意力残差结构则提供一种输入依赖型机制，用以突出贡献最大的特征尺度。

实验结果表明，相较于基于相同多尺度特征库构建的对标基线模型，所提方法能够有效加速高频分量的收敛。

构建随机傅里叶特征库

傅里叶特征库（Fourier Feature Bank）是一种用于增强神经网络高频表征能力的特征集合，核心是通过傅里叶变换相关的映射，将输入数据映射到包含多尺度频率信息的高维特征空间，为模型提供丰富的频谱信息以适配高频、振荡或不连续目标的拟合需求。

具体构建步骤如下：

- 基础频率采样
- 多尺度扩展
- 相位与幅值调制
- 特征归一化

构建随机傅里叶特征库

基础频率采样

设输入数据维度为d_{in}，即输入向量$x \in \mathbb{R}^{d_{\text{in}}}$。我们从零均值高斯分布中随机采样$M_{\text{base}}$个独立的基础频率向量，数学定义为：

$$\omega_m \sim \mathcal{N}\left(0, \sigma^{-2} \boldsymbol{I}_{\boldsymbol{d}_{\text{in}}}\right)$$

其中：

- M_{base} 表示基础频率向量的数量；
- σ 是控制频率范围的缩放参数；
- σ越小，采样频率的向量幅度越大，基础频率谱越宽；
- $\boldsymbol{I}_{\boldsymbol{d}_{\text{in}}}$ 是d_{in}维单位矩阵。保证采样频率向量在各维度上独立同分布。

构建随机傅里叶特征库

将所有采样得到的基础频率向量按行堆叠，形成基础频率矩阵:

$$\Omega_{\text{base}} \in \mathbb{R}^{M_{\text{base}} \times d_{\text{in}}}$$

该矩阵包含了特征库的核心频谱“原子”，后续所有多尺度频率均基于此矩阵通过缩放变换生成。

多尺度扩展

首先定义一组二进制尺度因子（Dyadic Scales）:

$$k = 0, 1, 2, ..., K-1$$

K 为尺度总数。每个尺度因子对应一个缩放系数：2^k

构建随机傅里叶特征库

对于每个基础频率向量ω_m和每个尺度因子 k，通过以下公式生成多尺度频率向量：

$$\tilde{\omega}_{m,k} = 2^k \cdot \omega_m$$

使用指数缩放确保是因为自然信号的频谱能量往往随频率升高呈幂律衰减，指数缩放可通过少量尺度层级，快速覆盖从低频到高频的宽范围频谱。将所有多尺度频率向量按行堆叠，形成多尺度频率矩阵：

$$\Omega \in \mathbb{R}^{M \times d_{\text{in}}}$$

其中$M = M_{\text{base}} \times K$，表示多尺度频率向量的总数量。

构建随机傅里叶特征库

相位与幅值调制

为了增强特征的表达能力，我们对每个多尺度频率向量引入随机相位偏移和幅值调制：

- 随机相位采样（Random Phase Sampling）

通过随机相位偏移，可以使不同频率分量的余弦映射结果相互独立，避免特征间的线性相关；

我们为每个$\tilde{\omega}_{m,k}$分配一个随机相位偏移$b_{m,k}$,服从均匀分布：

$$b_{m,k} \sim \text{Uniform}(0, 2\pi)$$

构建随机傅里叶特征库

- 依赖频率的振幅包络(frequency-dependent amplitude envelope)

我们为每个多尺度频率向量引入幅值衰减因子，控制高频分量的贡献强度，避免高频噪声干扰，数学定义为：

$$a_{m,k} = \exp\left(-\beta \times \|\tilde{\omega}_{m,k}\|_2\right), \quad \beta > 0$$

其中β是学习的非负标量，用于自适应调整幅值衰减速率。β越大，高频分量的幅值衰减越快。

构建随机傅里叶特征库

特征归一化

最后我们生成最终的傅里叶特征向量：

$$\phi(x) = \sqrt{\frac{2}{M}} \left[\cos\left(\tilde{\omega}_{m,k}^T x + b_{m,k} \right) \right]_{m,k}$$

$$\phi(x) \in \mathbb{R}^M$$

其中$\sqrt{\frac{2}{M}}$是归一化系数，确保随着特征数量 M 的增加不会导致特征幅值过大。该归一化系数保证了随机傅里叶特征的期望范数为 1，即：

$$\mathbb{E}[\![|\phi(x)\|_2^2] = 1$$

交叉注意力残差块

重塑分组(grouping-by-reshape) $\phi(x)$

为了减小交叉注意力机制的计算量，我们把$\phi(x)$中的傅里叶基分成多个 token,选择 token 宽度为d_q使得M可被d_q整除，RFF 特征可重塑为：

$$H(x) \in \mathbb{R}^{N_{\text{tok}} \times d_q}, \quad \text{where} \quad N_{\text{tok}} = \frac{M}{d_q}$$

交叉注意力残差块

交叉注意力迭代

加深神经网络第l层的权重为W^l,偏置为b^l，$(l = 0, 1, ..., L)$激活函数为σ. 定义初始查询向量为:

$$Q^{(0)}(x) = \sigma\left(W^{(0)}\phi(x) + b^{(0)}\right) \in \mathbb{R}^{d_q}$$

注意这里我们必须使$W^{(0)}$的形状为$\left(d_q, M\right)$，以确保矩阵乘法的维度匹配。并且后续层的$W^{(l)} \in R^{d_q \times d_q}$.

我们在每一层定义独立可训练的投影矩阵$W_Q^{(l)}, W_K^{(l)}, W_V^{(l)} \in \mathbb{R}^{d_q \times d_q}$,然后定义查询、键、值向量的计算公式:

$$Q_l(x) = Q^{(l)}(x)W_Q^{(l)}, \quad K_l(x) = H(x)W_K^{(l)}, \quad V_l(x) = H(x)W_V^{(l)}$$

交叉注意力残差块

其中$Q_l \in \mathbb{R}^{d_q}, K_l \in \mathbb{R}^{N_{\text{tok}} \times d_q}, V_l \in \mathbb{R}^{N_{\text{tok}} \times d_q}$.

然后定义 Cross-Attention 计算公式:

$$\mathrm{CA}\big(Q^{(l)}, H(x)\big) = \mathrm{softmax}\left(\frac{Q_l K_l^T}{\sqrt{d_q}}\right) V_l$$

这一公式返回一个$\mathbb{R}^{d_q}$向量，它的每一个分量的意义是，当前层对于对应 token 的需求，而每一个 token 都对应了一组傅里叶特征。这样Q^l在每一层都能根据输入x动态调整自己对不同傅里叶特征的关注度，定义调整过的查询向量为:

$$\tilde{Q}^{(l)}(x) = Q^{(l)}(x) + \mathrm{CA}\big(Q^{(l)}(x), H(x)\big), \quad l = 0, 1, ..., L-1$$

交叉注意力残差块

然后我们将调整过的查询向量传递到下一层:

$$Q^{(l+1)}(x) = \sigma\left(W^{(l+1)}\tilde{Q}^{(l)}(x) + b^{(l+1)}\right), \quad l = 0, 1, ..., L-1$$

最终网络输出为: $Q^{(L)}(x)$.

可以看出，我们在前向传播过程中，在每一层补充了一定的原始傅里叶特征信息，这些信息是经过输入依赖的交叉注意力机制加权筛选过的，从而使得网络能够动态地利用最有用的傅里叶特征来加速高频分量的学习。

自适应频率增强

尽管交叉注意力机制能够实现对随机傅里叶特征库的输入自适应加权，但纯随机频率对于主导模态稀疏且具有任务特异性的目标而言，可能难以有效拟合。

因此我们提出一种自适应频率增强（Adaptive Frequency Enhancing, AFE）策略：利用离散傅里叶变换（DFT）从初步近似解中提取后验频率，实现对 token 库的频谱丰富化。

设$u^{(0)}$表示在初始多尺度特征库下训练得到的初步近似解。我们在Ω的均匀网格上计算$u^{(0)}$的 DFT，得到傅里叶系数$\hat{u}_k^{(0)}$,定义集合B是k的索引集。我们定义

$$\zeta = \max_{k \in B} |\hat{u}_k^{(0)}|$$

自适应频率增强

ζ表示初步解的最大的频率幅值。我们希望提取那些幅值大于阈值$\lambda\zeta$的频率，作为增强频率，其中$\lambda \in (0,1)$是一个超参数。 定义

$$\mathcal{K}_{\text{post}} = \left\{ k \in B : |\hat{u}_k^{(0)}| > \lambda\zeta \right\}$$

我们原来的多尺度频率 token 库为

$$H_{\text{base}}(x) \in \mathbb{R}^{N_{\text{base}} \times d_q}$$

现在我们把大于阈值的频率都转化为角频率:

$$\omega_k^{\text{post}} = \frac{2\pi}{L} k, \quad k \in \mathcal{K}_{\text{post}}$$

其中 L 是输入域的长度。

自适应频率增强

定义$M_{\text{post}} = |\mathcal{K}_{\text{post}}|$,然后我们构造后验的傅里叶特征:

$$\phi_{\text{post}}(x) = \sqrt{\frac{2}{M_{\text{post}}}} \left[\cos\left(\omega_k^{\text{post}} x + b_k^{\text{post}}\right)\right]_{k \in \mathcal{K}_{\text{post}}}, \quad \phi(x) \in \mathbb{R}^{M_{\text{post}}}$$

其中$b_k^{\text{post}} \sim \text{Uniform}(0, 2\pi)$. 然后我们将后验特征重塑为 token 形式:

$$H_{\text{post}}(x) \in \mathbb{R}^{N_{\text{post}} \times d_q}, \quad \text{where} \quad N_{\text{post}} = \frac{M_{\text{post}}}{d_q}$$

直接竖直拼接原始 token 库和后验 token 库，得到增强后的多尺度傅里叶特征 token 库:

自适应频率增强

$$H_{\text{aug}}(x) = \left[H_{\text{base}}(x); H_{\text{post}}(x)\right] \in \mathbb{R}^{N_{\text{aug}} \times d_q}$$

$W_Q^{(l)}, W_K^{(l)}, W_V^{(l)} \in \mathbb{R}^{d_q \times d_q}$仍然保持不变,$H(x)$变为$H_{\text{aug}}(x)$,相应的三个投影向量的形状也发送改变

$$Q_l \in \mathbb{R}^{d_q}, \quad K_l \in \mathbb{R}^{N_{\text{aug}} \times d_q}, \quad V_l \in \mathbb{R}^{N_{\text{aug}} \times d_q}$$

要平滑地融合后验频率，我们在每个交叉注意力块的 logit 层引入了加性注意力掩码：

$$\mathrm{A}^{(l)} = \frac{Q_l K_l^T}{\sqrt{d_q}} + \mathcal{M}^{(l)}, \quad \mathcal{M} = [0; \eta_l \mathbf{1}]$$

自适应频率增强

$\eta_l < 0$并且随着训练的进行逐渐增大至 0，这是因为我们定义

$$\mathrm{CA}\big(Q^{(l)}, H_{\mathrm{aug}}(x)\big) = \mathrm{softmax}\big(A^{(l)}\big)V_l$$

而当η_l绝对值较大时，经过 softmax 后，后验频率对应的 token 权重接近 0，从而实现了平滑引入后验频率的效果。

该流程可重复执行，直至后验索引集$\mathcal{K}_{\mathrm{post}}$趋于稳定