arXiv:2302.01925v2 [cs.LG] 3 Apr 2024

为 Transformer 学习线性相对位置编码的傅里叶变换

Krzysztof Marcin Choromanski[*1,2] **Shanda Li**[*3] **Valerii Likhosherstov**[4] **Avinava Dubey**[1] **Shengjie Luo**[5] **Di He**[5] **Yiming Yang**[3] **Tamas Sarlos**[1] **Thomas Weingarten**[1] **Adrian Weller**[6,7]GoogleResearch [2]哥伦比亚大学 [3]卡内基梅隆大学[4] Waymo [5]北京大学 [6]剑桥大学 [7]艾伦·图灵研究所{kchoro,avinavadubey}@google.com shandal@cs.cmu.edu

*这两位作者贡献相同。作者署名按字母顺序排列。

摘要

我们提出了一类新的线性 Transformer，称为 FourierLearner-Transformer（FLT），它们融合了多种相对位置编码机制。这些机制包括应用于序列数据的常规相对位置编码技术，以及适用于嵌入高维欧几里得空间的几何数据的新型相对位置编码。FLT 通过学习相对位置编码的谱表示，隐式构建最优的相对位置编码机制。不同于其他将高效低秩线性注意力与相对位置编码结合的架构，FLT 在内存使用方面仍然实用，而且无需对相对位置编码掩码的结构作出额外假设。此外，FLT 支持应用某些结构归纳偏置技术来指定掩码策略，例如，它们提供了一种学习本文提出的所谓局部相对位置编码的方法，并且在语言建模方面，相较于其他几种线性 Transformer，准确率有所提升。我们还在其他数据模态和任务上对 FLT 进行了全面测试，包括图像分类、三维分子建模和可学习优化器。据我们所知，在三维分子数据领域，FLT 是首个提供线性注意力并纳入相对位置编码掩码的 Transformer 架构。

27th人工智能与统计国际会议（AISTATS）2024 论文集，西班牙瓦伦西亚。PMLR：第 238 卷。版权所有 © 2024，归作者所有。

1 引言

Transformer 彻底改变了机器学习的格局，为人们处理自然语言处理（NLP）中的复杂任务 [Devlin 等人, 2019], 计算机视觉（CV） [Dosovitskiy 等人, 2021], 分子建模 [Jumper 等人, 2021],以及其他领域的方式带来了范式转变。

Transformer 中最大的计算瓶颈，同时也是其成功的关键，便是注意力模块。注意力模块在输入序列中不同词元之间传递信号，其时间复杂度和空间复杂度相对于输入长度均为平方级 L，这限制了其扩展至长序列的能力。因此，设计高效的注意力模块一直是活跃的研究领域。近年来，随着 Transformer 模型的规模从“仅有”**117M** 参数的 GPT-1 架构增长到拥有**175B** 参数的 GPT-3，“高效”Transformer 的研究变得愈发重要；这一增长 $\mathbf{1000}\times$ 仅在两年内便已实现 [Brown 等人, 2020]。

一类高效 Transformer 基于稀疏注意力 [Li 等人, 2019, Vaswani 等人, 2021a, Zaheer 等人, 2020, Roy 等人, 2021, Vyas 等人, 2020, Kitaev 等人, 2020, Sun 等人, 2022]。这些方法并不旨在近似常规注意力，而是提出更简单、更易处理的注意力机制，有时还会施加额外约束（例如，相同的查询和键 [Kitaev 等人, 2020]）。另一类广受欢迎的高效 Transformer 基于核化注意力 [Choromanski 等, 2021, Tsai 等人, 2019, Katharopoulos 等人, 2020]。其核心思想是寻找注意力矩阵的近似低秩分解，并利用矩阵乘法的结合律来提升注意力机制的空间和时间复杂度。Performer [Choromanski 等, 2021] 是

这一模型类别的成功范例。与前面讨论的方法不同，Performer 的近似注意力矩阵（它从不被显式构造，而是以隐式方式使用）是原始注意力矩阵的无偏估计，通过 Softmax 核对词元之间的相似度进行编码。Performers 已被集成到许多 Transformer 架构中，以提供线性空间和时间复杂度 [Yuan 等，2021，Horn 等，2021，Tay 等，2021，Xiao 等人，2022]。

遗憾的是，Performers 的简洁性是有代价的。众所周知，对于 Performers 而言，融入结构归纳先验（在常规注意力架构中，通常通过各种加性相对掩码机制来实现）并非易事。我们将这些方法称为相对位置编码（相对位置编码） [Shaw 等，2018、Raffel 等，2020、Li 等，2021、Luo 等人、2022]。相对位置编码对于提升 Transformer 在语言长程建模中的表现至关重要 [Dai 等人、2019],语音 [Liutkus 等人、2021], 视觉 [Wu 等人、2021], 以及基因组数据 [Žiga Avsec 等人、2021]。然而，乍看之下，Performers 与通用相对位置编码技术并不兼容，因为这些技术似乎要求显式构造注意力矩阵，以便应用相对位置编码掩码；而 Performers 为了提升计算效率，恰恰避免了这一步。人们为使 Performers 与相对位置编码兼容付出了大量努力（详见第2节），但迄今为止，这些尝试都未能同时满足以下两项要求：(a) 获得切实的计算效率提升；(b) 对具有非平凡拓扑结构的输入纳入通用相对位置编码方法。

本文提出了一类名为 FourierLearner-Transformer（FLT）的新型线性 Transformer，它们融入了多种相对位置编码机制。这些机制包括应用于序列数据的常规相对位置编码技术，以及适用于嵌入高维欧几里得空间的几何数据（例如分子结构）的新型相对位置编码。FLT 通过学习谱表示，隐式构建最优的相对位置编码机制，并具有可证明的一致收敛保证。与其他将高效低秩线性注意力与相对位置编码相结合的架构不同，FLT 在内存使用量方面仍然实用，且无需对相对位置编码掩码的结构做出额外假设。此外，FLT 支持应用某些结构归纳偏置技术来指定掩码策略。例如，它们提供了一种学习本文所称的局部相对位置编码的方法；与其他几种用于语言建模的线性 Transformer 相比，这种方法能提升准确率。我们还全面测试了 FLT 在其他数据模态和任务上的表现，例如图像分类和分子建模。

据我们所知，对于三维分子数据，FLT 是首个同时提供线性注意力并融入相对位置编码掩码的 Transformer 架构，这拓展了相对位置编码增强的线性注意力的应用范围。

总而言之，我们的主要贡献如下：

- 我们提出了相对位置编码增强的线性注意力方法 FourierLearner-Transformers（FLTs）。FLTs 不仅适用于序列数据（例如文本），也适用于嵌入高维欧几里得空间的几何数据（例如三维分子数据），显著拓宽了相对位置编码增强的线性注意力的适用范围。

- 我们对FLTs进行了详尽的理论分析，包括其近似的一致收敛性和样本复杂度界（第 4.1节）。我们讨论了几种具体实现，尤其是FLTs采用所谓的高斯混合相对位置编码、平移不变核相对位置编码和局部相对位置编码的实现（第 4.3节）。

- 我们对FLTs 在语言建模（第 5.1节）、图像分类（第 5.2节）和分子性质预测（第 5.3节）任务上的表现进行了广泛评估。实验表明，FLTs 能够轻松应用于广泛的数据模态，并展现出优异的性能和效率。

2 相关工作

采用相对位置编码的核化注意力。 最早尝试解决将核化注意力 Transformer 与相对位置编码结合这一问题的工作之一是 [Liutkus 等人，2021], 他们提出了两种变体，分别是sineSPE 和 convSPE。这两种变体都将相对位置编码掩码建模为具有托普利茨掩码结构的平稳位置核。尽管它们的复杂度与序列长度 L呈线性关系，但还会额外依赖正弦分量的数量 T（对于 sineSPE）/卷积滤波器长度 P（对于 convSPE）。在实践中，由于计算预算有限， T 或 P 必须足够小。此外，它们将相对位置编码掩码限制为有效的核矩阵，而我们的FLT 则不需要此类假设。sineSPE和 convSPE 在语言建模（第 5.1节）上的表现都显著逊于FLT。此外，它们无法用于更通用的相对位置编码，例如词元嵌入在更高维欧几里得空间中的情况，如三维分子数据的相对位置编码。

最近， [Luo 等，2021，Choromanski 等，2022a]证明，相对位置编码机制可以与 Performers 结合，实现 $O(L\log(L)$）时间复杂度。该方法基于一个巧妙的观察：只要指数化的相对位置编码掩码支持快速矩阵-向量

乘法，序列数据的相对位置编码就满足这一条件，因为相应的掩码具有托普利茨结构。然而，该方法的空间复杂度较高，实际应用中内存消耗也很大（第5.1节）。此外，该方法严重依赖序列数据的结构，因此不适用于相对位置编码掩码不具有托普利茨结构的三维分子数据。

随机傅里叶特征（RFFs）。关于随机傅里叶特征的研究文献浩如烟海[Rahimi 和 Recht，2007， Avron 等人，2017，Szabó 和 Sriperumbudur，2019， Li 和 Li，2021，Choromanski 等人，2022b， Likhosherstov 等人，2022，Chowdhury 等人， 2022]。然而，关于可学习的随机傅里叶特征变体的研究 [Sinha 和 Duchi，2016] 相对较新。此外，先前的研究大多局限于在正定（PD）核的背景下应用随机傅里叶特征，而我们的工作突破了这一限制，因为相对位置编码不必由正定核定义。还有一些论文探讨了开发基于 Transformer 的模型，这类模型的注意力机制在频谱域中运行[Tamkin 等人，2020， Moreno-Pino 等人， 2023]，但它们并未研究高效的相对位置编码建模。

长序列建模。 将深度学习模型应用于长序列是一个活跃的研究方向。我们研究用于长序列建模的高效 Transformer 模型。虽然我们的重点是 Transformer 领域，但值得注意的是， [Geng 等人，2021、Bello，2021、Gu 等人， 2022]。除了效率之外，[O'Connor 和 Andreas，2021、Liu 等人，2024] 探究长序列语言模型对上下文的运用；[Press 等人，2022、Ruoss 等，2023、Li 等人， 2024]设计具备长度泛化能力的序列模型（即在短序列上训练，并泛化至长序列）； [Yun 等，2020、Yang 等人，2024]研究这些模型的理论能力。值得注意的是，现有工作大多聚焦于单一数据模态，而FLT则在广泛的模态范围内进行评估。

3 预备知识

Transformer 中的通用相对位置编码机制。考虑一个输入序列 $\mathbf{X} \in \mathbb{R}^{L \times d_{\text{in}}}$，其中 L 和 d_{in}分别表示词元的数量和嵌入维度。Transformer 中的自注意力模块将输入线性投影为三个矩阵 $\mathbf{Q}$，$\mathbf{K}$，$\mathbf{V} \in \mathbb{R}^{L \times d}$，分别称为查询、键和值。我们还为所有词元关联位置特征 $\mathbf{r}_1, ..., \mathbf{r}_L \in \mathbb{R}^{\ell}$，用于定义下文的相对位置编码机制：

定义 3.1（通用相对位置编码用于注意力）。 通用相对位置编码 增强型注意力形式如下，其中 $\mathbf{N} = [f(\mathbf{r}_i - \mathbf{r}_j)]_{i,\ j \in [L]} \in \mathbb{R}^{L \times L}$ 是所谓的 相对位置编码掩码[1] 而 $f\colon\ \mathbb{R}^{\ell} \to \mathbb{R}$ 是一个（可能可学习的）函数：

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{N}) = \mathbf{D}^{-1}\mathbf{A}\mathbf{V},$$
$$\text{where } \mathbf{A} = \exp\left(\mathbf{N} + \frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}\right), \mathbf{D} = \text{diag}(\mathbf{A}\mathbf{1}_L). \quad (1)$$

这里$\exp(\cdot)$是逐元素应用的， $\mathbf{1}_L$ 是长度为 L的全一向量，而$\text{diag}(\cdot)$ 是一个以输入向量为对角线的对角矩阵。计算式（1）的时间复杂度为 $O(L^2 d)$。

讨论。定义 3.1 具有很强的通用性，因为可以灵活选择位置的表示 $\mathbf{r}_i$ 和函数 f。例如，对于文本等序列数据，序列中的位置索引可作为位置特征（$\mathbf{r}_i = i$），而相对位置编码掩码则是一个可学习的托普利茨矩阵（$f(i-j) = c_{i-j}$，其参数为 $\{c_k\}_{k=-(L-1)}^{L-1}$） [Raffel 等人， 2020]。对于三维分子结构等几何数据，可以将 $\mathbf{r}_j$ 视为词元（例如原子）的三维坐标 $\mathbf{r}_j$，并使用特定领域的 f [Shi 等人， 2022]。我们强调，这种通用表述是新颖且重要的。它促使通用性极强的 FLT 得以发展，可适用于广泛的数据和任务；与之相对，现有方法严重依赖序列数据的结构和托普利茨相对位置编码掩码（第 2 节）。

核化线性注意力。 核化注意力技术（例如 Performers）利用注意力矩阵$\mathbf{A}$ 的分解，避免显式构造$\mathbf{A}$，从而避免 L 中的二次复杂度。对于 Softmax注意力，可通过随机特征将 Softmax核 $\exp(\mathbf{x}^{\top}\mathbf{y})$ 线性化，即针对某些随机映射构造ϕ：$\mathbb{R}^d \to \mathbb{R}^m$，使得 $\exp(\mathbf{x}^{\top}\mathbf{y}) = \mathbb{E}[\phi(\mathbf{x})^{\top}\phi(\mathbf{y})]$。

定义 $\mathbf{Q}', \mathbf{K}' \in \mathbb{R}^{L \times m}$ 为行分别表示为 $\phi(\mathbf{q}_i^{\top} d^{-\frac{1}{4}})^{\top}$ 和 $\phi(\mathbf{k}_i^{\top} d^{-\frac{1}{4}})^{\top}$ 的矩阵。于是，上述 Softmax核的线性化直接得出以下不含相对位置编码掩码的注意力近似算法：

$$\widehat{\text{Att}_{\text{K}}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \widehat{\mathbf{D}}^{-1}(\mathbf{Q}'(\mathbf{K}'^{\top}\mathbf{V}))$$
$$\text{where } \widehat{\mathbf{D}} = \text{diag}(\mathbf{Q}'(\mathbf{K}'^{\top}\mathbf{1}_L)). \quad (2)$$

这里，$\widehat{\text{Att}_{\text{K}}}$ 表示近似注意力，括号表示计算顺序。与常规注意力的时间复杂度 $O(L^2 d)$ 和空间复杂度 $O(L^2 + Ld)$ 相比，该机制的时间复杂度和空间复杂度分别为 $O(Lmd)$ 和 $O(Lm + md + Ld)$。因此，对于 $m \ll L$，Performers 能显著提升计算效率。

[1] We use $[L]$ to denote $\{1, \cdots, L\}$ in this paper.

4 方法

4.1 EfficientRPE 增强注意力

公式（2）中给出的算法未纳入相对位置编码机制。本小节首先在定理4.1 中提出一种新颖的方法，用于推导定义 N 3.1中一般相对位置编码掩码的（近似）低秩分解。接下来，我们介绍 FourierLearner-Transformer（FLT），这是一种基于该分解、兼容 Performer 的相对位置编码注意力机制。

定理 4.1。给定 $f\colon \mathbb{R}^\ell \to \mathbb{R}$ 以及在定义 *3.1* 中定义的 $\mathbf{N} = [f(\mathbf{r}_i - \mathbf{r}j)] \in \mathbb{R}^{L\times L}$，用 g表示 f的傅里叶变换。假设 p 是一个支集位于 $\mathbb{R}^\ell$上的概率密度函数。对 ξ_1、$\cdots$、ξ_r 独立同分布地$\sim p$进行采样，并定义如下随机特征映射（其中 $\mathbf{i} = \sqrt{-1}$）：

$$\varphi(\mathbf{z}) = \frac{1}{\sqrt{r}}\left[\mathrm{e}^{2\pi \mathbf{i}\mathbf{z}^\top \boldsymbol{\xi}_1}\sqrt{\frac{g(\boldsymbol{\xi}_1)}{p(\boldsymbol{\xi}_1)}}, \cdots, \mathrm{e}^{2\pi \mathbf{i}\mathbf{z}^\top \boldsymbol{\xi}_r}\sqrt{\frac{g(\boldsymbol{\xi}_r)}{p(\boldsymbol{\xi}_r)}}\right]^\top;$$

$$\psi(\mathbf{z}) = \frac{1}{\sqrt{r}}\left[\mathrm{e}^{-2\pi \mathbf{i}\mathbf{z}^\top \boldsymbol{\xi}_1}\sqrt{\frac{g(\boldsymbol{\xi}_1)}{p(\boldsymbol{\xi}_1)}}, \cdots, \mathrm{e}^{-2\pi \mathbf{i}\mathbf{z}^\top \boldsymbol{\xi}_r}\sqrt{\frac{g(\boldsymbol{\xi}_r)}{p(\boldsymbol{\xi}_r)}}\right]^\top,$$

定义 $\mathbf{N}_1 = [\varphi(\mathbf{r}_1)\ \cdots、\ \varphi(\mathbf{r}_L)]^\top \in \mathbb{R}^{L\times r}$ 以及 $\mathbf{N}_2 = [\psi(\mathbf{r}_1)\ \cdots、\ \psi(\mathbf{r}_L)]^\top \in \mathbb{R}^{L\times r}$。则 $\mathbb{E}[\mathbf{N}_1\mathbf{N}_2] = \mathbf{N}$。

适用于 Performer 的相对位置编码注意力。定理 4 1 。表明 $\mathbf{N} = \widehat{\mathbf{N}}_1\mathbf{N}_2$ 是 N的低秩无偏估计量。因此，可以获得具有线性复杂度的适用于 Performer 的相对位置编码注意力机制。具体而言，令 $\widehat{\mathbf{Q}} = [\mathbf{N}_1, \mathbf{Q}d^{-\frac{1}{4}}] \in \mathbb{R}^{L\times(m+r)}$、$\widehat{\mathbf{K}} = [\mathbf{N}_2, \mathbf{K}d^{-\frac{1}{4}}] \in \mathbb{R}^{L\times(m+r)}$，其中拼接沿第二个轴进行。则

$$\widehat{\mathbf{A}} \overset{\text{def}}{=} \exp\left(\widehat{\mathbf{N}} + \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) = \exp\left(\widehat{\mathbf{Q}}\widehat{\mathbf{K}}^\top\right). \qquad (3)$$

在公式（3）中，RPE掩码注意力现已转换为常规的Softmax注意力，并可按照公式（2）所述进行“Performer化”。这一观察自然导出了一种高效的RPE增强注意力算法，其伪代码实现见算法1。该算法的时间复杂度和空间复杂度分别为 O（L（m+r）d）和 O（L（m+r）+（m+r）d+Ld）。

在算法1中，我们不再学习 f 并尝试计算其傅里叶变换 g，以对**N**进行低秩分解，而是提出直接学习 g，并将我们的方法称为**FourierLearner-Transformer**（**FLT**）。注意，FLT 能够有效地学习 f的谱表示。

我们指出，我们的表述具有足够的通用性，能够涵盖实践中使用的多种相对位置编码变体：

用于序列数据的常规相对位置编码。在这种设定下，输入序列不具有更丰富的几何结构，因此向量$\mathbf{r}j$可以视为序列中词元的索引，即$\mathbf{r}_j = j$。因此，FLT 学习一个函数 $g : \mathbb{R} \to \mathbb{C}$（第5.1、5.2节）。

三维数据的RPE。对于这种输入类型（例如三维分子数据），将$\mathbf{r}j$ 识别为原子的三维坐标是很自然的。因此，FLT 学习一个函数 g： $\mathbb{R}^3 \to \mathbb{C}$。注意，现有方法 [Liutkus 等，2021、Luo 等，2021、Choromanski 等，2022a] 均不适用，而FLT在此设置下表现良好（第5.3节）。

最后，我们指出FLT 需要指定某个分布 p 其支撑集位于 $\mathbb{R}^\ell$，以满足定理4.1中的假设。实际应用中，p 的选择需要满足以下条件：我们能够从中高效采样，并计算其概率密度函数。在实验中，我们对 p使用零均值、单位方差或可学习方差的高斯分布。

4.2 FLT 的理论分析

我们对 FLT 的相对位置编码近似进行了理论研究。具体而言，我们证明了以下定理，该定理指出，在温和的假设条件下，估计得到的相对位置编码掩码**N** 能够以高概率将真实的相对位置编码掩码**N**近似到任意精度。此外，该定理还给出了实现这种高精度近似的样本复杂度界。

定理4.2（近似的一致收敛与样本复杂度）。给定 L 个向量$\mathbf{r}_1$，...，$\mathbf{r}_L \in \mathbb{R}^\ell$，定义相对位置编码注意力掩码 $\mathbf{N} = [f(\mathbf{r}_i - \mathbf{r}j)]_{i,\ j\in[L]}$。假设 $c = \||g(\mathbf{x})|/p(\mathbf{x})\|_\infty$，其中$g$ 是 f 的傅里叶变换，p 是定义在 $\mathbb{R}^\ell$上的某个概率密度函数。

对于任意 $\varepsilon, \delta > 0$，如果随机特征的数量$r = \Theta\left(\frac{c^2}{\varepsilon^2}\log\frac{L}{\delta}\right)$，那么 FLT 的相对位置编码近似器 $\widehat{\mathbf{N}}$ 满足

$$\mathbb{P}\left(\|\mathbf{N} - \widehat{\mathbf{N}}\|_{\max} \le \varepsilon\right) > 1 - \delta,$$

其中 $\|\cdot\|_{\max}$ 表示矩阵的最大范数。

我们还证明了估计所得相对位置编码的方差界，并在补充材料中给出结果。关于理论结果的证明和详细讨论也可参见补充材料。

4.3 傅里叶变换的拓扑

在第4.1 节的分析中，我们从未依赖 f的任何结构属性。特别是，矩阵**N** 不必是有效的正定核矩阵，甚至不必具有对称性。不过，如果需要，

算法 1 傅里叶学习器变换器：线性复杂度的 RPE增强注意力 要求：输入查询、键、值 Q、K、V $\in \mathbb{R}^{L\times d}$ 以及位置 R $\in \mathbb{R}^{L\times \ell}$；用于注意力的随机特征映射$\phi$ （见第 3 节）；RPE 函数的傅里叶变换 g_θ （可能由 θ 参数化）。输出：近似的 RPE增强注意力（定义 3.1） 1：# 应用随机特征映射以近似 RPE。 2：# 定理 4.1 中定义了 φ 和 ψ ，并按列应用； g_θ 在 φ 和 ψ中调用。 3：$\mathrm{N}_1 \leftarrow \varphi(\mathrm{R}), \mathrm{N}_2 \leftarrow \psi(\mathrm{R})$ 4：# 沿第二个轴拼接。 5：$\widehat{\ }\mathrm{Q} \leftarrow [\mathrm{N}_1, \mathrm{Q}d^{-\frac{1}{4}}]$, $\widehat{\ }\mathrm{K} \leftarrow [\mathrm{N}_2, \mathrm{K}d^{-\frac{1}{4}}]$6：# 应用随机特征映射 ϕ。 ⌢ ⌢7：$\mathrm{Q}' \leftarrow \phi(\mathrm{Q}), \mathrm{K}' \leftarrow \phi(\mathrm{K})$ 8：# 核化线性注意力（第 3 节）。方括号表示计算顺序。 9：$\mathrm{B}_1 \leftarrow \mathrm{Q}'(\mathrm{K}'^{\top}\mathrm{V}), \mathrm{B}_2 \leftarrow \mathrm{Q}'(\mathrm{K}'^{\top}\mathbf{1}_L)$, $\mathrm{O} \leftarrow \mathrm{diag}(\mathrm{B}_2)^{-1}\mathrm{B}_1$. 10：返回 O

所需的归纳偏置可以通过用于训练 g 的特定参数化方案融入 FLT，正如我们在本小节中所讨论的。

高斯混合相对位置编码。 我们考虑过的 g 最通用的参数化形式之一，是所谓的高斯混合 变体：

$$g(\boldsymbol{\xi}) = \sum_{t=1}^{T} w_t \exp\left(-\frac{\|\boldsymbol{\xi}-\boldsymbol{\mu}_t\|^2}{2\sigma_t^2}\right).$$

因此，傅里叶变换 g 由 $(2+\ell)T$ 个数字参数化： w_1, ..., w_T, σ_1, ..., $\sigma_T \in \mathbb{R}$, $\boldsymbol{\mu}_1$, ..., $\boldsymbol{\mu}_T \in \mathbb{R}^\ell$。在 $T=1$ 的特殊情况下，傅里叶变换会变成一个经过归一化的高斯核，因此将 f 定义为另一个高斯核。

用于相对位置编码掩码的平移不变核。 将 FLT相对位置编码机制用于构造N 任意 平移不变核[Rahimi 和 Recht, 2007]，非常简单。根据 Bochner 定理，对于平移不变核：$\mathrm{K}: \mathbb{R}^\ell \times \mathbb{R}^\ell \to \mathbb{R}$，存在相应的概率分布 p_{K} 和某个正数常量 $C>0$，使得

$$\mathrm{K}(\mathbf{x},\mathbf{y}) = C\int_{\mathbb{R}^d} e^{\mathbf{i}(\mathbf{x}-\mathbf{y})^\top \boldsymbol{\xi}} p_{\mathrm{K}}(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}$$

因此，为了通过平移不变核 K 得到的核矩阵 $[\mathbf{K}(\mathbf{s}_i, \mathbf{s}_k)]_{i,\ k=1,\ ...,\ L}$获得相对位置编码掩码 N 的无偏近似，只需取$\mathbf{r}_j = \frac{1}{2\pi}\mathbf{s}j$, $g(\boldsymbol{\xi}) = Cp_{\mathrm{K}}(\boldsymbol{\xi})$，其中 $j=1$, ...r。即使已经选定某一类平移不变核，FLT 仍提供了一种方法，可通过学习适当参数化的 g 来学习该类核的具体形式。

局部相对位置编码。 通过相应的结构化参数化傅里叶变换 g，FLT 也能够对各种方案进行建模，其中相对位置编码机制只需在局部应用，而对于彼此距离足够远的词元，则使用常规注意力。我们将这类策略称为局部相对位置编码。序列数据和高维几何数据都可以推导出局部相对位置编码。

最基本的局部相对位置编码取$\mathbf{r}_j = j$ ，并针对注意力半径 $v>0$ 和 $C\in\mathbb{R}$，将 f 定义为[2]

$$f_{v,C}(\Delta r) = C\cdot\mathbb{I}[|\Delta r|\leq v]. \tag{4}$$

这种相对位置编码机制会以某个乘法系数增强或减弱彼此相近词元之间的常规注意力分数，其作用可能类似于局部注意力 [Vaswani 等人，2021b]。事实证明，这种 f 的傅里叶变换具有一种特别简单的形式：

$$g_{f_{v,C}}(\xi) = C\cdot\frac{\sin(2\pi v\xi)}{\pi\xi}.$$

有趣的是，式（4）中的相对位置编码可以轻松推广到更高维的局部相对位置编码。在这种情况下，我们考虑如下形式的位置编码函数 f： $\mathbb{R}^\ell \to \mathbb{R}$

$$f_{\mathbf{v},C}(\Delta\mathbf{r}) = \prod_{j=1}^{\ell} C\cdot\mathbb{I}[|\Delta r^{(j)}|\leq v_j] \quad (\Delta\mathbf{r}\in\mathbb{R}^\ell),$$

其中 $\Delta r^{(j)}$ 表示 $\Delta\mathbf{r}$ 的第 j 个分量。相应的傅里叶变换 g 可以分解为

$$g_{f_{\mathbf{v},C}}(\boldsymbol{\xi}) = C\cdot\prod_{j=1}^{\ell}\frac{\sin(2\pi v_j\xi_j)}{\pi\xi_j}.$$

这一结果还可以进一步推广。考虑以下形式的函数 g：

$$g_{k_1,\ldots,k_\ell}^{v_1,\ldots,v_\ell}(\boldsymbol{\xi}) = C\cdot\prod_{j=1}^{\ell}\frac{\sin^{k_j}(2\pi v_j\xi_j)}{\pi\xi_j}$$

[2]请注意，除了使用式（4）中的一个指示函数外，也可以对多个指示函数进行线性组合，并为其设置可学习的半径和一组系数。

g 的逆傅里叶变换可写为

$$f(\Delta \mathbf{r}) = M \cdot \prod_{j=1}^{d} f_j^{v_j}(\Delta r_j),$$

其中 M 是一个常数，且每个 $f_j^{v_j}$ 均满足：(a) 连续；(b) 对称；(c) 具有紧支撑，其长度取决于 v_j；(d) 分段为阶数 $k_j - 1$的多项式。此类函数 f 是自然的候选函数，可用于为位置嵌入在 $\mathbb{R}^\ell$ 中的词元以及任意 $\ell \geq 1$构造连续的局部相对位置编码机制。图 $\ell = 2$展示了由 FLT支持的 4 种 C.1中的局部相对位置编码变体。

上述理论结果可通过直接积分并认识到以下函数的 N 维傅里叶变换直接得到： $h(x_1, \cdots, x_N)$ 定义为 $= h_1(x_1) \cdot \cdots \cdot h_N(x_N)$ 可以表示为各个分量 h_j 的一维傅里叶变换的乘积。

备注。我们指出，上述三种参数化方案在参数使用上都很高效。在所有实验中，FLT 为相对位置编码引入了额外的 < 0.03M 个参数。请注意，额外参数的数量不会随输入序列长度增加。

5 实验

在本节中，我们在多种任务上给出实验结果，以展示 FLT 架构的有效性。我们首先研究序列文本数据上的语言建模，这是高效 RPE 增强注意力的标准设置，也便于与现有基线方法进行全面比较。接下来，我们考虑计算机视觉领域，并在多个图像分类数据集上测试 FLTs。最后，为了展示 FLTs 如何拓展 RPE 增强的高效 Transformer 的应用范围，我们在具有复杂 RPE 掩码的分子性质预测任务上进行实验，而现有的高效 RPE 增强注意力基线方法无法处理这些掩码。完整的实验设置、各项任务的超参数以及硬件细节见补充材料。

5.1 语言建模

我们在 WikiText-103 语言建模任务上开展实验，以展示所提出方法在自然语言处理应用中的有效性。大多数现有基线方法都适用于序列文本数据。因此，在本小节中，我们针对模型质量和效率，提供与基线方法的全面实证比较。

对比方法。 在本实验中，我们研究了采用两种相对位置编码变体的 FLT：高斯混合相对位置编码和局部相对位置编码。我们将模型与以下强基线方法进行比较：

表1：**语言模型困惑度分数**在 WikiText-103 验证集上的结果。最低困惑度以**粗体**标出。

模型	困惑度
线性变换 [Katharopoulos 等人， 2020]	38.4
RFA-高斯 [Peng 等人， 2021]	33.6
RFA-反余弦 [Peng 等人， 2021]	36.0
RFA-GATE-高斯 [Peng 等人， 2021]	31.3
RFA-GATE-反余弦 [Peng 等人， 2021]	32.8
Performer [Choromanski 等人， 2021]	31.1
CosFormer [Qin 等人， 2022]	30.7
Performer-sineSPE [Liutkus 等人， 2021]	38.0
Performer-convSPE [Liutkus 等人， 2021]	37.8
对数线性 Performer [Luo 等人， 2021]	30.6
FLT（高斯混合相对位置编码）（我们的方法）	30.3
FLT（局部相对位置编码）（我们的方法）	**30.1**

- 线性 *Transformer* [Katharopoulos 等人， 2020], 它使用核化低秩注意力，并以$\mathrm{elu}(\cdot) + 1$ 作为特征映射。
- 随机特征注意力 （RFA） [Peng 等人 ， 2021],它有两种变体（高斯和反余弦），并可选用门控机制。
- 常规的*Performer* [Choromanski 等人， 2021],其应用 FAVOR+机制来近似注意力矩阵。
- *CosFormer* [Qin 等人， 2022], 其设计了一种线性算子和一种基于余弦的距离重加权机制，用于近似注意力矩阵。
- *Performer-SPE* [Liutkus 等人， 2021], 其将一类特殊的相对位置编码融入低秩注意力，并有两个变体（ sineSPE 和 convSPE）。
- 对数线性 *Performer* [Luo 等人， 2021] 其将 Performers 扩展为可处理任意 Toeplitz 相对位置编码注意力掩码。

实现细节。 所有测试模型均为基于核化低秩注意力的高效Transformer，包含 6 个解码器层。有关模型配置和训练的更多细节，请参阅补充材料。我们使用验证困惑度作为评估指标；困惑度越低，性能越好。

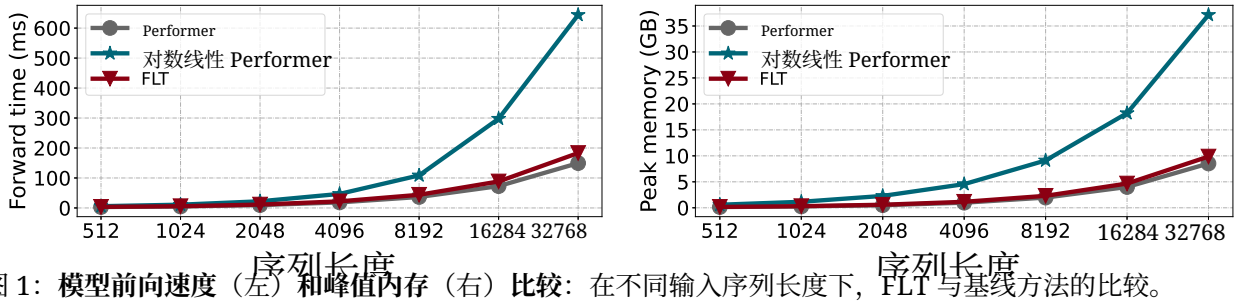


图 1：**模型前向速度**（左）**和峰值内存**（右）**比较**：在不同输入序列长度下，FLT 与基线方法的比较。

结果。 结果如表 1所示。我们提出的 FLT的两个变体均优于所有基线方法。与不使用相对位置编码的高效 Transformer相比，FLT的性能显著更强。例如，我们提出的FLT 结合局部相对位置编码时，其验证困惑度比常规 Performer 低 1.0，表明将相对位置编码融入注意力模块能够显著提升性能。

与其他采用相对位置编码的高效 Transformer 变体相比，我们的 FLT 仍然极具竞争力。例如，我们的 FLT 困惑度低于强大的对数线性 Performer 基线方法。需要注意的是，对数线性 Performer 依赖计算开销更大的快速傅里叶变换，因此在实践中效率较低。具体而言，FLT 的时间复杂度和空间复杂度分别为 $O(L(m+r)d)$ 和$O(L(m+r)+(m+r)d+Ld)$，而对数线性 Performer 的时间复杂度和空间复杂度分别为$O(Lmd\log L)$ 和 $O(Lmd)$。因此，在这项任务中，与现有采用相对位置编码的高效 Transformer 变体相比，我们的 FLT 兼具更优的质量和效率。

此外，我们进一步研究了FLT[3]的注意力矩阵。我们将一个FLT模型在 WikiText-103 语言建模任务上训练后，不同注意力头的注意力矩阵可视化，如附录C.2中的图 5 所示。可视化结果表明，一些注意力头更关注附近的词元，而另一些则呈现全局注意力模式。从定量结果来看，距离最远和最近的 10% 词元所获得的平均注意力概率分别为 0.068 和 0.279。因此，FLT 在语言中学会了局部性偏置，同时保留了捕捉全局上下文并利用远距离词元信息的优势。

计算成本比较。 如上所述，与最强的基线方法——对数线性 Performer 相比，FLT 在时间和空间复杂度方面表现出色。为展示 FLT 在实践中的效率，我们构建了一个包含 12 个

[3]FLT 在训练期间不会显式构建注意力矩阵，因此避免了二次计算复杂度。不过，我们仍然可以实例化由 FLT 近似得到的注意力矩阵

表 2：**图像分类准确率**比较。由于对数线性 Performer 的内存复杂度过高且会导致内存溢出，因此未予列出。最佳表现以**粗体**标出。

	ImageNet	Places365	Fashion-MNIST
Performer	75.1%	55.0%	91.1%
CosFormer	76.2%	55.6%	91.6%
FLT（我们的方法）	**77.4**%	**56.0**%	**92.1**%

注意力头的隐藏维度设为 768，前馈神经网络的维度设为 3072。我们将长度各异、批量大小为 8 的输入送入模型，并测量其效率。在不同输入序列长度下，我们报告 5 次运行中的平均前向传播时间和最大峰值内存消耗，结果见图1。我们将FLT与最强基线方法——对数线性 Performer——进行比较，并将常规 Performer 作为参考。显然，与常规 Performer 相比，FLT 仅带来可忽略的内存开销；在实际应用中，无论是模型前向传播时间还是峰值内存，其扩展性都远优于对数线性 Performer。因此，我们的实验表明，FLT 在序列文本数据上比基线方法准确性更高、可扩展性更强。

5.2 图像分类

我们对FLT视觉变换器（ViTs）的各种变体 [Dosovitskiy 等人， 2021] 在多个图像分类数据集上进行了全面基准测试，包括 ImageNet、Places365 和 Fashion-MNIST。有关这些数据集的详细信息，请参阅补充材料。

对比方法与实现细节。我们将 FLT 与常规 Performer 以及第 5.1 节中最具竞争力的对比方法 *CosFormer* 和 对数线性 *Performer*进行了比较。所有测试的视觉变换器均由 12 层组成，每层包含 12 个注意力头。有关模型配置和训练的更多细节请参阅补充材料。对于

表 3：**FLT 与常规 Performer 的比较**，用于 OC20 IS2RE 任务。后缀“-*k*L”表示模型由 k 层组成，例如，FLT-10L 表示一个由 10 层组成的 FLT。评估指标为能量的平均绝对误差（MAE，越低越好）和阈值内能量（EwT）的百分比（越高越好）。我们用**粗体**标出了最佳表现。

	能量平均绝对误差（eV）	阈值内能量（%）
Performer-12L	0.5454	4.90
FLT-10L（我们的模型）	0.5157	**5.44**
FLT-12L（我们的模型）	**0.5046**	5.33

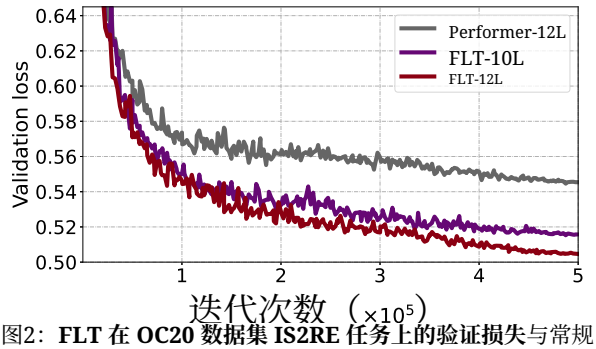


图2：**FLT 在 OC20 数据集 IS2RE 任务上的验证损失**与常规 Performer 的对比。

我们的 FLT变体采用高斯混合相对位置编码（第 4.3 节），高斯混合模式的数量T 设为 25，相对位置编码所用的随机特征数量 r 设为 64。

结果。结果如表 2 所示。对数线性 *Performer* 架构在 $m = 128$ 时会耗尽内存；而当 m 降低后（批量大小固定为 4096），才能适应分配的内存，因此未纳入比较。与另外两种基线方法相比，我们的 FLT 在全部三个数据集上都取得了最佳表现。例如，在 ImageNet 上，FLT 的**2.3%** 准确率比常规 Performer 高；甚至比表现强劲的 CosFormer 基线方法还高出**1.2%**。结果表明，FLT 在图像数据上同样表现出色。

5.3 分子性质预测

正如前文讨论所指出的，FLT 拓展了相对位置编码增强的高效 Transformer 的适用范围，并可用于嵌入高维欧几里得空间的几何数据。为验证这一主张，本小节进一步在分子性质预测任务上评估我们的 FLT 模型，以展示其处理三维输入数据和复杂（非 Toeplitz）相对位置编码掩码的能力。据我们所知，在这一应用场景中，FLT 是首个提供相对位置编码增强的可扩展注意力机制的 Transformer，其计算复杂度相对于输入词元数量呈线性增长。

我们使用公开的大规模电催化剂数据集 Open Catalyst2020（OC20），并专注于 IS2RE 任务：给定固体催化剂及其吸附分子的初始结构，预测弛豫后结构的能量[Chanussot* 等人， 2021]。

对比方法。 前面实验中考虑的现有技术不适用于这种设置。因此，我们只将 FLT 与不使用相对位置编码的标准 Performer 进行比较。

对于 FLT 模型，我们考虑使用高斯基函数来近似相对位置编码掩码。这类基函数广泛用于分子建模的神经网络中 [Gasteiger 等人，2021，Shi 等，2022，Luo 等人， 2023]。具体而言，相对位置编码掩码定义为 $\mathbf{N} = [f(\mathbf{r}_i - \mathbf{r}j)]_{i,\ j\in[L]} \in \mathbb{R}^{L\times L}$，其中 $\mathbf{r}_i \in \mathbb{R}^3$ 是第 i 个输入原子的位置， L 是输入原子的总数，并且

$$f(\mathbf{r}) = \sum_{t=1}^{T} \frac{w_t}{(\sqrt{2\pi}\sigma_t)^3} \exp\left(-\frac{\|\mathbf{r}\|^2}{2\sigma_t^2}\right).$$

请注意，相对位置编码仅计算原子之间的相对距离，因此自然保留了许多不变性质和等变性质。很容易看出， f 的傅里叶变换为

$$g(\boldsymbol{\xi}) = \sum_{t=1}^{T} w_t \exp\left(-2\pi^2\sigma_t^2\|\boldsymbol{\xi}\|^2\right),$$

这使我们能够**使用**第4节所述的技术，近似计算 FLT 中的相对位置编码掩码。

实现细节。我们采用了 3D-Graphormer 的大部分训练策略 [Shi 等人， 2022]。具体而言，我们训练了一个包含 12 层的标准 Performer，以及两个FLT ，层数分别为 10 层和 12 层。有关模型配置和训练的更多细节，请参阅补充材料。我们在域内验证集上评估模型性能。我们使用能量的平均绝对误差（MAE）和真实能量的阈值内能量（EwT）百分比来评估预测能量的准确率。

结果。结果见表3。此外，我们还在图2 中展示了模型的验证损失曲线，以便进行更全面的比较。显然，我们的FLT 模型在两项评估指标上均表现更佳，并能生成更准确的能量预测。例如，12 层FLT的能量平均绝对误差比

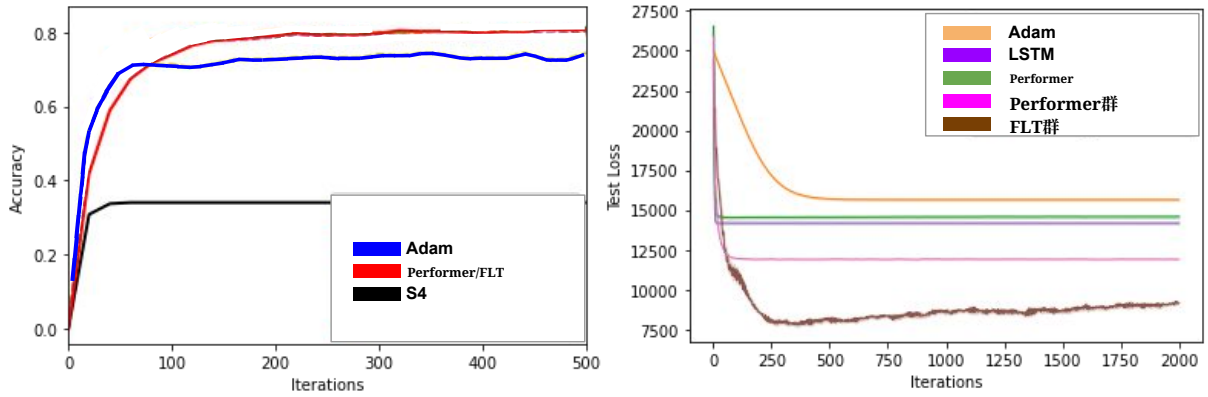


图 3：**可学习优化器实验结果。左：** Adam 和使用 FLT及 S4 的可学习优化器在 ImageNet 上训练 ViT-Base 分类器的任务中的表现。**右：** Adam 和各种可学习优化器在优化 Rastrigin 型函数任务中的表现（根据与 [Jain 等人， 2023]）。

12 层常规 Performer 低 0.04 eV 以上，这表明使用相对位置编码能够有效提升模型的预测能力。有人可能会认为，在 FLT 中使用相对位置编码会带来一些计算开销，并增加模型参数的数量。然而，值得注意的是，层数更少的 10 层 FLT 也能显著优于 12 层常规 Performer，同时运行速度更快，使用的参数也更少。

5.4 可学习优化器

FLT 也由作者和其他研究人员独立地在更长上下文中与其他类型的高效架构进行了比较，包括长短期记忆网络[Hochreiter 和 Schmidhuber， 1997] 以及状态空间模型 [Gu 等人， 2022]。相应的任务既实用又具有挑战性：在可学习优化器中将 Transformer 用作记忆模型（上下文长度最长可达 2000）。在这种设置下，长程时间推理（用于理解优化的历史）和空间推理（用于更全局地理解损失函数的地形）至关重要[Jain 等人，2023，Gärtner 等人， 2023]。

对比方法和实现细节。在第一个实验中，FLT-based 可学习优化器与基于 S4 的可学习优化器进行比较 [Gu 等人， 2022] 以及用于 ImageNet 分类的标准非可学习 Adam 优化器。

在第二个实验中，进一步将 FLT 应用于基于种群（群体）的可学习优化器，其中利用 FLT 实现的掩码机制来调节种群成员之间的相互关注方式。基线方法包括标准的非可学习 Adam 优化器，以及基于长短期记忆网络和 Performer 的可学习优化器。评估在类 Rastrigin 函数上进行。

结果。第一个/第二个实验的结果分别显示在图 3 的左/右图中。我们注意到，在两个实验中，FLT 方法都比其他所有变体取得了显著提升。结果表明，我们的模型在捕捉可学习优化器中的长程依赖方面始终有效。

6 结论

我们提出了 FourierLearner-Transformer（FLT），能够将相对位置编码（RPE）机制高效地适配到 Performers 中。Performers 是具有线性空间和时间复杂度的核化隐式注意力 Transformer。与其他将 Performers 与相对位置编码结合的架构不同，FLT 在不对相对位置编码掩码施加额外结构假设的情况下，仍能保持注意力模块的线性复杂度。我们提供了理论分析，并证明 FLT 能够准确近似 RPE。我们还进行了大量实验，展示 FLT 在多种任务和数据模态上的效率与性能，包括文本、图像、分子和优化器记忆。

致谢

我们感谢 Deepali Jain 就在可学习优化器中使用 FLT 进行讨论，并感谢其为作者提供的 FLT 变体提出并开展了基于群体（群体智能）的方法实验。我们也感谢审稿人提出的宝贵意见。

本研究部分由美国能源部通过布鲁克海文国家实验室提供支持，合同编号：384608。

参考文献

[Avron 等人， 2017] Avron, H.、Kapralov, M.、Musco, C.、Musco, C.、Velingker, A. 和 Zandieh, A.（2017）。用于核岭回归的随机傅里叶特征：近似界与统计保证。载于 Precup, D. 和 Teh, Y. W.（编），第*34*届国际机器学习会议论文集，*ICML 2017*，澳大利亚新南威尔士州悉尼，*2017*年*8*月*6*日至*11*日，第70卷机器学习研究论文集，第253–262页。PMLR。

[Bello， 2021] Bello，I.（2021）。Lambda 网络：无需注意力的长程相互作用建模。载于国际学习表征会议。

[Brown 等人， 2020] Brown，T. B.、Mann，B.、Ryder，N.、Subbiah，M.、Kaplan，J.、Dhariwal，P.、Neelakantan，A.、Shyam，P.、Sastry，G.、Askell，A.、Agarwal，S.、Herbert-Voss，A.、Krueger，G.、Henighan，T.、Child，R.、Ramesh，A.、Ziegler，D. M.、Wu，J.、Winter，C.、Hesse，C.、Chen，M.、Sigler，E.、Litwin，M.、Gray，S.、Chess，B.、Clark，J.、Berner，C.、McCandlish，S.、Radford，A.、Sutskever，I. 和 Amodei，D.（2020）。语言模型是少样本学习器。载于 Larochelle，H.、Ranzato，M.、Hadsell，R.、Balcan，M. 和 Lin，H.（编），神经信息处理系统进展第*33*卷：神经信息处理系统进展*2020*年度会议，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上举行。

[Chanussot* 等人， 2021] Chanussot*、L.、Das*、A.、Goyal*、S.、Lavril*、T.、Shuaibi*、M.、Riviere、M.、Tran、K.、Heras-Domingo、J.、Ho、C.、Hu、W.、Palizhati、A.、Sriram、A.、Wood、B.、Yoon、J.、Parikh、D.、Zitnick、C. L. 和 Ulissi、Z.（2021）。Open catalyst 2020（oc20）数据集和社区挑战赛。*ACS Catalysis*。

[Choromanski 等，2022a] Choromanski、K.、Lin、H.、Chen、H.、Zhang、T.、Sehanobish、A.、Likhosherstov、V.、Parker-Holder、J.、Sarlós、T.、Weller、A. 和 Weingarten、T.（2022a）。从块 Toeplitz 矩阵到图上的微分方程：迈向可扩展掩码 Transformer的一般理论。载于 Chaudhuri、K.、Jegelka、S.、Song、L.、Szepesvári、C.、Niu、G. 和 Sabato、S. 编，国际机器学习会议，*ICML 2022*，*2022*年*7*月*17*日至*23*日，巴尔的摩，马里兰州，美国，第162卷，机器学习研究论文集，第3962–3983页。PMLR。

[Choromanski 等人， 2021] Choromanski, K. M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlós, T., Hawkins, P., Davis, J. Q., Mohiuddin, A., Kaiser, L., Belanger, D. B., Colwell, L. J. 和 Weller, A.（2021）。重新思考注意力与 Performers。在第九届国际学习表示会议，*ICLR 2021*，线上会议，奥地利，*2021*年*5*月*3*日至*7*日。OpenReview.net。

[Choromanski 等人，2022b] Choromanski, K. M., Lin, H., Chen, H., Sehanobish, A., Ma, Y., Jain, D., Varley, J., Zeng, A., Ryoo, M. S., Likhosherstov, V., Kalashnikov, D., Sindhwani, V. 和 Weller, A.（2022b）。混合随机特征。发表于第十届国际学习表征会议，*ICLR 2022*，线上会议，*2022*年*4*月*25*日至*29*日。OpenReview.net。

[Chowdhury 等人， 2022] Chowdhury, S. P., Solomou, A., Dubey, A. 和 Sachan, M.（2022）。关于学习 Transformer 核。机器学习研究汇刊。

[Dai 等人， 2019] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. 和 Salakhutdinov, R.（2019）。Transformer-XL：超越固定长度上下文的注意力语言模型。发表于第*57*届计算语言学协会年会论文集，第2978–2988页，意大利佛罗伦萨。计算语言学协会。

[Deng 等人， 2009] Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. (2009). ImageNet：一个大型分层图像数据库。载于*2009*年*IEEE*计算机学会计算机视觉与模式识别会议（*CVPR 2009*），*2009*年*6*月*20*日至*25*日，美国佛罗里达州迈阿密，第248–255页。IEEE计算机学会。

[Devlin 等人，2019] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT：用于语言理解的深度双向 Transformer 预训练。载于 Burstein, J., Doran, C., and Solorio, T. 编，*2019*年北美计算语言学协会会议论文集：人类语言技术，*NAACL-HLT 2019*，美国明尼苏达州明尼阿波利斯，*2019*年*6*月*2*日至*7*日，第*1*卷（长篇与短篇论文），第4171–4186页。计算语言学协会。

[Dosovitskiy 等人， 2021] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N.（2021）。一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer。载于第九届国际学习表征会议，*ICLR 2021*，线上会议，奥地利，*2021*年*5*月*3*日至*7*日。OpenReview.net。

[Gärtner 等人，2023] Gärtner，E.、Metz，L.、Andriluka，M.、Freeman，C. D. 和 Sminchisescu，C.（2023）。

基于Transformer的学习优化。在*IEEE/CVF*计算机视觉与模式识别会议，第11970–11979页。

[Gasteiger 等人， 2021] Gasteiger, J.、Becker, F. 和 Günnemann, S.（2021）。GemNet：用于分子的通用方向图神经网络。神经信息处理系统进展，34：6790–6802。

[Geng 等人， 2021] Geng, Z.、Guo, M.-H.、Chen, H.、Li, X.、Wei, K. 和 Lin, Z.（2021）。注意力是否优于矩阵分解？载于国际学习表征会议。

[Gu 等人， 2022] Gu, A.、Goel, K. 和 Ré, C.（2022）。利用结构化状态空间高效建模长序列。载于第十届国际学习表征会议，*ICLR 2022*，线上会议，*2022*年*4*月*25*日至*29*日。OpenReview.net。

[Hochreiter 和 Schmidhuber，1997] Hochreiter, S. 和 Schmidhuber, J. (1997). 长短期记忆。*Neural Computation*, 9(8):1735–1780.

[Horn 等人， 2021] Horn, M., Shridhar, K., Groenewald, E. 和 Baumann, P. F. M. (2021). 可核化注意力中的平移等变性。*CoRR*, abs/2102.07680.

[Jain 等人， 2023] Jain, D., Choromanski, K. M., Dubey, K. A., Singh, S., Sindhwani, V., Zhang, T. 和 Tan, J. (2023). Mnemosyne：学习使用 Transformer 训练 Transformer。载于第三十七届神经信息处理系统大会。

[Jumper 等人， 2021] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. 等人 (2021). 使用 AlphaFold 实现高精度蛋白质结构预测。*Nature*, 596(7873):583–589.

[Katharopoulos 等人， 2020] Katharopoulos、A.、Vyas、A.、Pappas、N. 和 Fleuret、F.（2020）。Transformer 即循环神经网络：采用线性注意力的快速自回归Transformer。载于国际机器学习会议，第 5156–5165 页。PMLR。

[Kitaev 等人， 2020] Kitaev、N.、Kaiser、L. 和 Levskaya、A.（2020）。Reformer：高效的Transformer。载于第八届国际学习表征会议，*ICLR 2020*，埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26*日至*30*日。OpenReview.net。

[Li 等人， 2021] Li, S., Chen, X., He, D., 和 Hsieh, C.-J.（2021）。视觉 Transformer 能够执行卷积吗？*arXiv* 预印本 *arXiv:2111.01353*。

[Li 等人， 2019] Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., 和 Yan, X.（2019）。增强时间序列预测中 Transformer 的局部性并突破内存瓶颈。神经信息处理系统进展，32。

[Li 等人， 2024] Li、S.、You、C.、Guruganesh、G.、Ainslie、J.、Ontanon、S.、Zaheer、M.、Sanghai、S.、Yang、Y.、Kumar、S. 和 Bhojanapalli、S.（2024）。相对位置的函数插值可改进长上下文 Transformer。载于第十二届国际学习表征会议。

[Li 和 Li，2021] Li, X. 和 Li, P.（2021）。随机傅里叶特征的量化算法。载于 Meila, M. 和 Zhang, T. 编，第*38*届国际机器学习会议论文集，*ICML 2021*，*2021*年*7*月*18*日至*24*日，线上会议，第139卷，机器学习研究论文集，第6369–6380页。PMLR。

[Likhosherstov 等人， 2022] Likhosherstov, V., Choro- manski, K. M., Dubey, K. A., Liu, F., Sarlos, T. 和 Weller, A.（2022）。Chefs' 随机表：非三角函数随机特征。载于 Koyejo, S., Mo- hamed, S., Agarwal, A., Belgrave, D., Cho, K. 和 Oh, A. 编，神经信息处理系统进展，第35卷，第34559–34573页。Curran Associates, Inc。

[Liu 等人， 2024] Liu, N. F., Lin, K., Hewitt, J., Paran- jape, A., Bevilacqua, M., Petroni, F. 和 Liang, P.（2024）。迷失在中间：语言模型如何利用长上下文。计算语言学协会会刊，12：157–173。

[Liutkus 等人， 2021] Liutkus, A., Cífka, O., Wu, S., Simsekli, U., Yang, Y. 和 Richard, G.（2021）。用于 Transformer 的线性复杂度相对位置编码。载于 Meila, M. 和 Zhang, T.（编），第*38*届国际机器学习会议论文集，*ICML 2021*，*2021*年*7*月*18*日至*24*日，线上会议，第139卷，机器学习研究论文集，第7067–7079页。PMLR。

[Luo 等人， 2023] Luo、Chen、Xu、Zheng、Liu、Wang 和 He（2023）。一个 Transformer 能够同时理解二维和三维分子数据。发表于第十一届国际学习表征会议。

[Luo 等人， 2021] Luo, S.、Li, S.、Cai, T.、He, D.、Peng, D.、Zheng, S.、Ke, G.、Wang, L. 和 Liu, T.-Y.（2021）。稳定、快速且准确：带有相对位置编码的核化注意力。神经信息处理系统进展，34：22795–22807。

[Luo 等人， 2022] Luo, S.、Li, S.、Zheng, S.、Liu, T.-Y.、Wang, L. 和 He, D.（2022）。你的 Transformer 可能没有你想象的那么强大。载于 Oh, A. H.、Agarwal, A.、Belgrave, D. 和 Cho, K. 编，神经信息处理系统进展。

[Moreno-Pino 等人， 2023] Moreno-Pino，F.、Olmos，P. M. 和 Artés-Rodríguez，A.（2023）。具有谱注意力的深度自回归模型。模式识别，133：109014。

[O'Connor 和 Andreas， 2021] O'Connor，J. 和 Andreas，J.（2021）。Transformer 语言模型可以使用哪些上下文特征？载于计算语言学协会第 *59* 届年会暨第 *11* 届自然语言处理国际联合会议（第一卷：长篇论文），第 851–864 页。

[Peng 等人，2021] Peng，H.、Pappas，N.、Yogatama，D.、Schwartz，R.、Smith，N. 和 Kong，L.（2021）。随机特征注意力。载于国际学习表征会议。

[Press 等人，2022] Press, O., Smith, N., 和 Lewis, M.（2022）。短序列训练，长序列测试：带线性偏置的注意力支持输入长度外推。发表于国际学习表征会议。

[Qin 等人， 2022] Qin, Z., Sun, W., Deng, H., Li, D., Wei, Y., Lv, B., Yan, J., Kong, L., 和 Zhong, Y.（2022）。CosFormer：重新思考注意力中的 Softmax。*CoRR*，abs/2202.08791。

[Raffel 等人， 2020] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., 和 Liu, P. J.（2020）。探索统一文本到文本转换器的迁移学习极限。机器学习研究杂志，21:140:1–140:67。

[Rahimi 和 Recht， 2007] Rahimi，A. 和 Recht，B.（2007）。用于大规模核机器的随机特征。载于 Platt，J. C.、Koller，D.、Singer，Y. 和 Roweis，S. T.（编），神经信息处理系统进展 *20*：第二十一届神经信息处理系统年会论文集，*2007*年*12*月*3*日至*6*日，加拿大不列颠哥伦比亚省温哥华，第 1177–1184 页。Curran Associates, Inc.

[Roy 等人， 2021] Roy，A.、Saffar，M.、Vaswani，A. 和 Grangier，D.（2021）。结合路由 Transformer 实现高效的基于内容的稀疏注意力。计算语言学协会会刊，9：53–68。

[Ruoss 等人， 2023] Ruoss、Delétang、Ge-newein、Grau-Moya、Csordás、Bennani、Legg 和 Veness（2023）。随机位置编码提升 Transformer 的长度泛化能力。*arXiv* 预印本 *arXiv:2305.16843*。

[Shaw 等人， 2018] Shaw、Uszkoreit 和 Vaswani（2018）。带有相对位置表示的自注意力。载于 Walker、Ji 和 Stent（编），*2018*年北美计算语言学协会年会：人类语言技术会议论文集（*NAACL-HLT*），美国路易斯安那州新奥尔良，*2018*年*6*月*1*日至*6*日，第*2*卷（短篇论文），第464–468页。计算语言学协会。[Shi 等人，2022] Shi、Zheng、Ke、Shen、You、He、Luo、Liu、He 和 Liu（2022）。在大规模分子建模数据集上对 Graphormer 进行基准测试。*arXiv* 预印本 *arXiv:2203.04810*。

[Sinha 和 Duchi，2016] Sinha，A. 和 Duchi（2016）。利用随机特征学习核。载于 Lee、Sugiyama、von Luxburg、Guyon 和 Garnett（编），神经信息处理系统进展 *29*：*2016*年神经信息处理系统年会，*2016*年*12*月*5*日至*10*日，西班牙巴塞罗那，第 1298–1306 页。

[Sun 等人， 2022] Sun、Yang 和 Yoo（2022）。利用学习哈希实现稀疏注意力。载于第十届国际学习表征会议，*ICLR 2022*，线上会议，*2022*年*4*月*25*日至*29*日。OpenReview.net。

[Szabó 和 Sriperumbudur， 2019] Szabó 和 Sriperumbudur（2019）。关于利用随机傅里叶特征近似核导数。载于 Chaudhuri 和 Sugiyama 编，第二十二届人工智能与统计国际会议，*AISTATS 2019*，*2019*年*4*月*16*日至*18*日，日本冲绳那霸，第 89 卷机器学习研究论文集，第 827–836 页。PMLR。

[Tamkin 等人， 2020] Tamkin、Jurafsky 和 Goodman（2020）。棱镜中的语言：一种用于多尺度语言表示的谱方法。神经信息处理系统进展，33：5492–5504。

[Tay 等人， 2021] Tay, Y., Dehghani, M., Aribandi, V., Gupta, J., Pham, P. M., Qin, Z., Bahri, D., Juan, D.-C., and Metzler, D. (2021). Omninet：来自 Transformer 的全向表示。载于机器学习国际会议，第 10193–10202 页。PMLR。

[Tsai 等人， 2019] Tsai, Y.-H. H., Bai, S., Yamada, M., Morency, L.-P., and Salakhutdinov, R. (2019). Transformer剖析：从核的视角统一理解 Transformer 的注意力。载于

*2019*年自然语言处理实证方法会议论文集暨第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*），第4344–4353页。

[Vaswani 等人，2021a] Vaswani、A.、Ramachandran、P.、Srinivas、A.、Parmar、N.、Hechtman、B. A. 和 Shlens、J.（2021a）。扩展局部自注意力，以实现参数高效的视觉骨干网络。载于*IEEE* 计算机视觉与模式识别会议（*CVPR 2021*），线上举行，*2021*年*6*月*19*日至*25*日，第12894–12904页。计算机视觉基金会 / IEEE。

[Vaswani 等人，2021b] Vaswani、A.、Ramachandran、P.、Srinivas、A.、Parmar、N.、Hechtman、B. A. 和 Shlens、J.（2021b）。扩展局部自注意力，以实现参数高效的视觉骨干网络。载于*IEEE* 计算机视觉与模式识别会议（*CVPR 2021*），线上举行，*2021*年*6*月*19*日至*25*日，第12894–12904页。计算机视觉基金会 / IEEE。

[Vyas 等人， 2020] Vyas、A.、Katharopoulos、A. 和 Fleuret、F.（2020）。采用聚类注意力的快速 Transformer。载于 Larochelle、H.、Ranzato、M.、Hadsell、R.、Balcan、M. 和 Lin、H.（编），神经信息处理系统进展第*33*卷：神经信息处理系统*2020*年年会，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上举行。

[Wu 等人， 2021] Wu、K.、Peng、H.、Chen、M.、Fu、J. 和 Chao、H.（2021）。重新思考并改进视觉 Transformer的相对位置编码。载于*2021 IEEE/CVF* 国际计算机视觉会议，*ICCV 2021*，蒙特利尔，魁北克省，加拿大，*2021*年*10*月*10*日至*17*日，第10013–10021页。IEEE。

[Xiao 等人， 2017] Xiao, H., Rasul, K., 和 Vollgraf, R. (2017)。Fashion-MNIST：一个用于机器学习算法基准测试的新型图像数据集。*CoRR*, abs/1708.07747。

[Xiao 等人， 2022] Xiao, X., Zhang, T., Choromanski, K., Lee, T. E., Francis, A. G., Varley, J., Tu, S., Singh, S., Xu, P., Xia, F., Persson, S. M., Kalash- nikov, D., Takayama, L., Frostig, R., Tan, J., Parada, C., 和 Sindhwani, V. (2022)。通过实时注意力学习用于真实世界导航的模型预测控制器。*CoRL 2022*, abs/2209.10780。

[Yang 等人， 2024] Yang, K., Ackermann, J., He, Z., Feng, G., Zhang, B., Feng, Y., Ye, Q., He, D., 和 Wang, L. (2024)。高效 Transformer 是否真的节省计算量？*arXiv* 预印本 *arXiv:2402.13934*。

[Yuan 等人， 2021] Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F. E. H., Feng, J., and Yan, S. (2021). Tokens-to-token ViT：训练视觉 Transformer：在 ImageNet 上从零开始训练。在 *2021* 年 *IEEE/CVF* 国际计算机视觉大会（*ICCV 2021*），加拿大魁北克省蒙特利尔，*2021* 年 *10* 月 *10* 日至 *17* 日，第 538–547 页。IEEE。

[Yun 等人， 2020] Yun, C., Chang, Y.-W., Bhojanapalli, S., Rawat, A. S., Reddi, S., and Kumar, S. (2020). O(n) 连接已足够 expressive：稀疏 Transformer 的通用可逼近性。神经信息处理系统进展, 33:13783–13794.

[Zaheer 等人， 2020] Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al. (2020). Big Bird：面向更长序列的 Transformer。神经信息处理系统进展, 33:17283–17297.

[Zhou 等人， 2018] Zhou, B., Lapedriza, À., Khosla, A., Oliva, A., and Torralba, A. (2018). Places：用于场景识别的千万级图像数据库。*IEEE* 模式分析与机器智能汇刊, 40(6):1452–1464.

[Žiga Avsec 等人， 2021] Žiga Avsec、Agarwal, V.、Visentin, D.、Ledsam, J. R.、Grabska-Barwinska, A.、Taylor, K. R.、Assael, Y.、Jumper, J. M.、Kohli, P.、以及 Kelley, D. R.（2021）。通过整合长程相互作用，从序列中进行有效的基因表达预测。《自然·方法》，18：1196–1203。

A 省略的理论结果与证明

我们在此给出省略的理论结果、证明，以及关于 FLT 的相对位置编码近似的讨论。我们首先证明主论文中提出的相对位置编码近似是无偏的。随后，我们给出该近似的方差界和近似保证。

为便于阅读，我们总是在给出证明之前陈述定理，即使该定理已在论文正文中出现过。

A.1 无偏 RPE 近似

定理 A.1. [4] 给定 $f\colon \mathbb{R}^\ell \to \mathbb{R}$ 以及 $\mathbf{N} = [f(\mathbf{r}_i - \mathbf{r}_j)] \in \mathbb{R}^{L\times L}$，如定义 *3.1* 中所定义，记 g 为 f的傅里叶变换。假设 p 是一个支撑于 $\mathbb{R}^\ell$上的概率密度函数。采样 $\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_r \overset{\text{独立同分布}}{\sim} p$，并定义以下随机特征映射（其中 $\mathbf{i} = \sqrt{-1}$）：

$$\varphi(\mathbf{z}) = \frac{1}{\sqrt{r}}\left[\mathrm{e}^{2\pi\mathbf{i}\mathbf{z}^\top\boldsymbol{\xi}_1}\sqrt{\frac{g(\boldsymbol{\xi}_1)}{p(\boldsymbol{\xi}_1)}}, \cdots, \mathrm{e}^{2\pi\mathbf{i}\mathbf{z}^\top\boldsymbol{\xi}_r}\sqrt{\frac{g(\boldsymbol{\xi}_r)}{p(\boldsymbol{\xi}_r)}}\right]^\top;$$
$$\psi(\mathbf{z}) = \frac{1}{\sqrt{r}}\left[\mathrm{e}^{-2\pi\mathbf{i}\mathbf{z}^\top\boldsymbol{\xi}_1}\sqrt{\frac{g(\boldsymbol{\xi}_1)}{p(\boldsymbol{\xi}_1)}}, \cdots, \mathrm{e}^{-2\pi\mathbf{i}\mathbf{z}^\top\boldsymbol{\xi}_r}\sqrt{\frac{g(\boldsymbol{\xi}_r)}{p(\boldsymbol{\xi}_r)}}\right]^\top,$$

定义 $\mathbf{N}_1 = [\varphi(\mathbf{r}_1)\ \cdots, \varphi(\mathbf{r}_L)]^\top \in \mathbb{R}^{L\times r}$和 $\mathbf{N}_2 = [\psi(\mathbf{r}_1)\ \cdots, \psi(\mathbf{r}_L)]^\top \in \mathbb{R}^{L\times r}$。那么

$$\mathbb{E}[\mathbf{N}_1\mathbf{N}_2] = \mathbf{N}.$$

证明。 根据$\mathbf{N}$的定义，只需证明

$$f(\mathbf{r}_i - \mathbf{r}_j) = \mathbb{E}\left[\varphi(\mathbf{r}_i)^\top\psi(\mathbf{r}_j)\right].$$

注意， g 是 f的傅里叶变换。因此，

$$f(\mathbf{x}) = \int_{\mathbb{R}^d}\mathrm{e}^{2\pi\mathbf{i}\mathbf{x}^\top\boldsymbol{\xi}}g(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \int_{\mathbb{R}^d}\mathrm{e}^{2\pi\mathbf{i}\mathbf{x}^\top\boldsymbol{\xi}}\frac{g(\boldsymbol{\xi})}{p(\boldsymbol{\xi})}\cdot p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \mathbb{E}_{\boldsymbol{\xi}\sim p}\left[\mathrm{e}^{2\pi\mathbf{i}\mathbf{x}^\top\boldsymbol{\xi}}\frac{g(\boldsymbol{\xi})}{p(\boldsymbol{\xi})}\right]. \tag{5}$$
$$\Rightarrow\quad f(\mathbf{r}_i - \mathbf{r}_j) = \mathbb{E}_{\boldsymbol{\xi}\sim p}\left[\mathrm{e}^{2\pi\mathbf{i}\mathbf{r}_i^\top\boldsymbol{\xi}}\sqrt{\frac{g(\boldsymbol{\xi})}{p(\boldsymbol{\xi})}}\cdot\mathrm{e}^{-2\pi\mathbf{i}\mathbf{r}_j^\top\boldsymbol{\xi}}\sqrt{\frac{g(\boldsymbol{\xi})}{p(\boldsymbol{\xi})}}\right].$$

同时，根据 φ 和 ψ的定义，我们有

$$\varphi(\mathbf{r}_i)^\top\psi(\mathbf{r}_j) = \frac{1}{r}\sum_{k=1}^{r}\mathrm{e}^{2\pi\mathbf{i}\mathbf{r}_i^\top\boldsymbol{\xi}_k}\sqrt{\frac{g(\boldsymbol{\xi}_k)}{p(\boldsymbol{\xi}_k)}}\cdot\mathrm{e}^{-2\pi\mathbf{i}\mathbf{r}_j^\top\boldsymbol{\xi}_k}\sqrt{\frac{g(\boldsymbol{\xi}_k)}{p(\boldsymbol{\xi}_k)}}$$

最后，注意到 $\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_m \sim \text{i.i.d. } p$。根据期望的线性性，我们有 $f(\mathbf{r}_i - \mathbf{r}j) = \mathbb{E}\left[\varphi(\mathbf{r}_i)^\top\psi(\mathbf{r}j)\right]$，由此得证。

□

A.2 RPE 近似的方差

引理 A.2。 假设 $c = \||g(\mathbf{x})|/p(\mathbf{x})\|_\infty$，其中 g 是相对位置编码函数 f 的傅里叶变换， p是某个概率分布的概率密度函数。那么对于任意$\mathbf{z} \in \mathbb{R}^\ell$，以下结论成立：

$$\mathrm{Var}_{\boldsymbol{\xi}\sim p}\left[\mathrm{e}^{2\pi\mathbf{i}\mathbf{z}^\top\boldsymbol{\xi}}\frac{g(\boldsymbol{\xi})}{p(\boldsymbol{\xi})}\right] \le c^2 - f(\mathbf{z})^2. \tag{6}$$

[4]这是定理 4.1 ，见主论文。

证明。 回顾一下，对于复随机变量 Z，其方差定义为

$$\mathrm{Var}[Z] = \mathbb{E}\left[(Z - \mathbb{E}\left[Z\right])(Z - \mathbb{E}\left[Z\right])^*\right] = \mathbb{E}\left[ZZ^*\right] - \mathbb{E}\left[Z\right]\mathbb{E}\left[Z\right]^*, \tag{7}$$

其中 * 表示共轭。

经过直接计算可得

$$\mathbb{E}\left[\left(\mathrm{e}^{2\pi\mathrm{i}\mathbf{z}^\top\boldsymbol{\xi}}\frac{g(\boldsymbol{\xi})}{p(\boldsymbol{\xi})}\right)\left(\mathrm{e}^{2\pi\mathrm{i}\mathbf{z}^\top\boldsymbol{\xi}}\frac{g(\boldsymbol{\xi})}{p(\boldsymbol{\xi})}\right)^*\right] = \mathbb{E}\left[\mathrm{e}^{2\pi\mathrm{i}\mathbf{z}^\top\boldsymbol{\xi}}\frac{g(\boldsymbol{\xi})}{p(\boldsymbol{\xi})}\cdot\mathrm{e}^{-2\pi\mathrm{i}\mathbf{z}^\top\boldsymbol{\xi}}\frac{g(\boldsymbol{\xi})^*}{p(\boldsymbol{\xi})}\right] \tag{8}$$

$$= \mathbb{E}\left[\frac{|g(\boldsymbol{\xi})|^2}{p(\boldsymbol{\xi})^2}\right] \leq c^2. \tag{9}$$

此外，式（5）意味着

$$\mathbb{E}\left[\mathrm{e}^{2\pi\mathrm{i}\mathbf{z}^\top\boldsymbol{\xi}}\frac{g(\boldsymbol{\xi})}{p(\boldsymbol{\xi})}\right] = f(\mathbf{z}). \tag{10}$$

Pl将上述两个结果代入式（7）可得式（6），从而完成证明 . □

定理 A.3 (RPE 近似的方差)。 在引理 *A.2* 的假设下，对于任意 $\mathbf{x}$，$\mathbf{y} \in \mathbb{R}^\ell$，定理 $\varphi(\mathbf{x})^\top\psi(\mathbf{y})$ 中给出的近似的方差满足

$$\mathrm{Var}[\varphi(\mathbf{x})^\top\psi(\mathbf{y})] \leq \frac{c^2 - f(\mathbf{x}-\mathbf{y})^2}{r}. \tag{11}$$

证明。注意到

$$\varphi(\mathbf{x})^\top\psi(\mathbf{y}) = \frac{1}{r}\sum_{t=1}^{r}\mathrm{e}^{2\pi\mathrm{i}(\mathbf{x}-\mathbf{y})^\top\boldsymbol{\xi}_t}\frac{g(\boldsymbol{\xi}_t)}{p(\boldsymbol{\xi}_t)}, \tag{12}$$

其中，随机特征 $\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_r$ 是来自分布 p 的 r 独立同分布样本。

Se将 $\mathbf{z} = \mathbf{x} - \mathbf{y}$ 代入式（6），并考虑 r 独立同分布样本，即可得到式（11） . □

A.3 相对位置编码近似的一致收敛与样本复杂度

定理 A.4 (一致收敛与近似的样本复杂度)。 [5] 给定向量 $\mathbf{r}_1$，...，$\mathbf{r}_L \in \mathbb{R}^\ell$，定义相对位置编码注意力掩码 $\mathbf{N} = [f(\mathbf{r}_i - \mathbf{r}_j)]_{i,j\in[L]}$。假设 $c = \||g(\mathbf{x})|/p(\mathbf{x})\|_\infty$，其中 g 是 f 的傅里叶变换，且 p 是 $\mathbb{R}^\ell$上的某个概率密度函数。

对于任意 ε、$\delta > 0$，若随机特征的数量为 $r = \Theta\left(\frac{c^2}{\varepsilon^2}\log\frac{L}{\delta}\right)$，则 FLT 的相对位置编码近似器 $\widehat{\mathbf{N}}$ 满足

$$\mathbb{P}\left(\|\mathbf{N} - \widehat{\mathbf{N}}\|_{\max} \leq \varepsilon\right) > 1 - \delta,$$

其中 $\|\cdot\|_{\max}$ 表示矩阵的最大范数。

证明。对于任意 i，$j \in \{1, \cdots, L\}$，假设 $c = \||g(\mathbf{x})|/p(\mathbf{x})\|_\infty$ 意味着几乎必然有

$$\left|\mathrm{e}^{2\pi\mathrm{i}(\mathbf{r}_i-\mathbf{r}_j)^\top\boldsymbol{\xi}_t}\frac{g(\boldsymbol{\xi}_t)}{p(\boldsymbol{\xi}_t)}\right| \leq c \qquad (\forall\, t \in \{1, \cdots, r\}), \tag{13}$$

其中 $\boldsymbol{\xi}_1$，$\cdots$，$\boldsymbol{\xi}_r$ 表示来自分布 p 的 r 随机特征。

注意，式（13）表明，估计得到的相对位置编码的实部和虚部都有界。应用霍夫丁不等式（分别针对实部和虚部）可得

$$\mathbb{P}\left(\left|\frac{1}{t}\sum_{t=1}^{r}\mathrm{e}^{2\pi\mathrm{i}(\mathbf{r}_i-\mathbf{r}_j)^\top\boldsymbol{\xi}_t}\frac{g(\boldsymbol{\xi}_t)}{p(\boldsymbol{\xi}_t)} - f(\mathbf{r}_i - \mathbf{r}_j)\right| > \varepsilon\right) < 4\mathrm{e}^{-\frac{r\varepsilon^2}{4c^2}}. \tag{14}$$

[5] 这是4.2 主论文中的定理。

对 $i,\ j \in \{1,\cdots,L\}$ 使用并集界，可得

$$\mathbb{P}\left(\exists\, i,j\in\{1,\cdots,L\},\ \text{s.t.}\ \left|\frac{1}{t}\sum_{t=1}^{r}\mathrm{e}^{2\pi\mathrm{i}(\mathbf{r}_i-\mathbf{r}_j)^\top\boldsymbol{\xi}_t}\frac{g(\boldsymbol{\xi}_t)}{p(\boldsymbol{\xi}_t)}-f(\mathbf{r}_i-\mathbf{r}_j)\right|>\varepsilon\right)<4L^2\mathrm{e}^{-\frac{r\varepsilon^2}{4c^2}}. \tag{15}$$

等价地，以至少 $4L^2\mathrm{e}^{-\frac{r\varepsilon^2}{4c^2}}$ 的概率，我们有 —

$$\left|\frac{1}{r}\sum_{t=1}^{r}\mathrm{e}^{2\pi\mathrm{i}(\mathbf{r}_i-\mathbf{r}_j)^\top\boldsymbol{\xi}_t}\frac{g(\boldsymbol{\xi}_t)}{p(\boldsymbol{\xi}_t)}-f(\mathbf{r}_i-\mathbf{r}_j)\right|>\varepsilon \qquad (\forall\, i,j\in\{1,\cdots,L\}) \tag{16}$$

$$\Rightarrow \qquad \|\mathbf{N}-\widehat{\mathbf{N}}\|_{\max}\leq\varepsilon. \tag{17}$$

Set

$$r=\frac{4c^2}{\varepsilon^2}\log\frac{4L^2}{\delta}=\Theta\left(\frac{c^2}{\varepsilon^2}\log\frac{L}{\delta}\right). \tag{18}$$

于是我们有

$$\mathbb{P}\left(\|\mathbf{N}-\widehat{\mathbf{N}}\|_{\max}\leq\varepsilon\right)>1-4L^2\mathrm{e}^{-\frac{r\varepsilon^2}{4c^2}}=1-\delta, \tag{19}$$

由此得证。 □

A.4 关于理论结果的讨论

^定理A.4 表明FLT估计得到的相对位置编码掩码 $\mathbf{N}$ 能够以高概率任意精确地近似真实的相对位置编码掩码 $\mathbf{N}$ 。此外，请注意，对于某个预训练模型而言，常数 c 可以视为固定值，且不依赖于 L。因此，要得到任意精确的相对位置编码近似器，所需的随机特征数量仅随 对数级增长于 L。这一性质尤其吸引人，因为它表明FLT在长序列场景下仍能保持准确，同时加速强大的 RPE增强注意力。

定理A.3 和A.4 还为寻找最优的 p提供了见解。请注意，方差和样本复杂度都随 $c=\||g(\mathbf{x})|/p(\mathbf{x})\|_\infty$变化，而较低的 c 可能带来更好的近似。具体而言，选择 $p(\mathbf{x})\propto|g(\mathbf{x})|$ 能够在 c 满足 p 为概率密度函数的约束下使其最小化。相对位置编码的平移不变核确实满足这一性质，并且在近似样本复杂度方面是最优的。

另一种简单而有效的方法是将 p 参数化为具有 可学习的 均值和方差的高斯分布。优化过程可以在高斯分布类中搜索最优的 p ，以获得良好的相对位置编码近似。这项技术对分子性质预测实验（主论文第 5.3 节）很有帮助。

最后，我们指出，定理 A.4 中的样本复杂度界的 对数 L 因子是出于技术原因引入的：收敛性分析针对应用了指数映射的随机特征进行，而该映射并无界。[6] 尽管如此，这个对数因子仍可忽略不计（与多项式依赖不同）。

B 详细实验设置

所有测试的 Transformer 变体均在包含 4 块 TPU v3 芯片 的 TPU pod 上使用 JAX进行训练和测试，并在图形处理器（V100）上进行训练和测试。

B.1 语言建模

在本实验中，我们研究了具有两种相对位置编码变体的 FLT：高斯混合相对位置编码和局部相对位置编码。主论文中已详细介绍基线方法。

[6]对于应用三角函数的随机特征，因此可以利用 ε-net 技巧，并结合强 Lipschitz 函数论证。类似的结果可见于 [Choromanski 等人, 2021]

所有测试模型均为基于核化低秩注意力的高效 Transformer，包含 6 个解码器层。每层有 8 个注意力头。隐藏维度设为 512。前馈子层的维度设为 2048。在注意力矩阵的低秩近似中，特征映射维度 m 设为 64。对于我们的 FLT 模型，相对位置编码（RPE）的随机特征数 r 设为 32。我们使用验证困惑度作为评估指标：困惑度越低，性能越好。

遵循现有研究 [Peng 等，2021、Luo 等人， 2021], 训练和评估期间的序列长度均设为 512。所有模型的训练过程均不使用前几个小批次中的上下文，以确保公平比较。丢弃率和权重衰减分别设为 0.1 和 0.01。批量大小设为 64。我们使用Adam 作为优化器，并将其超参数 ε 设为$1\mathrm{e}-6$ 和$(\beta_1,\ \beta_2)$ 设为 (0.9, 0.98)。模型训练 150k 步，先经过 6k 步的预热阶段，随后采用逆平方根学习率调度器，峰值学习率设为$2\mathrm{e}-3$。

对于采用高斯混合相对位置编码的 FLT变体，相对位置编码函数的傅里叶变换，即函数 g，按式（4.3）进行参数化：

$$g(\boldsymbol{\xi}) = \sum_{t=1}^{T} w_t \exp\left(-\frac{\|\boldsymbol{\xi}-\boldsymbol{\mu}_t\|^2}{2\sigma_t^2}\right).$$

对于采用局部相对位置编码的 FLT 变体，函数 g 的参数化形式为

$$g(\xi) = \sum_{t=1}^{T} w_t \cdot \frac{\sin(2\pi v_t \xi)}{\pi\xi}, \tag{20}$$

其中 $w_1,\cdots,w_T$和 $v_1,\cdots,v_T$是可学习参数，而 T是预定义的超参数。在这种情况下，隐式的相对位置编码函数 f 为

$$f(\Delta r) = \sum_{t=1}^{T} w_t \cdot \mathbb{I}[|\Delta r| \leq v_t]. \tag{21}$$

对于这两种 FLT 变体，不同注意力头使用不同的相对位置编码掩码，但在不同层之间共享。随机特征 $\xi_1,\cdots,\xi_r$ 从标准高斯分布中采样得到。

B.2 图像分类

表4 预　　列出了图像分类实验中所用数据集的基本统计信息。

表 4：使用 FourierLearner-Transformer 进行图像分类任务所用数据集的详细信息。

数据集名称	类别数	训练集大小	测试集大小
ImageNet2012 [Deng 等人， 2009]	1K	1.2M	100K
Places365 [Zhou 等人， 2018]	365	1.8M	328K
Fashion-MNIST [Xiao 等人， 2017]	10	60K	10K

所有测试模型均由 12 层组成，每层包含 12 个注意力头。前馈子层的维度设为 3072。在我们的 FLT 中，我们采用可学习 ReLU 作为核化线性注意力的特征映射。具体而言，特征映射为 ϕ：$\mathbf{x} \mapsto \mathbf{W}\mathbf{x}$ 其中 $\mathbf{W}$ 是一个可学习矩阵。对于所有模型，我们采用 0.1 的丢弃率，且不使用注意力丢弃。我们使用权重衰减为 0.05 的 Adam 优化器，批量大小设为标准值 4096。所有 Transformer 均在 TPU 架构上训练至收敛。

B.3 分子性质预测

我们采用了 3D-Graphormer 的大多数训练策略 [Shi 等人， 2022]。具体而言，我们训练了一个具有 12 层的常规 Performer，以及两个分别具有 10 层和 12 层的 FLT。遵循现有工作[Jumper 等人，2021，Shi 等人， 2022], 模型输出会被重复输入模型四次。在每一层中，

表 5：图像分类的超参数。

参数	值
批量大小	4096
优化器	AdamW
基础学习率	1。5e -4
权重衰减	0.05
优化器动量	$(\beta_1, \beta_2) = (0.9, 0.95)$
学习率调度	余弦衰减
预热轮数	40
数据增强	随机调整大小裁剪
计算资源	8×8 TPUv3

有 48 个注意力头。隐藏维度设为 768。前馈子层的维度设为 2048。在注意力矩阵的低秩近似中，特征映射维度 m 设为 64。对于我们的 FLT 模型，相对位置编码 r 的随机特征数量设为 16，相对位置编码 T中的高斯基函数数量设为 32。随机特征 ξ_i从高斯分布 $\mathcal{N}(0, \sigma_i^2\mathbf{I})$ 中采样，其中σ_i 是可学习的。

我们在域内验证集上评估所测试模型的性能，其中验证样本来自与训练分布相同的分布。我们使用能量的平均绝对误差（MAE）以及真实能量的阈值内能量（EwT）百分比来评估预测能量的准确率。

对于所有模型，注意力丢弃率和权重衰减分别设为 0.1 和 0.001。批量大小设为 64。我们使用 Adam 作为优化器，并将其超参数 ε 设为 1e -6，将 (β_1, β_2) 设为 (0.9, 0.98)。峰值学习率设为 3e -4，并采用 10K 步的预热阶段。预热阶段结束后，学习率线性衰减至零。所有模型总共训练 500k 步。

B.4 可学习优化器

在所有实验中，均采用了来自 [Jain 等人， 2023] 的元训练流程和训练方案。遵循 [Jain 等人， 2023], 所有可学习模型均作为相应可学习优化器中的记忆单元使用，并以完全相同的方式在一小组无关的优化任务上进行元训练。此外，所有基于注意力的记忆机制均源自 [Jain 等人， 2023]。

C 可视化

C.1 局部相对位置编码

在图4中，我们可视 化通过傅里叶变换可由 FLTs 建模的局部相对位置编码的形状。

C.2 注意力矩阵

在图 5中，我们将一个经过 WikiText-103 语言建模训练的 FLT 模型的注意力矩阵可视化。具体而言，我们将训练集中的一个序列输入模型，并将第一层中 8 个注意力头的注意力矩阵可视化。可以看到，一些注意力头更加关注附近的词元，而另一些则呈现出全局注意力模式。距离最远/最近的 10% 词元上的平均注意力概率分别为 0.068/0.279。该结果表明，FLT 能够学习语言中的局部性偏置，同时保持捕捉全局上下文并利用远距离词元信息的优势。

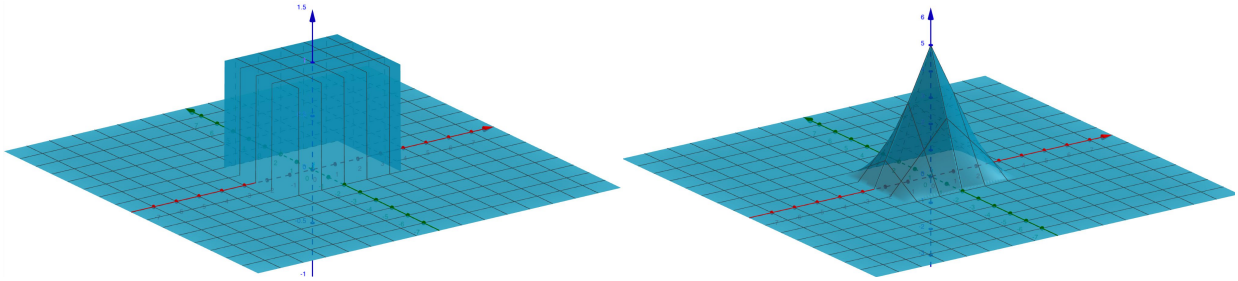


图 4：**第 4.3** 节讨论的局部相对位置编码机制，并通过 FLT 实现支持。两个示例中的词元位置均由两个坐标描述（$\ell = 2$）。 x 和 y 坐标编码差向量 $\Delta\mathbf{r} = (\Delta r_1, \Delta r_2)^\top$。 z坐标给出函数 f 的值。**左：**（不连续）$f_{\mathbf{v},C}(\Delta\mathbf{r}) = C \cdot \mathbb{I}_{\{|\Delta r_1| \leq v_1\}} \mathbb{I}_{\{|\Delta r_2| \leq v_2\}}$，其中某个$\mathbf{v} = (v_1, v_2)^\top$。**右：**（连续） $f_{\mathbf{v}}(\Delta\mathbf{r}) = \mathbb{I}_{\{|\Delta r_1| \leq v_1\}} \mathbb{I}_{\{|\Delta r_2| \leq v_2\}} (-|\Delta r_1| + v_1)(-|\Delta r_2| + v_2)$。这两个局部相对位置编码函数在有界区域之外均为零。

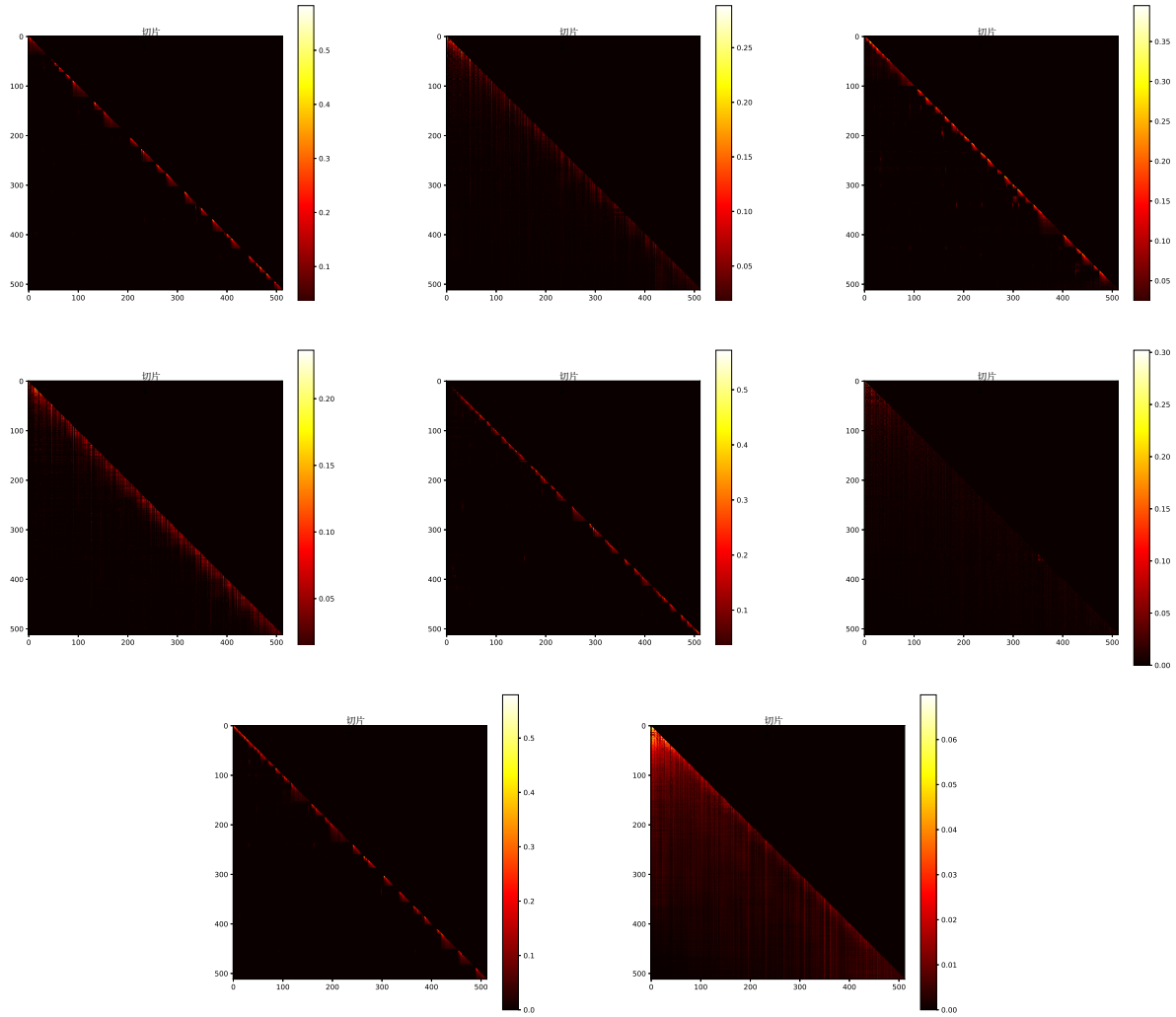


图 5：**注意力矩阵可视化**FLT 第一层中采用局部相对位置编码的 8 个注意力头