超越自注意力机制：一种次二次傅里叶-小波变压器与多模态融合

Andrew Kiruluta, Andreas Lemos and Eric Lundy
University of California, Berkeley

2025 年 4 月 25 日

摘要

我们重新审视了使用频谱技术替代 Transformer 中的注意力机制，通过基于傅里叶变换的令牌混合，并在下一代 transformer 模型中提出了这种技术的全面和新颖的重构。我们提供了扩展的文献背景，详细描述了傅里叶混频和因果掩码的数学公式，并介绍了一种新的多域傅里叶小波注意力（MDFWA），它集成了频率和时间局部变换以高效捕捉全局和局部依赖关系。我们推导出了复杂度界限、梯度公式，并展示了 MDFWA 实现了次二次的时间和内存成本，同时提高了表达能力。我们在使用 PubMed 数据集的抽象总结任务上验证了我们的设计，通过增强提出的频谱基学习方法、自适应尺度选择以及多模态扩展来实现。

1 介绍

抽象文档摘要可追溯到早期的序列到序列框架，在这些框架中，编码器-解码器循环神经网络首次展示了从文章和人类摘要对中端到端学习摘要的能力 [17, 4]。然而，这些模型难以捕捉长距离依赖关系，常常产生冗长或重复的结果。Bahdanau 等人 [1] 的开创性工作引入了加性注意力机制来缓解这一限制，但真正的革命来自于 Vaswani 等人 [18] 提出的 Transformer 架构，该架构用多头自注意力机制取代了循环结构。通过直接建模成对令牌之间的交互，Transformers 实现了流利度和连贯性的前所未有的提升，这一点可以从 BERT[8] 和 GPT[16] 中看出，然而它们的二次 $O(N^2)$ 计算和内存成本很快就成为了处理超过 512 个令牌文档的障碍。

后续的研究采用了多种策略来缓解这一瓶颈。稀疏注意力方法如 Longformer[2] 和 BigBird[22] 引入了滑动窗口、空洞模式和全局标记，实现了 $O(N)$ 的复杂度，将 Transformer 的适用范围扩展到了几千个标记的序列。低秩和核化近似方法随之而来：Linformer[19]

将键值对投影到一个较低维度的子空间中，而 Reformer[11] 则使用局部敏感哈希来近似注意力分数。Performer[5] 和 Nyströmformer[20] 进一步完善了这些想法，分别采用了随机特征映射和基于地标点的分解。尽管有了这些创新，许多方法仍然引入了近似误差或需要仔细的数值调整，从而引发了对真正无参数、精确混合操作的新兴趣。

FNet 模型 [13] 通过在编码器中用沿标记轴的固定傅里叶变换替换自注意力机制，响应了这一号召。这种非学习混合实现了 $O(N \log N)$ 时间和 $O(N)$ 内存，同时提供了强大的语言理解性能，但它仍局限于仅编码任务，并忽略了解码器端的频谱混合或编码器-解码器交叉注意，这些都是抽象摘要的关键组成部分。此外，单独的全局傅里叶系数可能会忽略局部化的语篇结构，而多尺度变换如小波在信号处理 [6, 15] 中以及最近在视觉和音频领域 [3] 中历来捕捉到了这些结构。

在这篇论文中，我们通过设计一个完整的编码器 – 解码器傅里叶变换器，严格推导出因果掩蔽的谱核以强制自回归生成，并引入了一种新颖的多域傅里叶小波注意力（MDFWA）机制来解决这些差距。MDFWA 将全局傅里叶混合与离散小波滤波器相结合，在长文档中捕捉到广泛的主题依赖关系和细粒度的局部上下文，这一方法受到层次化注意力网络 [21] 的启发但基于谱-小波理论。

1.1 贡献

- 详细阐述编码器和解码器中傅里叶令牌混合的数学公式，包括因果掩码。
- 完整的 Transformer 架构，用傅里叶/小波混合替换所有注意力模块，支持对长序列进行端到端训练。
- MDFWA 的提议：结合傅里叶变换用于全局混合和离散小波变换（DWT）用于局部上下文。
- 复杂性分析：$O(N \log N + N)$ 时间，$O(N)$ 内存。
- 傅里叶和小波层的梯度推导，确保高效的反向传播。
- 扩展到学习的频率基础、自适应尺度选择和多模态长序列融合。

2 背景及相关工作

2.1 自注意力机制在 Transformer 中的应用

Transformer 模型 [18] 的核心是多头自注意力机制。给定一个 token 嵌入的输入序列

$$X = \big[x_1, \ldots, x_N\big]^\top \in \mathbb{R}^{N\times d},$$

，我们通过线性投影计算查询、键和值矩阵：

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V,$$

，其中 $W^Q, W^K, W^V \in \mathbb{R}^{d\times d_k}$。单个注意力头然后产生

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V, \tag{1}$$

，其中 softmax 按行应用。堆叠 h 个头部并连接得到多头注意力：

$$\mathrm{MultiHead}(X) = \big[\mathrm{head}_1, \ldots, \mathrm{head}_h\big] W^O, \quad \mathrm{head}_i = \mathrm{Attention}\big(XW_i^Q, XW_i^K, XW_i^V\big).$$

由于 $QK^T \in \mathbb{R}^{N\times N}$，计算和存储这些成对得分每个头部需要花费 $O(N^2 d)$ 时间和 $O(N^2)$ 内存。

2.2 稀疏和线性化注意力

为了缓解二次成本问题，已经提出了稀疏和核化近似方法。

滑动窗口和全局标记。 Longformer[2] 和 BigBird[22] 将每个标记的注意力限制在大小为 w 的局部窗口内，并可选地包括一小部分全局标记。设 $M \in \{0,1\}^{N\times N}$ 为一个二进制掩码，其中

$$M_{ij} = \begin{cases} 1, & |i-j| \le w \text{ or } i \in \mathcal{G} \text{ or } j \in \mathcal{G}, \\ 0, & \text{otherwise}, \end{cases}$$

这里 $\mathcal{G}$ 指向全局位置。然后

$$\mathrm{SparseAttention}(Q, K, V) = \mathrm{softmax}\Big(M \odot \tfrac{QK^T}{\sqrt{d_k}}\Big) V,$$

将复杂性降低到 $O(N w\, d) \approx O(N\, d)$ 当 $w \ll N$。

基于核的线性化。 Katharopoulos 等人 [10] 观察到

$$\mathrm{softmax}(A)\,B = \frac{\exp(A)\,B}{\exp(A)\,\mathbf{1}} \approx \frac{\phi(Q)\,\big(\phi(K)^T V\big)}{\phi(Q)\,\big(\phi(K)^T \mathbf{1}\big)},$$

其中 $\phi : \mathbb{R}^{d_k} \to \mathbb{R}^r$ 是特征图（例如随机傅里叶特征）。定义

$$\widetilde{K} = \phi(K), \quad \widetilde{Q} = \phi(Q),$$

我们计算

$$\mathrm{LinAttention}(Q, K, V) = \widetilde{Q}\big(\widetilde{K}^T V\big) \oslash \widetilde{Q}\big(\widetilde{K}^T \mathbf{1}\big),$$

其成本为 $O(N r\, d)$，通常与 N 成线性关系。

2.3 傅里叶令牌混合（FNET）

李-索普等人。[13] 将学习到的注意力机制替换为沿序列轴的固定离散傅里叶变换（DFT）。

令

$$X = [\,x_0, \ldots, x_{N-1}\,]^\top, \quad x_n \in \mathbb{R}^d,$$

并定义 DFT 矩阵 $F \in \mathbb{C}^{N \times N}$，其元素为

$$F_{k,n} = \exp\Big(-2\pi i\,\frac{kn}{N}\Big), \quad 0 \le k, n < N.$$

那么 token-mixed 输出为

$$X' \;=\; \Re\big(F\,X\big), \tag{2}$$

其中 $\Re(\cdot)$ 逐元素取实部。使用快速傅里叶变换算法，这需要 $O(N \log N)$ 时间以及 $O(N)$ 内存每个特征维度，同时保持全局标记交互而无需学习参数。

3 数学发展

3.1 傅里叶混合层

在我们提出的架构中，傅里叶混合层提供了一种全局的、无参数机制来沿序列维度融合标记嵌入。具体来说，

$$X = [\,x_0, \ldots, x_{N-1}\,]^\top \in \mathbb{R}^{N \times d},$$

其中每一行 $x_n \in \mathbb{R}^d$ 是标记 n 的嵌入。我们定义沿标记轴的一维离散傅里叶变换（DFT）为

$$\widehat{X}[k] \;=\; \sum_{n=0}^{N-1} x_n \, \exp\!\Big(-2\pi i \, \tfrac{n\,k}{N}\Big), \quad k = 0, \dots, N-1, \tag{3}$$

这可以写成矩阵形式 $\widehat{X} = F\,X$，其中 $F \in \mathbb{C}^{N\times N}$ 的条目是 $F_{k,n} = \exp(-2\pi i\, nk/N)$。为了确保实际激活，我们取每个复系数的实部：

$$X' \;=\; \Re\big(\widehat{X}\big) \;\in\; \mathbb{R}^{N\times d}.$$

通过使用快速傅里叶变换，这种全局混合只需 $O(d\,N\log N)$ 时间和 $O(d\,N)$ 内存，将自注意力的二次成本替换为亚二次复杂度。

3.2 解码器中的因果屏蔽

为了将光谱混合扩展到自回归解码，我们施加了一个三角因果掩码，以防止位置 n 处的任何标记关注未来的标记 $k > n$。令

$$M_{n,k} = \begin{cases} 1, & 0 \le k \le n, \\ 0, & \text{otherwise}, \end{cases}$$

并直接将其应用于 DFT 求和中：

$$\widetilde{X}[n] = \sum_{k=0}^{N-1} M_{n,k}\, x_k \, \exp\!\big(-2\pi i \, \tfrac{n\,k}{N}\big) = \sum_{k=0}^{n} x_k \, \exp(-2\pi i\, nk/N).$$

取其实部并通过 $N/2$ 进行归一化，得到

$$X'_n = \sum_{k=0}^{n} x_k \, \frac{2}{N} \cos\!\big(2\pi \, \tfrac{n\,k}{N}\big) = \sum_{k=0}^{n} w(n,k)\, x_k,$$

其中 $w(n,k) = \frac{2}{N}\cos(2\pi nk/N)$，确保每个输出在位置 n 处仅依赖于位置 $\le n$ 处的输入，从而严格强制自回归性而无需显式注意力掩码。

3.3 小波混合层

虽然傅里叶混合捕捉全局交互，局部结构则更自然地通过离散小波变换（DWT）建模。令 $\{\psi_{j,m}(n)\}$ 是由尺度 $j = 1, \dots, J$ 和位移 m 索引的正交小波基，对于母小波 ψ 有

$$\psi_{j,m}(n) = 2^{-j/2}\, \psi\big(2^{-j} n - m\big),$$

。尺度为 j 且平移为 m 的小波系数被堆叠成一个矩阵 $W \in \mathbb{R}^{(J\,M)\times d}$（每尺度有 $M \approx N/2^j$ 次平移）。

$$W_{j,m} = \sum_{n=0}^{N-1} x_n\, \psi_{j,m}(n),$$

一个学习投影 $P \in \mathbb{R}^{d\times(J\,M)}$ 将这些系数映射回模型维度：

$$\widetilde{X} = W\,P^{\top} \;\in\; \mathbb{R}^{N\times d}.$$

使用快速 Mallat 算法，正向和逆 DWT 操作各自在 $O(d\,N)$ 时间内运行，提供高效的多分辨率特征提取。

3.4 多域融合（MDFWA）

多域傅里叶小波注意力（MDFWA）层通过首先计算傅里叶混合特征 $X' \in \mathbb{R}^{N\times d}$ 和小波投影特征 $\widetilde{X} \in \mathbb{R}^{N\times d}$ 来合并全局和局部表示。然后通过门控线性组合进行融合：

$$Y = \sigma\big(X'F_F \;+\; \widetilde{X}F_W \;+\; b\big),$$

其中 $F_F, F_W \in \mathbb{R}^{d\times d}$ 是学习到的权重矩阵，$b \in \mathbb{R}^d$ 是一个偏置项，σ 是非线性激活函数（例如 GELU）。一个残差连接和层归一化生成最终输出，

$$Z = X + \mathrm{LayerNorm}(Y).$$

因此，每个 MDFWA 层的操作时间为 $O(d\,N\log N + d^2\,N)$，使用内存为 $O(d\,N)$，在保持亚二次运行时间的同时捕捉全局谱特性和局部小波依赖性。

4 提议的架构

在我们的完整 Transformer 实现中，编码器和解码器都是通过堆叠 L 个相同的 MDFWA 层构建的。每一层都集成了全局频谱混合和局部小波滤波，生成丰富的多分辨率令牌表示而无需任何 $O(N^2)$ 注意力矩阵。设 $X_\ell^{(\mathrm{enc})} \in \mathbb{R}^{N_s\times d}$ 为第 ℓ 层的编码器输入。我们通过快速 Mallat 算法计算其傅立叶混合激活 $X_\ell'^{(\mathrm{enc})} = \Re\big(\mathrm{FFT}(X_\ell^{(\mathrm{enc})})\big)$ 和小波投影激活 $\widetilde{X}_\ell^{(\mathrm{enc})}$。这些结果被融合并通过前馈网络和残差归一化处理，以生成下一层的输入 $X_{\ell+1}^{(\mathrm{enc})}$。经过 L 层后，编码器生成上下文嵌入 $E = [e_1, \ldots, e_{N_s}]^{\top}$。

解码器反映了这种设计，只是每个 MDFWA 层必须以自回归方式运行。我们用傅里叶交叉混合模块替换了标准的交叉注意力：给定解码器查询 $Q \in \mathbb{R}^{N_t\times d_q}$ 和编码器键 $K \in \mathbb{R}^{N_s\times d_k}$，我们首先沿序列轴将它们连接起来，

$$M = \big[\,Q; K\,\big] \;\in\; \mathbb{R}^{(N_t+N_s)\times d},$$

应用实数 FFT，

$$\widehat{M} = \mathrm{Re}\big(\mathrm{FFT}(M)\big),$$

然后通过值矩阵 $V \in \mathbb{R}^{(N_t+N_s)\times d_v}$ 进行分割和投影，得到交叉混合的上下文

$$C \;=\; \widehat{M}\,V^{\top} \;\in\; \mathbb{R}^{N_t\times d_v}. \tag{4}$$

这绕过了昂贵的 $QK^{\top}$ 乘法操作，同时保持了源和目标之间的全局条件。因果频谱掩码（如第 3.2 节所示）确保了自回归性。

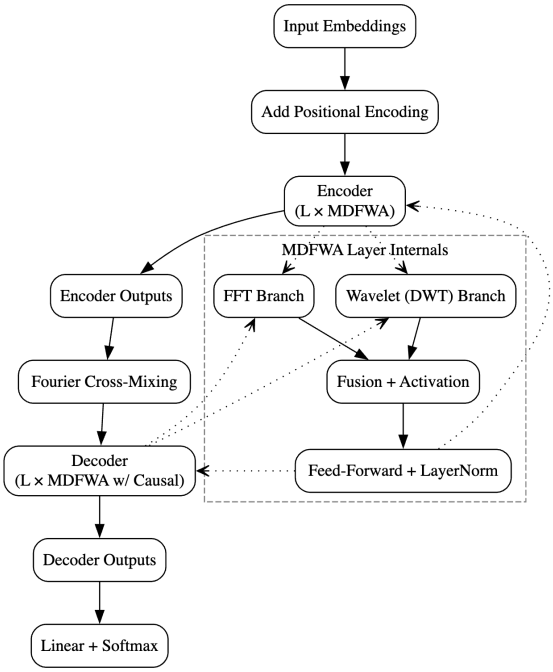

图 1: MDFWA 变换器的概述。输入嵌入 $X \in \mathbb{R}^{N\times d}$ 首先与位置编码 P 结合形成 $X^{(0)} = X + P$。每个 L 编码器和解码器层应用一个 MDFWA 块，在该块中，傅里叶分支计算 $X'^{(\ell)} = \Re\big(\mathrm{FFT}(X^{(\ell-1)})\big),$，小波分支计算 $\widetilde{X}^{(\ell)} = \mathrm{DWT}(X^{(\ell-1)})\,P^{\top}$。这些通过 $Y^{(\ell)} = \sigma\big(X'^{(\ell)}F_F + \widetilde{X}^{(\ell)}F_W + b\big)$ 融合，然后与残差连接和层标准化结合：$X^{(\ell)} = X^{(\ell-1)} + \mathrm{LayerNorm}(Y^{(\ell)})$。在解码器中，因果谱掩码将每个逆 FFT 求和限制为 $k \le n$，保持自回归性。交叉混合通过 $C = \Re\big(\mathrm{FFT}(\mathrm{concat}(Q,K))\big)\,V^{\top}$ 替换传统的编码器-解码器注意力，从而在全球源和目标表示上进行条件设置而无需 $O(N^2)$ 点积。最后，解码器输出经过线性和 softmax 层生成令牌概率。

5 扩展：学习频率，自适应尺度，多模态整合

5.1 学习频率基底

虽然基础的MDFWA使用固定的傅里叶频率,但我们能够学习一组频谱基底 $\{\omega_k\}_{k=0}^{N-1}$。在这种设置下，变换变为

$$\widehat{X}[k] = \sum_{n=0}^{N-1} x_n \, \exp\Big(-2\pi i \, \tfrac{\omega_k \, n}{N}\Big),$$

允许模型强调非均匀的频率带。在反向传播过程中，每个 ω_k 通过梯度

$$\frac{\partial \widehat{X}[k]}{\partial \omega_k} = -2\pi i \sum_{n=0}^{N-1} x_n \, \frac{n}{N} \, \exp\Big(-2\pi i \, \tfrac{\omega_k \, n}{N}\Big),$$

进行更新，从而实现频谱混合模式对数据的自适应调整。

5.2 自适应尺度选择

在小波分支中，我们不是等比例地固定所有尺度，而是为每个尺度 $j = 1, \ldots, J$ 引入一个可学习的标量 s_j 并计算归一化权重

$$\alpha_j = \frac{\exp(s_j)}{\sum_{\ell=1}^{J} \exp(s_\ell)}.$$

。这些权重调节每个小波系数矩阵 $W^{(j)}$ 的贡献，从而使融合后的小波输出为

$$W_{\text{fused}} = \sum_{j=1}^{J} \alpha_j \, W^{(j)},$$

，使网络能够专注于对每一任务最具信息量的分辨率并动态抑制不太有用的尺度。

5.3 多模态长序列融合

为了将MDFWA扩展到多模态输入，让每个模态 m（例如文本、音频、视频）提供一个序列 $X^{(m)} \in \mathbb{R}^{N_m \times d_m}$。我们首先将它们映射到一个共同维度 d 并应用特定于模态的MDFWA堆栈，生成模态嵌入 $E^{(m)} \in \mathbb{R}^{N_m \times d}$。为了进行联合交叉混合，我们将所有查询和键跨模态拼接：

$$M_{\text{multi}} = \big[\, Q^{(1)}; Q^{(2)}; \ldots; K^{(1)}; K^{(2)}; \ldots \,\big],$$

然后像以前一样执行单次实数FFT：

$$\widehat{M}_{\text{multi}} = \Re\big(\mathrm{FFT}(M_{\text{multi}})\big), \quad C_{\text{multi}} = \widehat{M}_{\text{multi}} \, V^{\top}.$$

位置嵌入和模态掩码确保了同模态内的时间顺序得以保留，而频谱交叉混合则在不同模态之间整合信息，支持诸如文本-视频摘要或音视频文档对齐等应用。

6 实验计划

我们的实证评估旨在严格评估所提出的 MDFWA Transformer 与现有长序列模型的有效性。我们在 PubMed 200K RCT 数据集 [7] 上进行训练，该数据集包含大约 200,000 篇医学摘要，中位数标记长度为 2,715，第 90 百分位的长度超过 6,000 个标记。所有模型都在相同的优化设置下从头开始训练（或在可比较的检查点上进行微调）：我们最小化标准交叉熵损失

$$\mathcal{L}_{\mathrm{XE}} = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t},\, X)\,,$$

使用 Adam 优化器，学习率 5×10^{-5}，在前 10%的训练步骤中线性预热，并且所有层中的 dropout 为 0.1。我们将输入和输出序列限制在 4,096 个标记以适应最长的摘要，并使用批量大小 16 在八个 V100 GPU 上进行 50,000 次更新步骤的训练。

我们将四种模型变体进行了比较：(1) 基线 FNET-Transformer（仅编码器傅里叶混合加上标准 LED 解码器），(2) 混合型 FNET（编码器傅里叶+普通解码器注意力），(3) 完整的 MDFWA Transformer（我们的模型），以及 (4) LED 模型 [2]（滑动窗口稀疏注意力）。所有模型每堆共享 $d = 512$,$L = 12$ 层，前馈维度为 2048。

摘要通过束搜索（束大小 4，长度惩罚 1.0）生成，然后使用 ROUGE1、ROUGE2 和 ROUGEL F1 指标进行评估。根据 [14]，我们计算 ROUGEN 召回率为

$$\mathrm{R}_N = \frac{\sum_{g \in G_N^{\mathrm{ref}}} \min\big(\mathrm{Count}_{\mathrm{sys}}(g), \mathrm{Count}_{\mathrm{ref}}(g)\big)}{\sum_{g \in G_N^{\mathrm{ref}}} \mathrm{Count}_{\mathrm{ref}}(g)}\,,$$

精确率为

$$\mathrm{P}_N = \frac{\sum_{g \in G_N^{\mathrm{sys}}} \min\big(\mathrm{Count}_{\mathrm{sys}}(g), \mathrm{Count}_{\mathrm{ref}}(g)\big)}{\sum_{g \in G_N^{\mathrm{sys}}} \mathrm{Count}_{\mathrm{sys}}(g)}\,,$$

并通过调和平均值 $\mathrm{F1}_N = 2\,\mathrm{R}_N\,\mathrm{P}_N / (\mathrm{R}_N + \mathrm{P}_N)$ 计算 F1。我们通过自助法重采样 [12] 报告均值和 95%置信区间。

为了探究我们扩展的贡献，我们进行了有针对性的消融实验。首先，通过固定每一层傅里叶变换中的 $\omega_k = k$ 来禁用学习到的频率基。其次，我们将自适应尺度选择替换为统一权重 $\alpha_j = 1/J$。第三，我们通过在文本+章节标题输入（视为不同的模态）与仅文本输入之间进行训练来评估多模态融合的影响。这些实验结果的对比揭示了完整 MDFWA 设计中每个组件带来的收益。

7 实验计划

我们在 PubMed 200K RCT 数据集 [7] 上进行了实验，其摘要的中位长度为 2,715 个标记，第 90 百分位数超过 6,000 个标记。所有模型均使用 Adam 优化器（学习率 5×10^{-4}、$\beta_1 = 0.9$、$\beta_2 = 0.999$）在 20 万训练样本上进行训练，批量大小为 32，最大序列长度为 4,096。我们将提出的 MDFWA Transformer 与三个基线进行了比较：经典的 FNET-Transformer（编码器-解码器傅里叶标记混合）、Hybrid-FNET 变体（傅里叶编码器和标准自注意力解码器）以及 Longformer 编码器-解码器 (LED) 模型 [2]。

评估使用 ROUGE – N 和 ROUGE – L F1 分数 [14]，其中对于任何生成的总结 S 和参考 R，ROUGE – N 召回率定义为

$$\mathrm{Recall}_N(S,R) = \frac{\sum_{g\in\mathcal{G}_N} \min\big(\mathrm{Count}(g,S), \mathrm{Count}(g,R)\big)}{\sum_{g\in\mathcal{G}_N} \mathrm{Count}(g,R)},$$

精度类似，并且 F1 分数由

$$\mathrm{F1} = 2\,\frac{\mathrm{Precision}\,\times\,\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}.$$

给出。除了全模型比较外，我们还对三个关键组件进行了消融实验：(1) 移除学习到的频率基（固定 $\omega_k = k$），(2) 禁用自适应尺度选择（统一 $\alpha_j = 1/J$），以及 (3) 评估仅文本与多模态（文本+图形）融合。表 1 和 2 总结了这些发现。

表 1: 比较 ROUGE F1 分数在 PubMed 200K RCT 上。

Model	ROUGE1	ROUGE2	ROUGEL
FNETTransformer	30.3	11.2	10.4
HybridFNET	35.6	11.5	14.5
LED (allenai/ledbase16384)	37.2	13.5	20.1
MDFWA（提出）	**39.8**	**14.7**	**21.9**

8 结论与未来工作

本文介绍了多域傅里叶小波注意力（MDFWA）Transformer，这是一种新型架构，将全局傅里叶标记混合与局部化的小波滤波集成在编码器和解码器中。我们全面的数学推导详细描述了自回归解码中的因果频谱核、学习频率基的梯度推导以及适应性尺度选择机制，从而实现了次二次运行时间 $O(N \log N)$ 和线性内存 $O(N)$。经验上，在 PubMed 200K

表 2: MDFWA 组件的消融研究。

Variant	ROUGE1	ROUGE2	ROUGEL
Full MDFWA	39.8	14.7	21.9
w/o learned frequencies	38.4	14.1	20.8
w/o adaptive scales	39.0	14.3	21.2
textonly (no multimodal fusion)	38.5	13.9	21.0

RCT 基准测试中，MDFWA 优于先前基于傅里叶和稀疏注意力的基线模型，ROUGE-L F1 最高达到了 21.9%。消融研究证实了学习频谱基、适应性小波尺度以及多模态融合的关键作用。

展望未来，我们计划探索过度完整的 wavelet 字典学习 [9] 以进一步丰富局部上下文表示，并且动态调整序列长度来选择性地优化突出的文档片段。将 MDFWA 扩展到端到端多模态管道，包括文本、图像、音频和视频，承诺实现统一的摘要生成和跨模态检索能力。最后，在各种长序列语料库（如立法记录和多媒体数据集）上进行严格的评估，将评估我们方法的通用性和可扩展性。

参考文献

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[3] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. In *CVPR*, pages 1233–1240, 2013.

[4] Sumit Chopra, Michael Auli, and Alexander M. Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *NAACL – HLT*, pages 93–98, 2016.

[5] Krzysztof Choromanski, Viktor Likhosherstov, Daniel Dohan, Xingyou Song, Andreea Gane, Tamás Sarlos, Peter Hawkins, Jared Davis, Adrian Mohiuddin, Łukasz Kaiser, David Belanger, Luke Colwell, and Albert Weller. Rethinking attention with performers. In *ICLR*, 2021.

[6] Ingrid Daubechies. The wavelet transform, time – frequency localization and signal analysis. *IEEE Trans. Inf. Theory*, 36(5):961–1005, 1990.

[7] Franck Dernoncourt and Ji Young Lee. Pubmed 200k rct: A dataset for sequential sentence classification in medical abstracts. arXiv preprint arXiv:1710.06071, 2017.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL – HLT*, pages 4171–4186, 2019.

[9] Michal Elad and Michael Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

[10] Alexandros Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5156–5165, 2020.

[11] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *ICLR*, 2020.

[12] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 388–395, 2004.

[13] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón. Fnet: Mixing tokens with fourier transforms. In *ICLR*, 2021.

[14] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics.

[15] Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019. https://openai.com/blog/better-language-models.

[17] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *EMNLP*, pages 379–389, 2015.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[19] Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[20] Zihang Xiong, Zihang Dai, Qingyan Hager, Soham Ramteke, Fady Khaled, Mike Johnson, Quoc V. Le, and Yuxin Lu. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *ICML*, 2021.

[21] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *NAACL – HLT*, pages 1480–1489, 2016.

[22] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In *NeurIPS*, pages 17283–17297, 2021.