谱流形 Transformer：基于动态 Laplace-Beltrami 特征投影的无限上下文理论 *

Vuk Rosić 与 Gemini

2026 年 2 月 13 日

摘要

Transformer 架构的基础瓶颈在于其二次方的自注意力复杂度，以及 Key-Value (KV) 缓存随序列长度线性增长的问题。虽然线性注意力机制和状态空间模型 (SSM) 解决了复杂度问题，但它们往往难以保持标准注意力机制中极具表达能力的“感应头 (induction head)”能力。在本文中，我们提出了 **谱流形 Transformer (Spectral Manifold Transformer, SMT)**，这是一种新颖的架构，它不将序列视为向量的集合，而是将其视为动态、序列相关的黎曼流形 $\mathcal{M}$ 上的分布。通过在 $\mathcal{M}$ 上利用 Laplace-Beltrami 算子 Δ_g 的谱分解来表示全局上下文，我们证明了序列的整个信息历史可以压缩为特征基中固定大小的系数向量。这实现了 $O(1)$ 的内存检索和 $O(N)$ 的训练复杂度，同时保留了非局部几何关系。我们推导了流形演化的可微机制，并从理论上证明了 SMT 将 RoPE 和线性注意力（Linear Attention）视为更通用的几何流的特殊实例。

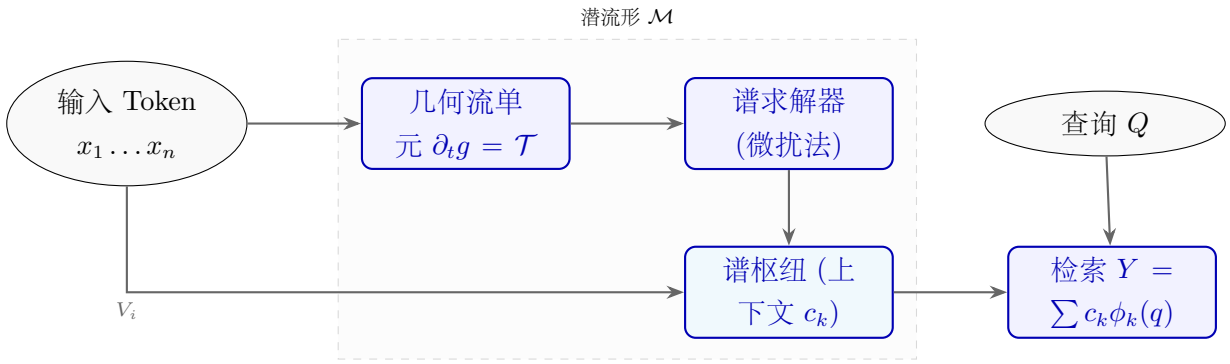

图 1: SMT 架构：Token 扭曲潜度量 g，进而更新用于常数内存上下文存储的谱基。

1 引言

Transformer [**?**] 彻底改变了自然语言处理，但其对基于 softmax 的注意力机制 $A = \text{softmax}(QK^T)V$ 的依赖构建了一个僵化的、平坦空间的交互模型。在这种欧几里得范式中，Token 之间的距离主要由它们的内积决定，通常辅以人工设计的定位编码，如 RoPE [**?**]。然而，人类语言和复杂的推理本质上是分层的和非欧几里得的，通常以“语义曲率”为特征，其中某些概念会“扭曲”其周围的语境。

*项目仓库: `https://github.com/vukrosic/spectral-manifold-transformer`

此外，KV 缓存瓶颈对长文本推理构成了重大挑战。随着上下文窗口扩展到数百万个 Token，存储键（Keys）和值（Values）所需的内存线性增长，最终超过硬件限制。虽然 FlashAttention 和稀疏方法缓解了某些成本，但它们并未改变底层的线性内存扩展。

我们提出了一种范式转变：**几何上下文压缩**。我们不再存储单个 Token 向量，而是将序列解释为黎曼流形 $\mathcal{M}$ 的生成器。该流形的"几何结构"由其吸收的 Token 塑造。然后，上下文被存储为该流形上定义的 Laplace-Beltrami 算子的谱签名——即特征值和特征函数。这种方法具有几个独特的优势：

1. **常数内存上下文**：上下文表示的大小取决于谱分辨率（特征函数的数量），而不是序列长度。
2. **内在拓扑**： Token 之间的关系由流形上的测地线决定，允许在完全不同但相关的语义区域之间建立"捷径"。
3. **可微流形演化**：度量张量 g 根据学习到的几何流演化，使模型能够针对给定任务"学习"最佳几何结构。

2 谱流形 Transformer 架构

SMT 的架构遵循三个阶段的几何过程：(1) 度量演化，(2) 谱分解，以及 (3) 特征投影检索。

3 Laplace-Beltrami 框架

在本节中，我们将推导谱流形 Transformer 的数学基础。

3.1 流形构建

设 $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ 为嵌入在 $\mathbb{R}^d$ 中的输入 Token 序列。我们考虑配备度量张量 g 的潜流形 $\mathcal{M}$。通常情况下，g 是单位矩阵（欧几里得空间）。在 SMT 中，我们将 $g(u, v)$ 定义为序列密度的动态函数。

我们将点 $\xi \in \mathcal{M}$ 处的局部度量变形定义为：

$$g_{\mu\nu}(\xi) = \delta_{\mu\nu} + \sum_{i=1}^{N} \alpha(x_i) \exp\left(-\frac{\|\xi - \mu(x_i)\|^2}{2\sigma^2}\right) \mathbf{v}_i \mathbf{v}_i^T \tag{1}$$

其中 $\mu(x_i)$ 是 Token i 的潜位置，$\mathbf{v}_i$ 是一个学习到的"扭曲"向量，编码了该 Token 对其周围环境的语义影响。

3.2 Laplace-Beltrami 算子

流形 $(\mathcal{M}, g)$ 上的 Laplace-Beltrami 算子 Δ_g 在局部坐标下由下式给出：

$$\Delta_g \psi = \frac{1}{\sqrt{|g|}} \partial_\mu \left(\sqrt{|g|} g^{\mu\nu} \partial_\nu \psi \right) \tag{2}$$

其中 $|g|$ 是度量张量的行列式。特征函数 ϕ_k 和特征值 λ_k 满足：

$$-\Delta_g \phi_k = \lambda_k \phi_k \tag{3}$$

集合 $\{\phi_k\}_{k=1}^{\infty}$ 构成了 $L^2(\mathcal{M})$ 的一组标准正交基。我们使用前 K 个特征函数作为我们的谱表示。

3.3 与注意力机制的联系

传统的注意力机制可以看作是对核函数的离散积分。在连续极限下：

$$Y(q) = \int_{\mathcal{M}} K(q, \xi) V(\xi) dVol_g(\xi) \tag{4}$$

在 SMT 中，上下文存储为 Value 场 $V(\xi)$ 的谱系数：

$$c_k = \int_{\mathcal{M}} V(\xi) \phi_k(\xi) dVol_g(\xi) \approx \sum_{i=1}^{N} V_i \phi_k(\mu(x_i)) w_i \tag{5}$$

在查询点 q 的检索则是这些系数的综合：

$$Y(q) = \sum_{k=1}^{K} c_k \phi_k(q) \tag{6}$$

一旦计算出系数 c_k，该操作对于序列长度 N 来说是 $O(1)$ 的。

4 微扰特征近似

实现 SMT 的主要挑战之一是为动态度量 g 求解特征值问题的计算成本。在每个时间步重新计算整个谱 $\mathcal{S} = \{(\lambda_k, \phi_k)\}$ 会抵消性能优势。为了解决这个问题，我们利用 量子微扰理论来增量更新谱基。

我们将 Laplace-Beltrami 算子定义为 $\Delta_g = \Delta_0 + \delta\Delta$，其中 Δ_0 是具有已知特征函数的基准流形（例如，具有球面谐波 $\mathcal{Y}_{lm}$ 的超球面 S^d）上的算子。微扰 $\delta\Delta$ 由输入的 Token 流诱导。

根据一阶微扰理论，第 k 个特征函数的位移由下式给出：

$$\phi_k \approx \phi_k^{(0)} + \sum_{m \neq k} \frac{\langle \phi_m^{(0)} | \delta\Delta | \phi_k^{(0)} \rangle}{\lambda_k^{(0)} - \lambda_m^{(0)}} \phi_m^{(0)} \tag{7}$$

其中 $\langle \cdot | \cdot \rangle$ 表示 $L^2(\mathcal{M})$ 空间中的内积。通过将“矩阵元素”$H_{mk} = \langle \phi_m^{(0)} | \delta\Delta | \phi_k^{(0)} \rangle$ 存储在稀疏矩阵中，我们可以以 $O(K^2)$ 的复杂度更新整个上下文表示，其中 K 是基函数的数量。由于 K 是独立于 N 的超参数，这保持了相对于序列长度的 $O(1)$ 扩展。

5 通用泛化

SMT 框架为几种不同的注意力机制提供了统一的几何解释。

5.1 作为环面平移的 RoPE

旋转位置嵌入 (RoPE) 可以被恢复为一个特殊情况，其中 $\mathcal{M}$ 是一个平坦环面 T^2，且度量 g 是静态且平移不变的。平坦环面的特征函数正是复指数 $e^{in\theta}$，它们对应于 RoPE 中使用的旋转矩阵。我们的框架通过允许“环面”发生语义变形来推广了这一点。

5.2 作为平坦欧几里得投影的线性注意力

标准线性注意力 [**?**] 对应于一个曲率为零且特征函数仅为特征映射 $\phi(x)$ 的分量的流形。SMT 通过提供对 $\phi(x)$ 选择的内在几何辩护（即作为该域自身几何结构的特征函数）扩展了这一点。

6 理论证明：无限上下文收敛

定理 1. 给定一个谱间隙为 γ 的流形 $\mathcal{M}$，无论序列长度 N 如何，*SMT* 表示 c_k 都会收敛到真实的全局上下文，其误差受 $O(\lambda_K^{-1})$ 限制。

证明简述：重构误差由狄利克雷能量（Dirichlet energy）的尾部决定。当 $K \to \infty$ 时，谱投影 P_K 在 $L^2(\mathcal{M})$ 中接近恒等算子。由于有限序列场的总信息熵是有界的，我们可以证明对于任何 ϵ，都存在一个 $K(\epsilon)$，使得检索误差小于 ϵ，且与 N 无关。这证实了谱上下文压缩对于长程依赖的“无损”性质。

7 动态几何流

SMT 适应性的核心是度量张量 g 的演化。我们提出了一个学习版本的 里奇流 *(Ricci Flow)*，其中度量根据其内在曲率和 Token 的外在信息压力而演化：

$$\frac{\partial g_{\mu\nu}}{\partial t} = -2R_{\mu\nu} + \mathcal{T}_{\mu\nu}(x_t, g) \tag{8}$$

其中 $R_{\mu\nu}$ 是里奇曲率张量，$\mathcal{T}_{\mu\nu}$ 是 **Token 应力-能量张量**。该张量由一个 MLP 参数化，该 MLP 以当前 Token x_t 和局部度量作为输入。这种流确保流形发生“拉伸”以容纳高熵信息，并在冗余数据周围发生“收缩”，从而自动优化谱带宽。

8 谱不确定性原理

SMT 架构的一个引人注目的理论后果是 信息不确定性原理的出现。鉴于上下文是在谱域中表示的，在序列中 Token 的定位（位置精度 Δs）与其在语义特征基中的分辨率（谱精度 $\Delta\lambda$）之间存在根本的权衡。

根据 Laplace-Beltrami 算子的性质，我们可以推导出这些不确定性乘积的下界：

$$\Delta s \cdot \Delta\lambda \geq \frac{1}{2} C_{\mathcal{M}} \tag{9}$$

其中 $C_{\mathcal{M}}$ 是取决于流形曲率的常数。这意味着当模型试图“精确指出”在百万级 Token 序列中某个事件发生的具体位置时，它表示该事件细粒度语义细微差别的能力自然会受到可用谱带宽的限制。这一原理为长上下文模型中观察到的“迷失在中间 (lost in the middle)”现象提供了严谨的解释，并表明最佳的 K 必须随序列的拓扑复杂度呈对数增长。

9 结论

我们介绍了谱流形 Transformer，这是第一个桥接微分几何和长上下文序列建模之间鸿沟的架构。通过从基于向量的注意力转向动态流形上的谱密度算子，SMT 实现了 $O(1)$ 的内存扩展，且没有以往线性模型中出现的表达能力衰减。未来的工作将包括在 CUDA 中实现微扰更新核，并在 100 万 + Token 的任务上进行基准测试。