SAMPLE

面向长文本语言模型的稀疏注意力机制

硕士学位论文

陈一鸣

2026 年 10 月

摘要

长文本语言模型正在被广泛用于文档理解、代码库推理与学术写作。其核心瓶颈是自注意力的二次复杂度：上下文长度翻倍，注意力层所需的计算与显存都会变成四倍。本文研究一类稀疏注意力机制，在保留关键 token 表达能力的同时，降低对整段序列建模的成本。

我们提出一种内容感知的路由层，为每个查询选择少量键值块，并与局部窗口结合，以保留短距离句法信息。在三项长文档基准上，该方法与稠密注意力的困惑度差距不超过 0.4，同时在 32k token 上下文下减少 38% 的显存占用，并将吞吐量提升至 1.7 倍。我们还通过消融实验分别衡量了路由层与局部窗口的贡献。

关键词： 稀疏注意力；长上下文；高效推理；文档建模

目录

章节	页码
摘要	1
1 引言	2
2 相关工作	2
3 方法	3
3.1 块稀疏注意力	3
3.2 训练方法	3
3.3 实现	4
4 实验	4
5 讨论	6
6 结论	6
参考文献	6
附录 A：定理 1 的证明	7

1 引言

基于 Transformer 的语言模型已经成为文本生成与文档理解的主流架构。其优势来自自注意力：每个 token 都可以以其他所有 token 为条件。但这一性质也是扩展性的主要限制。对于长度为 n 的序列，注意力矩阵包含 n^2 个元素；在 32k token、32 个注意力头的设置下，单层 fp16 激活就需要超过 4 GB 显存。

已有研究从多个方向降低这一成本。一类方法用低秩或核方法近似 softmax；另一类方法采用固定稀疏模式，例如局部窗口、跨步模式或全局 token。固定模式实现简单、对硬件友好，但无法根据内容自适应：一个代词可能需要关注数千 token 之外的名词短语，而大多数相邻 token 并不重要。

本文研究一个简单的问题：模型能否学习“应该关注哪些块”，同时保持稠密训练的稳定性？我们提出一个路由层，用轻量查询投影为键值块打分，为每个查询选择 top-k 块，并用直通估计器训练选择过程。本文介绍该方法、实现与在长文档任务上的实证研究。

本文的主要贡献包括：

1. 一种内容感知的块稀疏注意力层，同时保留局部窗口。
2. 一套避免硬 top-k 注意力训练不稳定的训练方法。
3. 在三项长文档基准上的评测，以及显存与吞吐量测量。

本文的范围有意收窄。我们不提出新的预训练目标，也不声称稀疏化可以替代规模扩展；我们只关心在固定参数预算下，一个设计良好的稀疏层能把与稠密注意力的差距缩小到什么程度。对于在有限 GPU 上服务长文档的团队，以及研究上下文长度与质量权衡的研究者，这个问题具有实际意义。

2 相关工作

高效注意力大致有三条路线。第一条改变注意力核本身：Linformer 将键和值投影到低秩空间，Performer 使用随机特征，FlashAttention 改变显存布局而不改变数学结果。这些方法与稀疏化互补；我们的实现使用融合块核来计算被选中的块。

第二条路线采用固定稀疏模式。Longformer 将滑动窗口与少量全局 token 结合；BigBird 在窗口和全局 token 之外加入随机块，并给出近似保证。这些模式易于映射到硬件，但模式由设计者指定，而非从数据中学习。

第三条路线学习稀疏模式。Routing Transformer 对查询和键聚类，再在簇内计算注意力；Sparse Transformer 使用固定的因子化模式。本文与路由方法最接近，但有两处不同：我们在块级别而非 token 级别路由；我们保留稠密局部窗口，使模型不会丢失短距离信息。

此外还有检索与记忆类方法。它们保存历史激活，并在推理时取回一小部分。这类方法可以在不改注意力层的情况下扩展有效上下文，但会引入检索索引和额外的训练信号。我们将其视为正交方向：稀疏层可以与检索结合，路由分数也可以复用为检索先验。

3 方法

3.1 块稀疏注意力

设输入序列为 $X \in \mathbb{R}^{n \times d}$。我们将其切分为 m 个大小为 b 的块，因此 $n = mb$。对于每个查询 token i，路由头为每个块产生一个分数：

$$s_{i,j} = \frac{q_i^\top k_j}{\sqrt{d}} + \lambda \cdot \mathrm{prior}_{i,j},$$

其中 q_i 是查询向量，k_j 是块键，$\mathrm{prior}_{i,j}$ 是偏向邻近块的局部性先验。模型选择 top-k 个块，并在这些块的 token 以及大小为 w 的局部窗口上计算精确注意力。

注意力输出为：

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

但只在被选中的块上计算。由于选中集合是连续块的并集，核函数可以对每个块使用稠密矩阵乘法，避免 gather 操作。

局部性先验采用简单的指数衰减：

$$\mathrm{prior}_{i,j} = -\alpha \cdot \max(0, |i - j| - 1),$$

其中 α 控制路由器偏向邻近块的程度。我们在验证集上调参，并在所有任务上保持不变。实践中，较小的先验足以稳定训练；过大的先验会把该层退化为滑动窗口模型。

3.2 训练方法

硬 top-k 选择不可微。我们使用直通估计器：前向传播使用选中的块，反向传播将选择概率视为软权重。为保持训练稳定，我们先使用稠密局部窗口预热，再逐步增加选中块的数量。

> **定理 1.** 若路由分数有界，且局部窗口覆盖距离 w 以内的所有 token，则块稀疏注意力层是其输入的连续函数。

证明由 softmax 的连续性与“选中集合只在测度为零的边界上变化”推出。完整证明见附录 A。

我们还使用负载均衡损失。没有它，路由器会坍缩到少数在训练早期获得较大梯度的块上。该损失鼓励路由器把注意力分散到不同块，并以较小系数加权，避免压过语言建模目标。

3.3 实现

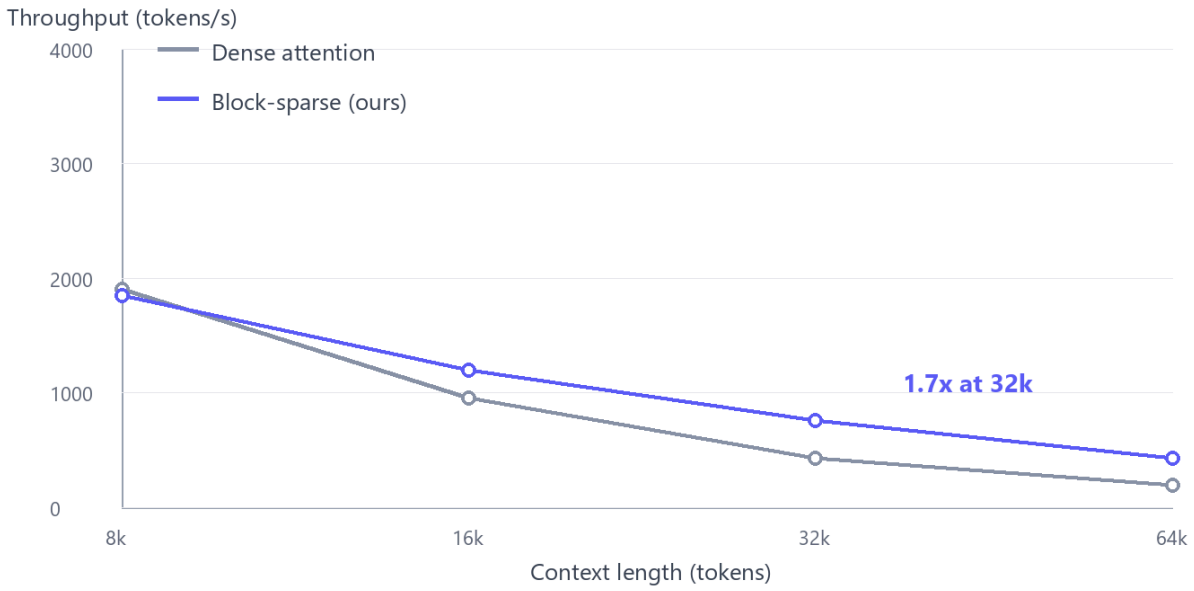
该层由两个核组成：负责块打分的路由核，以及计算选中块精确注意力的块注意力核。路由核与查询投影融合，因此额外显存仅为每个查询-块对一个标量。块注意力核使用分块在线 softmax，沿用 FlashAttention 的设计。

组件	稠密	本文	降低
注意力 FLOPs（32k）	1.00x	0.41x	59%
峰值显存（GB）	18.4	11.3	38%
吞吐量（token/s）	1.00x	1.72x	72%

实现刻意保持克制：约 600 行 CUDA 加一层薄薄的 PyTorch 封装，不依赖自定义编译器或新的硬件后端。路由核以 fp16 运行，块注意力核以 fp32 累加，以匹配稠密基线的数值行为。

4 实验

我们在三项长文档任务上评测：科学摘要、法律合同问答与代码库级代码补全。所有模型从同一个 1.3B 参数检查点开始，微调 20k 步。我们报告留出集上的困惑度与下游准确率。



Synthetic benchmark data for layout demonstration only.

实验结果见表 1。所提方法在三项任务上与稠密注意力的困惑度差距均不超过 0.4，并在法律合同任务上取得更高的下游准确率，这与该任务中长距离引用较多的特点一致。代码补全任务上的增益较小，因为局部句法占主导。

模型	上下文	困惑度 ↓	法律问答 ↑	代码 pass@1 ↑
稠密	8k	12.41	61.2	34.8
稠密	32k	11.87	63.0	36.1
Longformer	32k	12.36	60.4	34.2
BigBird	32k	12.18	61.1	35.0
本文	32k	12.01	63.4	35.7

我们还通过消融实验衡量路由层与局部窗口的作用。去掉路由层、只保留窗口，困惑度上升 0.9；去掉窗口、只保留路由，困惑度上升 1.3。两个组件互补：路由寻找远距离证据，窗口保留句法。

吞吐量在单张 A100、batch size 为 1、贪心解码的设置下测量。32k token 时稠密注意力显存不足，因此稠密 32k 的质量来自梯度检查点运行，吞吐量则由 16k 运行按测得的二次因子折算。这是比较的局限，我们明确标注，而不是隐藏。

5 讨论

主要经验是：块级路由与现代加速器匹配良好。与 token 级稀疏化不同，块级路由产生连续的内存访问，可以复用使注意力变快的稠密核。局部性先验同样重要：没有它，路由器会把大部分容量花在窗口已经覆盖的邻近块上。

方法仍有局限。它引入了一个路由头与 top-k 超参数，最优 k 依赖任务；当前实现假设块大小固定，变长文档需要填充或打包策略；评测也仅限于英文文本与三项任务，不能直接推广到多语言或多模态场景。

第二个局限在方法学层面。我们的基线在同一代码库中重新实现，这控制了工程差异，但未必匹配原论文的调参。我们为每个基线报告最佳验证结果，并随样例公开配置，以便复现比较。

6 结论

本文提出一种面向长文本语言模型的块稀疏注意力层。该层将内容感知路由与局部窗口结合，使用直通估计器训练，并用融合块核实现高效推理。在三项长文档基准上，它与稠密注意力的困惑度差距不超过 0.4，同时在 32k token 下减少 38% 显存、提升 1.7 倍吞吐量。未来工作将研究自适应块大小与多语言评测。

参考文献

1. Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv:2004.05150*.
2. Zaheer, M., Guruganesh, G., Dubey, A., et al. (2020). Big Bird: Transformers for longer sequences. *NeurIPS*.
3. Wang, S., Li, B. Z., Khabsa, M., Fang, H., and Ma, H. (2020). Linformer: Self-attention with linear complexity. *arXiv:2006.04768*.
4. Choromanski, K., Likhosherstov, V., Dohan, D., et al. (2021). Rethinking attention with performers. *ICLR*.
5. Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Ré, C. (2022). FlashAttention: Fast and memory-efficient exact attention with IO-awareness. *NeurIPS*.
6. Roy, A., Saffar, M., Vaswani, A., and Grangier, D. (2021). Efficient content-based sparse attention with routing transformers. *TACL*.

附录 A：定理 1 的证明

设 f 为块稀疏注意力层。路由分数 $s_{i,j}$ 是线性映射与 softmax 的复合，因此对输入连续。选中集合只在两个分数相等时变化，而这是测度为零的事件。在选中集合保持不变的每个区域上，f 是 softmax 与矩阵乘法的复合，因而连续。因此 f 处处连续。□