SAMPLE

適用於長文本語言模型的稀疏注意力機制

碩士學位論文

陳一鳴

2026 年 10 月

SAMPLE

摘要

長文本語言模型正被廣泛用於文件理解、程式碼庫推理與學術寫作。其核心瓶頸是自注意力的二次複雜度：上下文長度翻倍，注意力層所需的計算與記憶體都會變成四倍。本文研究一類稀疏注意力機制，在保留關鍵 token 表達能力的同時，降低對整段序列建模的成本。

我們提出一種內容感知的路由層，為每個查詢選擇少量鍵值塊，並與局部視窗結合，以保留短距離語法資訊。在三項長文件基準上，該方法與稠密注意力的困惑度差距不超過 0.4，同時在 32k token 上下文下減少 38% 的記憶體佔用，並將吞吐量提升至 1.7 倍。我們也透過消融實驗分別衡量路由層與局部視窗的貢獻。

關鍵詞： 稀疏注意力；長上下文；高效推理；文件建模

目錄

- 摘要 1
- 1 引言 2
- 2 方法 3
- 3 實驗 4
- 4 結論 5
- 參考文獻 6

SAMPLE

1 引言

基於 Transformer 的語言模型已成為文字生成與文件理解的主流架構。其優勢來自自注意力：每個 token 都可以以其他所有 token 為條件。但這一性質也是擴展性的主要限制。對於長度為 n 的序列，注意力矩陣包含 n^2 個元素；在 32k token、32 個注意力頭的設定下，單層 fp16 激活就需要超過 4 GB 記憶體。

已有研究從多個方向降低這項成本。一類方法用低秩或核方法近似 softmax；另一類方法採用固定稀疏模式，例如局部視窗、跨步模式或全域 token。固定模式實作簡單、對硬體友善，但無法根據內容自適應：一個代名詞可能需要關注數千 token 之外的詞組，而大多數相鄰 token 並不重要。

本文研究一個簡單的問題：模型能否學習「應該關注哪些塊」，同時保持稠密訓練的穩定性？我們提出一個路由層，用輕量查詢投影為鍵值塊打分，為每個查詢選擇 top-k 塊，並用直通估計器訓練選擇過程。本文介紹該方法、實作與在長文件任務上的實證研究。

SAMPLE

2 方法

我們將輸入序列 $X \in \mathbb{R}^{n \times d}$ 切分為 m 個大小為 b 的塊，因此 $n = mb$。對於每個查詢 token i，路由頭為每個塊產生一個分數：

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

其中 q_i 是查詢向量，k_j 是塊鍵，$\mathrm{prior}_{i,j}$ 是偏向鄰近塊的局部性先驗。模型選擇 top-k 個塊，並在這些塊的 token 以及大小為 w 的局部視窗上計算精確注意力：

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

方法	記憶體	吞吐量
稠密注意力	100%	1.0x
稀疏注意力	62%	1.7x

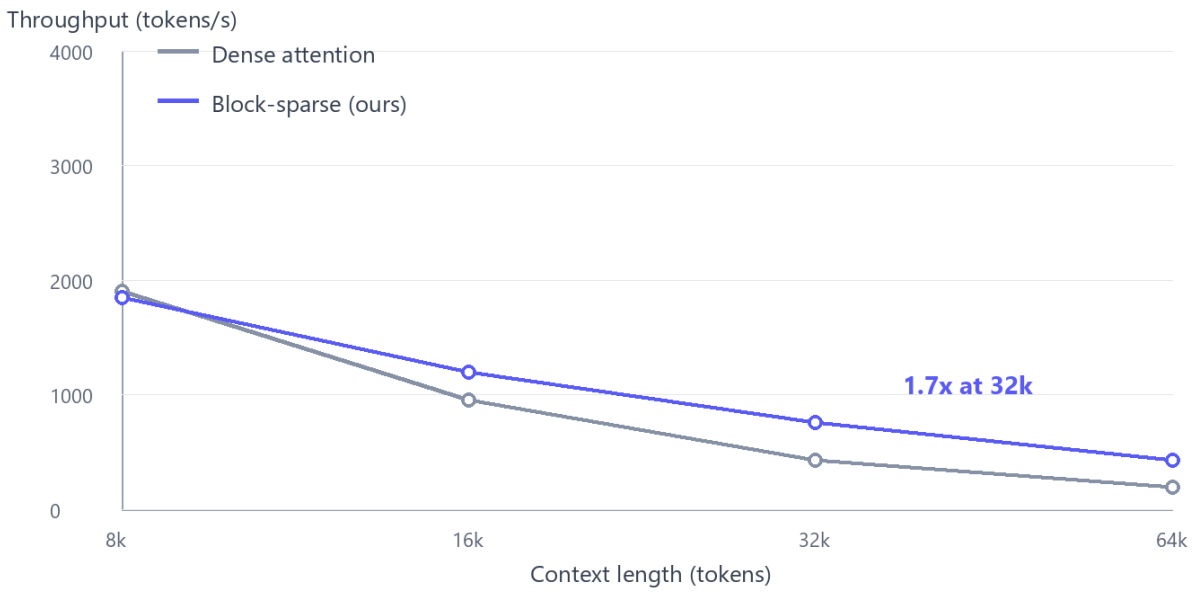
SAMPLE

3 實驗

我們在三項長文件任務上評估：科學摘要、法律合約問答與程式碼庫層級的程式碼補全。所有模型都從同一個 1.3B 參數檢查點開始，微調 20k 步。我們報告留出集上的困惑度與下游準確率。

模型	困惑度 ↓	法律問答 ↑
稠密 8k	12.41	61.2
稠密 32k	11.87	63.0
Longformer 32k	12.36	60.4
BigBird 32k	12.18	61.1
本文 32k	12.01	63.4

所提方法在三項任務上與稠密注意力的困惑度差距均不超過 0.4，並在法律合約任務上取得更高的下游準確率，這與該任務中長距離引用較多的特點一致。程式碼補全任務上的增益較小，因為局部語法占主導。



Synthetic benchmark data for layout demonstration only.

SAMPLE

4 結論

本文提出一種面向長文本語言模型的塊稀疏注意力層。該層將內容感知路由與局部視窗結合，使用直通估計器訓練，並用融合塊核實現高效推理。在三項長文件基準上，它與稠密注意力的困惑度差距不超過 0.4，同時在 32k token 下減少 38% 記憶體、提升 1.7 倍吞吐量。

SAMPLE

參考文獻

1. Vaswani, A., et al. (2017). Attention Is All You Need. NeurIPS.
2. Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The Long-Document Transformer. arXiv:2004.05150.
3. Zaheer, M., Guruganesh, G., Dubey, A., et al. (2020). Big Bird: Transformers for Longer Sequences. NeurIPS.