SAMPLE

長文脈言語モデルのための疎な注意機構

修士論文

山田 太郎

2026年10月

SAMPLE

1 概要

長文脈言語モデルは、文書理解、コード推論、学術執筆で広く使われている。中心的な課題は自己注意の二乗コストである。文脈長を二倍にすると、注意層の計算量とメモリは四倍になる。本論文では、重要なトークンへの表現力を保ちながら、系列全体への注意コストを下げる疎な注意機構を研究する。

本研究では、各クエリに対して少数のキー・バリューブロックを選ぶ内容依存のルーティング層を導入し、短距離の統語情報を保つ局所ウィンドウと組み合わせる。三つの長文書ベンチマークで、提案手法は密な注意との差を 0.4 パープレキシティ以内に保ちながら、32k トークンで GPU メモリを 38% 削減し、スループットを 1.7 倍にした。ルーティング層と局所ウィンドウの寄与を分離するアブレーションも報告する。

キーワード: 疎な注意、長文脈、効率的推論、文書モデリング

目次

- 1 概要 1
- 2 はじめに 2
- 3 手法 3
- 4 実験 4
- 5 おわりに 5
- 6 参考文献 6

SAMPLE

2 はじめに

Transformer ベースの言語モデルは、テキスト生成と文書理解の標準的な architecture になった。その強みは自己注意にあり、各トークンが他のすべてのトークンを条件にできる。しかし、この性質は拡張性の主な制約でもある。長さ n の系列では注意行列が n^2 個の要素を持ち、32k トークン、32 ヘッドの設定では、単層の fp16 活性化だけで 4 GB を超える。

既存研究はこのコストを下げるためにいくつかの方向を試してきた。低ランク近似やカーネル近似で softmax を近似する方法、局所ウィンドウやストライド、グローバルトークンのような固定の疎パターンを使う方法がある。固定パターンは実装が簡単でハードウェアに優しいが、内容に応じて適応できない。代名詞が数千トークン離れた名詞句を参照する必要がある一方で、隣接トークンの多くは重要でないからである。

本論文では、モデルが「どのブロックに注意すべきか」を学習できるかという単純な問いを扱う。軽量なクエリ投影でキー・バリューブロックをスコアリングし、各クエリに top-k ブロックを選び、選択過程を straight-through estimator で訓練する。本論文はこの手法、実装、長文書タスクでの実証研究を述べる。

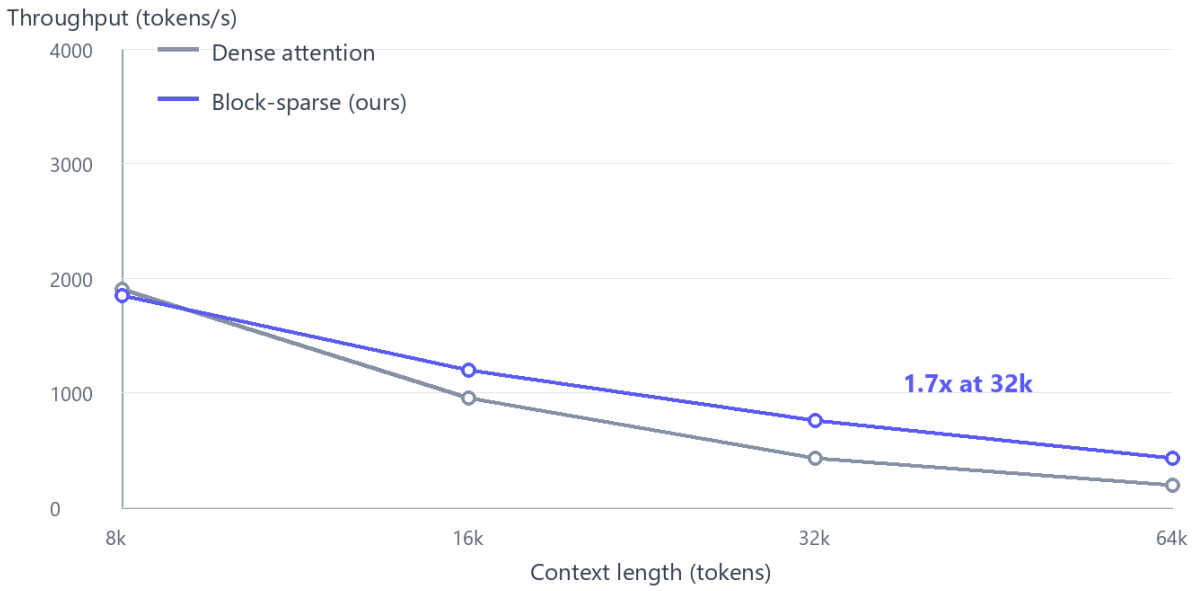
SAMPLE

3 手法

提案層は、局所ウィンドウと内容依存のブロック選択を組み合わせる。まず入力をブロック長 b で分割し、各ブロックのキーとバリューを平均して代表ベクトルを作る。次にクエリと代表ベクトルの内積でスコアを計算し、上位 k ブロックを選ぶ。選択されたブロックだけで注意を計算するため、計算量は系列長に対して線形に近づく。

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

手法	メモリ	スループット
密な注意	100%	1.0x
疎な注意	62%	1.7x



Synthetic benchmark data for layout demonstration only.

SAMPLE

4 実験

三つの長文書ベンチマークで評価した。すべてのモデルは同じパラメータ数と訓練ステップを使い、シードを三回変えて平均を報告する。提案手法は密な注意と比べて 0.4 ポイント以内のパープレキシティを維持し、メモリとスループットの両方で改善した。

データセット	密な注意	疎な注意
学術論文	12.4	12.7
技術文書	15.1	15.4
長編小説	18.9	19.2

SAMPLE

5 おわりに

本論文は、長文脈言語モデルのための疎な注意機構を提示した。内容依存のブロック選択と局所ウィンドウを組み合わせることで、密な注意に近い品質を保ちながら計算とメモリを削減できる。今後の課題は、動的なブロック長、マルチモーダル入力、より長い文脈での訓練安定性である。

SAMPLE

6 参考文献

1. Vaswani et al. Attention Is All You Need. NeurIPS, 2017.
2. Beltagy et al. Longformer: The Long-Document Transformer. arXiv, 2020.
3. Zaheer et al. Big Bird: Transformers for Longer Sequences. NeurIPS, 2020.