Theta-CS：コーシー=シュワルツ不等式による前段スクリーニング

高槻高等学校　宮﨑俊輔

1. 導入

[2]TransformerのAttention機構は全トークン間の関係を内積計算で求めるため高精度だが、計算コストが膨大である。
[1]Top-θ は各headに静的閾値を設けて不要なAttention Scoreを刈ることで、推論時の計算量を削減する手法である。さらに補正によって精度を維持しつつ訓練を要さない疎化を実現する。本研究では、コーシー＝シュワルツ不等式を用いた安全な前段ふるいを導入し、閾値疎化の前に不要な内積計算を削減する。

2. 方法論

2.1,2.2は本研究の前提となるアーキテクチャである。

2.1 Attention機構 [2]

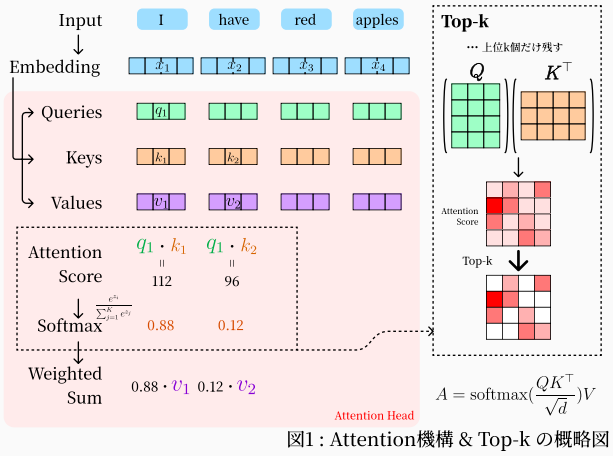
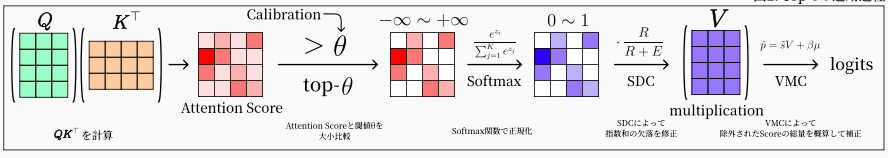


図1 : Attention機構 & Top-k の概略図

2.2 Top-θ [1]

… 事前計算されたk個の要素を保持する静的閾値θを使用することで、Top-kとほぼ同一の集合を保持する。

図2: Top-θの適用過程



・SDC (Softmax Denominator Compensation)
softmax関数の分母から省略された"指数の合計"を修正する

$$\mathrm{softmax}(a)_i = \mathrm{softmax}(\tilde{a})_i \cdot \frac{R}{R+E}$$

・VMC (V-Mean-Compensation)
除外されたScoreの総量を概算し、不足部を埋めて補正する

$$\mu = \frac{1}{n}\sum_{i=0}^{n-1} V_i \ , \quad \beta = 1 - \sum_{i=0}^{k-1} \tilde{s}_i \quad \rightarrow \quad \hat{p} = \tilde{s}\boldsymbol{V} + \beta\mu$$

Value行の平均　　残ったScoreの合計　　最終出力 (logits)

Good:
・訓練不要。短時間のしきい値校正だけで使える
・計算量とKVキャッシュの読み取りを削減可能

Bad:
・除外されたScoreの内積計算が無駄になる

Algorithm 1 Calibrate($\mathcal{C}, k, \alpha$) - 1 head threshold

Require: $\mathcal{C}$ (Calibration set of inputs)
Require: $k \in \mathbb{N}$ (elements to keep per attention row)
Require: $\alpha \in \mathbb{R}$ (calibration offset in std _devs)
$\Theta_r \leftarrow \emptyset, \forall r$ {empty sets of observed thresholds}
for $\boldsymbol{X} \in \mathcal{C}$ **do**
  if is_prefill($\boldsymbol{X}$) **then**
    $\boldsymbol{A} = Q(\boldsymbol{X})K^\dagger(\boldsymbol{X})$
    $n = \mathrm{num\_rows}(\boldsymbol{A})$
    for $r = k$ to $n-1$ **do**
      $\Theta_r = \Theta_r \cup \{\mathrm{quantile}_{\frac{k}{r}}(\boldsymbol{A}_r)\}$
      $\boldsymbol{A}_r = \mathrm{top}_k(\boldsymbol{A}_r)$
    end for
    $\boldsymbol{S} = \mathrm{row\_softmax}(\boldsymbol{A})$
  else {generative decoding, $X \in \mathbb{R}^d$}
    $\boldsymbol{a} = Q(\boldsymbol{X})K^\dagger(\boldsymbol{X})$
    $n = \mathrm{length}(\boldsymbol{a})$
    $\Theta_n = \Theta_n \cup \{\mathrm{quantile}_{\frac{k}{n}}(\boldsymbol{a})\}$
    $\boldsymbol{a} = \mathrm{top}_k(\boldsymbol{a})$
    $\boldsymbol{s} = \mathrm{softmax}(\boldsymbol{a})$
  end if
end for
return $\theta_r = \mathrm{mean}(\Theta_r) + \alpha \cdot \mathrm{std\_dev}(\Theta_r), \forall r$

[1]　図3: 閾値θを求める疑似コード

2.3 Theta-CS

… QK内積を計算する前にコーシー=シュワルツ不等式に従って明らかに小さい"あり得ない候補"を排除する

コーシー=シュワルツ不等式

$$|\langle x, y\rangle| \le \|x\|_2 \|y\|_2$$

→

行列にして、

$$\|QK^\top\|_2 \le \|Q\|_2 \|K\|_2$$

ゆえに、

$$\|Q\|_2\|K\|_2 \le \theta \Rightarrow \|QK^\top\|_2 \le \theta$$

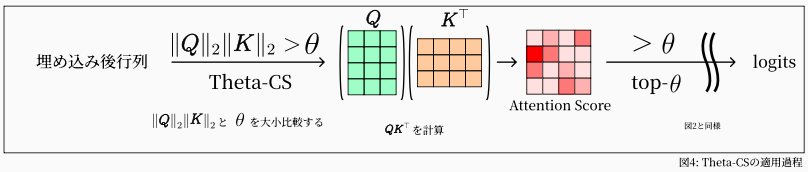


図4: Theta-CSの適用過程

ノルム

$\vec{x} = (x_1, x_2, \cdots, x_n)$

$$\|\vec{x}\|_p = \sqrt[p]{|x_1|^p + |x_2|^p + \cdots + |x_n|^p}$$

Good:
・内積計算の前に候補を剪定することで計算量を削減
・実装が容易
・上限値に基づいてスクリーニングするため安全

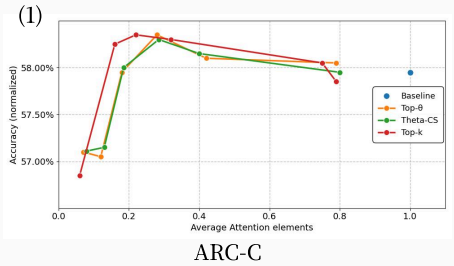
Algorithm 2 Theta-CS ProFilter (pre-softmax, 1 head) with Top-θ thresholds

Require: $\{\theta_r\}_{r\ge 1}$ **(pre-softmax thresholds from Top-θ calibration)**
Require: $\delta \in \mathbb{R}_{\ge 0}$ **(safety margin)**
$\boldsymbol{k\_norms} \leftarrow$ empty vector {cache of $\|k_i\|$ aligned with KV cache}
for $\boldsymbol{X} \in \mathcal{C}$ **or streaming inputs do**
  if is_prefill($\boldsymbol{X}$) **then**
    $(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \leftarrow \mathrm{project}(\boldsymbol{X})$ {shape: $[n, d]$ each}
    $\boldsymbol{k\_norms} \leftarrow \|\boldsymbol{K}\|_{\mathrm{row}\ 2}$ {one-time compute for current prompt; append if continuing}
    $n = \mathrm{rows}(\boldsymbol{Q})$
    for $r = 1$ to n **do**
      $q \leftarrow \boldsymbol{Q}_r, \quad q\_n \leftarrow \|q\|$
      $\theta \leftarrow \theta_r, \quad \tau \leftarrow \theta - \delta$
      $\boldsymbol{m}_r \leftarrow (q\_n \cdot \boldsymbol{k\_norms} \ge \tau)$ {Cauchy–Schwarz screen}
      if allfalse($\boldsymbol{m}_r$) **then**
        $\boldsymbol{m}_r[\arg\max(\boldsymbol{k\_norms})] \leftarrow$ true
      end if
      $\boldsymbol{K}^{(\mathrm{keep})} \leftarrow \boldsymbol{K}[\boldsymbol{m}_r], \quad \boldsymbol{V}^{(\mathrm{keep})} \leftarrow \boldsymbol{V}[\boldsymbol{m}_r]$
      $\boldsymbol{a}^{(\mathrm{keep})} \leftarrow \boldsymbol{K}^{(\mathrm{keep})} q^\top$ {compute dot only on kept keys}
      $\boldsymbol{a}^{(\theta)} \leftarrow \mathrm{threshold\_pre}(\boldsymbol{a}^{(\mathrm{keep})}, \theta)$ {Top-θ pre-softmax step}
      $(\tilde{\boldsymbol{s}}, \mathrm{SDC\_state}) \leftarrow \mathrm{apply\_SDC\_and\_softmax}(\boldsymbol{a}^{(\theta)}, \theta)$
      $\hat{y}_r \leftarrow \mathrm{apply\_VMC}(\tilde{\boldsymbol{s}}, \boldsymbol{V}^{(\mathrm{keep})}, \mathrm{SDC\_state})$
    end for
  else
    {generative decoding; new token $x \in \mathbb{R}^d$}
    $(q, k, v) \leftarrow \mathrm{project}(x)$
    append k, v to KV cache; append $\|k\|$ to $\boldsymbol{k\_norms}$
    $q\_n \leftarrow \|q\|, \quad n \leftarrow \mathrm{length}(\boldsymbol{k\_norms})$
    $\theta \leftarrow \theta_n, \quad \tau \leftarrow \theta - \delta$
    $\boldsymbol{m} \leftarrow (q\_n \cdot \boldsymbol{k\_norms} \ge \tau)$
    if allfalse($\boldsymbol{m}$) **then**
      $\boldsymbol{m}[\arg\max(\boldsymbol{k\_norms})] \leftarrow$ true
    end if
    $\boldsymbol{K}^{(\mathrm{keep})} \leftarrow \boldsymbol{K}[\boldsymbol{m}], \quad \boldsymbol{V}^{(\mathrm{keep})} \leftarrow \boldsymbol{V}[\boldsymbol{m}]$
    $\boldsymbol{a}^{(\mathrm{keep})} \leftarrow \boldsymbol{K}^{(\mathrm{keep})} q^\top$
    $\boldsymbol{a}^{(\theta)} \leftarrow \mathrm{threshold\_pre}(\boldsymbol{a}^{(\mathrm{keep})}, \theta)$
    $(\tilde{\boldsymbol{s}}, \mathrm{SDC\_state}) \leftarrow \mathrm{apply\_SDC\_and\_softmax}(\boldsymbol{a}^{(\theta)}, \theta)$
    $\hat{y} \leftarrow \mathrm{apply\_VMC}(\tilde{\boldsymbol{s}}, \boldsymbol{V}^{(\mathrm{keep})}, \mathrm{SDC\_state})$
  end if
end for
return $\{\hat{y}_r\}$ or $\hat{y}$ **(same interface as Top-θ after pre-softmax)**

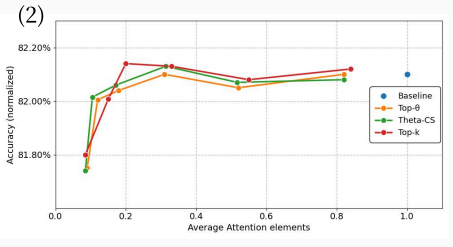
図5: θ-CSの擬似コード

3. 評価実験

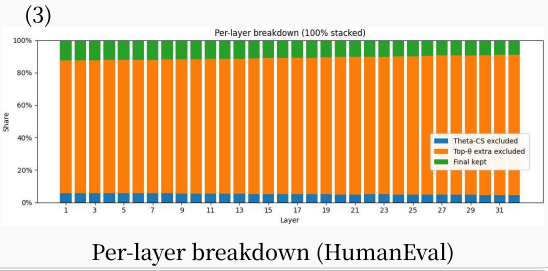
設定:　モデル ：LLaMA3-8B-Instruct
ベンチマーク： ARC-C(prefill), HellaSwag(prefill), HumanEval(generative)



ARC-C



HellaSwag



Per-layer breakdown (HumanEval)

4. 考察

(1)(2)から、
・Theta-CS と Top-θ の精度がほとんど重なっていて安全性がある。

(3)から、
・先行除外率は5%程度であり、$L^2 \times L^2$のため高次元では上界が緩くなりすぎるためだろう。

5. 今後の展望

・今回はコーシー=シュワルツ不等式に則り $L^2 \times L^2$ で上界を設定したが、下にあるようにコーシー=シュワルツ不等式を一般化したヘルダー不等式などまだまだ発展の余地があり、今後も数学的視点からの計算量削減に関する検証を進める。

ヘルダー不等式

$1 \le p, q \le \infty,\ 1/p + 1/q = 1$

$$|\langle x, y\rangle| \le \|x\|_p \|y\|_q$$

→

$p = 1, q = \infty$ で、

$$|QK^\top| \le \|Q\|_1 \|K\|_\infty$$

ゆえに、

$$\|Q\|_1\|K\|_\infty \le \theta \Rightarrow |QK^\top| \le \theta$$

参考文献
[1] "Top-Theta Attention"(Renzo Andri et al., 2025)
[2] "Attention Is All You Need"(Ashish Vaswani et al., 2017)
[3] https://jalammar.github.io/illustrated-transformer/
[4] https://mathlandscape.com/holder/
[5] https://mathlandscape.com/cauchy-schwarz/
[6] https://manabitimes.jp/math/640