arXiv:2304.04397v1 [cs.DS] 10 Apr 2023

过参数化特征维度下的随机与确定性注意力稀疏化算法*

Yichuan Deng† Sridhar Mahadevan‡ Zhao Song§

摘要

大型语言模型（LLM）已在不同领域展现出强大的能力。作为 LLM 的一项重要子程序，注意力计算也引起了理论研究的兴趣。近期，[Alman 和 Song 2023] 以及 [Brand、Song 和 Zhou 2023]分别从算法和困难性角度研究了注意力矩阵的静态计算与动态维护。在这项工作中，我们考虑注意力问题的稀疏化。我们作出一项简化假设，即 logit 矩阵是对称的。令 n 表示句子长度，令 d 表示嵌入维度。给定一个矩阵 $X \in \mathbb{R}^{n \times d}$，假设 $d \gg n$ 和$\|XX^\top\|_\infty < r$，其中 $r \in \in (0, 0.1)$，那么我们的目标是找到 $Y \in \mathbb{R}^{n \times m}$（其中 $m \ll d$），使得

$$\|D(Y)^{-1}\exp(YY^\top) - D(X)^{-1}\exp(XX^\top)\|_\infty \leq O(r)$$

针对这个问题，我们给出了两个结果。

- 我们的第一个结果是一种随机化算法。它的运行时间为 $\widetilde{O}(\mathrm{nnz}(X) + n^\omega)$，成功概率为 $1-\delta$，并选取 $m = O(n\log(n/\delta))$。这里，$\mathrm{nnz}(X)$ 表示 X 中非零项的数量。我们用 ω 表示矩阵乘法的指数。目前 $\omega \approx 2.373$。~
- 我们的第二个结果是一种确定性算法。它的运行时间为 $O(\min\{\sum_{i\in[d]} \mathrm{nnz}(X_i)^2, dn^{\omega-1}\} + n^{\omega+1})$，并选取 $m = O(n)$。这里，X_i 表示矩阵 X 的第 i 列。

我们的主要发现对应用于大语言模型的任务具有以下启示：对于任意超大的特征维度，我们都可以将其缩减到几乎与句子长度呈线性关系的规模。

*作者感谢 Lichen Zhang 提供的大力帮助，并提出了分析中主要的误差控制思路。

†ycdeng@cs.washington.edu。华盛顿大学。
‡smahadev@adobe.com.Adobe研究.
§zsong@adobe.com Adobe研究.

1 引言

注意力机制已成为许多自然语言处理（NLP）应用和大语言模型（LLM）中的重要工具。例如，Transformer[VSP+17]、GPT-1 [RNS+18]、BERT [DCLT18], GPT-2 [RWC+19]、GPT-3 [BMR+20]、PaLM [CND+22] 和 OPT [ZRG+22] 等大语言模型显著推进了该领域的技术水平。这些模型依赖注意力机制来捕捉序列中不同词元之间的依赖关系和关联，使其非常适合语言建模 [MMS+19]、机器翻译[HWL21], 和情感分析 [UAS+20]等任务。近期的一项突破性成果 ChatGPT 是由 OpenAI 基于 GPT-3 构建的聊天机器人，展示了注意力机制的强大能力 [Cha22]。就在最近，OpenAI 发布了关于 GPT-4 [Ope23], 的技术报告；在许多现实任务中，其性能显著优于 GPT 的先前版本。

在 LLM 中，注意力机制是一个基础组件 [VSP+17、RNS+18、DCLT18、RWC+19、BMR+20]。注意力矩阵通常使用一个注意力矩阵$A \in \mathbb{R}^{n \times n}$来表示，其中 n 是输入序列的长度。注意力矩阵的每个元素 $A_{i,\ j}$ 表示分配给第 j个输入元素（对应第 i 个输出词元）的注意力分数或权重。注意力分数通常通过将 softmax函数应用于查询-键点积矩阵 $QK^\top$来计算，其中 Q、$K \in \mathbb{R}^{n \times d}$ 分别是查询矩阵和键矩阵，而d 是输入嵌入的维度。（这里我们关注 $d \gg n$ 的情况。）更准确地说，注意力分数可以计算为 $A = \exp(QK^\top)$。随后，注意力机制的输出通过对输入序列元素进行加权求和来计算，其中注意力分数充当权重。形式上，注意力矩阵的计算定义为

定义 1.1（注意力矩阵计算）。设 Q、 $K \in \mathbb{R}^{n \times d}$ 为两个矩阵。对于矩阵$V \in \mathbb{R}^{n \times d}$，注意力矩阵定义为

$$\mathrm{Att}(Q, K, V) := D^{-1}AV,$$

其中，矩阵 $A \in \mathbb{R}^{n \times n}$ 和 $D \in \mathbb{R}^{n \times n}$ 定义为

$$A := \exp(QK^\top) \quad and \quad D := \mathrm{diag}(A\mathbf{1}_n).$$

注意力计算是 LLM 模型的主要瓶颈 [VSP+17、RNS+18、DCLT18、RWC+19、BMR+20、WLK+20、KKL20]。虽然注意力机制在广泛的 NLP 任务中展现出卓越的性能，但计算注意力矩阵的成本可能高得令人难以承受，尤其是在特征维度较大时。维度较大时，直接计算矩阵的成本会很高。由于 A 的形状为 $n \times n$，直接计算 $D^{-1}AV$至少需要 $\Omega(n^2)$ 的时间。然而，在某些特定的精度保证下，对注意力计算进行近似就足以满足大多数 LLM 的需求 [CGRS19、KKL20、WLK+20、DKOD20、KVPF20、CDW+21、CDL+21]。近期的一项工作 [AS23] 研究了何时可以在不显式计算 A 的情况下计算 $\mathrm{Att}(Q, K, V)$（只能访问矩阵 Q、 K、 V）。他们 [AS23]给出了两个结果：从正面来看，他们证明，如果条目满足某种性质，就存在一种算法，可以在 $n^{1+o(1)}$ 时间内计算注意力。从负面来看，他们证明，如果条目不满足该性质，那么除非强指数时间假设不成立，否则不存在能在真正次二次时间内将注意力计算到误差不超过 $1/\mathrm{poly}(n)$ 的算法。在 [BSZ23], 的工作中，他们将静态注意力计算推广为动态注意力维护问题。如果带提示的MV猜想成立，他们给出了一种条件最优的注意力维护动态算法。

[1]其困难性结果基于 OMV 猜想的一种变体[HKNS15，LW17，CKL18] ，称为带提示的MV猜想 [BNS19]。

受以往关于注意力的研究及其应用启发，我们研究如下问题，

Is 对于较大的特征维度，是否可以更快地计算注意力矩阵?

为应对这一挑战，我们寻求一种快速计算注意力矩阵的算法。在本工作中，我们通过假设 $X = Q = K$ （忽略 V 的影响），聚焦于一种特定形式的注意力矩阵计算：

定义 1.2（对称注意力近似）。设 $X \in \mathbb{R}^{n \times d}$ 是一个矩阵，目标是近似

$$\mathsf{D}^{-1}(X)\exp(XX^\top),$$

其中 $\mathsf{D}(X) := \mathrm{diag}(\exp(XX^\top)\mathbf{1}_n)$。

对于这一特定的在 c 设置下，我们在本文中对上述问题给出了肯定的回答 .

1.1 我们的结果

在这项工作中，我们提出了一种以计算高效的方式近似对称注意力矩阵的算法。该算法以一个大型特征矩阵 X 作为输入，其维度为$n \times d$，并满足约束 $\|XX^\top\|_\infty \leq r$。随后，它输出一个压缩矩阵 $Y \in \mathbb{R}^{n \times m}$，其具有$m = O(n \log(n/\delta))$，其中 $\delta \in (0,1)$是失败概率，用户在实践中可以自行选择。该算法以 X 的近乎稀疏时间运行，并输出对 X的对称注意力的近似。

受 [DSW22], 的方法启发，我们正式证明：

定理 1.3 （随机化算法，定理 3.5）。 设 $r \in$ （0，0。1）。给定一个矩阵 $X \in \mathbb{R}^{n \times d}$，满足 $d \gg n$ 且 $\|XX^\top\|_\infty \leq r$，存在一种随机化算法，其运行时间为

$$\widetilde{O}(\mathrm{nnz}(X) + n^\omega)$$

并输出一个矩阵 $Y \in \mathbb{R}^{n \times m}$，带有 $m = O(n\log(n/\delta))$，使得

$$\|\mathsf{D}(Y)^{-1}\exp(YY^\top) - \mathsf{D}^{-1}(X)\exp(XX^\top)\|_\infty \leq O(r)$$

以概率 $1 - \delta$ 成立。

此处 $\mathsf{D}(X) := \mathrm{diag}(\exp(XX^\top)\mathbf{1}_n)$。我们使用 ω 表示矩阵乘法的指数。目前 $\omega \approx 2.373$。

紧随 Brand 的线性规划去随机化算法发展趋势，以及 Brand 的突破性成果 [Bra20]。在这项工作中，我们还考虑了如何设计一种用于注意力矩阵的确定性稀疏化算法。受另一种方法 [SXZ22、Zha22] 所提出的确定性稀疏化方法启发，我们提出了另一种确定性算法，如下所示。

定理 1.4 (确定性算法，定理 3.6 的非正式版本)。 设 $r \in (0,0.1)$。给定一个矩阵 $X \in \mathbb{R}^{n \times d}$ 满足 $d \gg n$ 且 $\|XX^\top\|_\infty \leq r$，存在一个运行时间为

$$\widetilde{O}(\min\{\sum_{i \in [d]} \mathrm{nnz}(X_i)^2, dn^{\omega-1}\} + n^{\omega+1})$$

并输出一个矩阵 $Y \in \mathbb{R}^{n \times m}$，其具有 $m = O(n)$，使得

$$\|\mathsf{D}(Y)^{-1}\exp(YY^\top) - \mathsf{D}^{-1}(X)\exp(XX^\top)\|_\infty \leq O(r).$$

与 [Bra20] 匹配 [CLS19], 的线性规划算法的结果不同，我们的确定性算法比随机化算法慢得多。我们不确定是否可以设计出一种能够达到我们的随机化算法运行时间的确定性算法。我们将此作为未来工作的一个开放问题。

1.2 相关工作

注意力计算理论 一系列研究专注于注意力计算[ZHDK23、AS23、BSZ23、LSZ23]。相关研究利用局部敏感哈希（LSH）技术进行注意力近似 [ZHDK23]。研究者发现，核密度估计（KDE）问题可用于简化 softmax 函数的分母，并据此提出了 KDEformer [ZHDK23]，该方法采用高效的核密度估计求解器，因此能够在次二次时间内近似注意力，并具有可证明的谱范数界。此外，近期研究还探索了静态和动态两种版本的注意力计算 [AS23、BSZ23]。文献 [LSZ23] 研究了正则化双曲回归问题，例如指数函数、双曲余弦函数和双曲正弦函数。

输入稀疏时间算法 关于输入稀疏时间运行的算法，已有悠久的研究历史 [CW13, NN13, RSW16, SWZ17, ALS+18, SWZ19, SYYZ22b, DSW22, LSZ23]。 [CW13] 提出了首个用于线性回归和低秩近似的输入稀疏时间算法。给定任意输入矩阵 A，稀疏嵌入矩阵 S 与 A的乘积可以在 nnz(A) 时间内计算。后续工作 [NN13] 提出了无偏稀疏范数近似投影（OSNAP），改进了 [CW13]中的维度。许多工作 [RSW16, SWZ17] 将弗罗贝尼乌斯范数低秩近似推广到了其他范数。近期工作 [SYYZ22b]提出了一种更快的算法，可以在输入稀疏时间内近似约翰椭球。另一项近期工作 [DSW22] 提出了一种算法，使其能够针对任意实矩阵，在该矩阵的输入稀疏时间内求解差异最小化问题。

草图与采样 草图与采样技术在数值线性代数中非常强大，已应用于许多基础任务，例如线性规划（LP）[CLS19, JSWZ21, Ye20, DLY21, GS22], 切平面法 [JLSW20], 经验风险最小化 [LSZ19, QSZZ23], 积分最小化 [JLSZ23], 半定规划 [SYYZ22a], 矩阵感知[DLS23], 联邦学习 [SWYZ22], Frank-Wolfe 方法 [SXYZ22], 矩阵补全 [GSYZ23], 动态稀疏化器 [DJS+22]、差分隐私 [EMM+23]、聚类 [EMZ19, CAMZ22, EMMZ22], 张量问题 [RSZ22, SYYZ23], 过参数化神经网络的训练 [SZZ21, Zha22]。

路线图。 本文结构安排如下。在第 2 节中，我们介绍本文的预备知识，包括记号、代数事实以及草图矩阵的定义。在第 3 节中，我们分析算法的正确性。在第 4 节中，我们介绍稀疏化工具。在第 5 节中，我们分析所使用的杠杆分数采样工具。

致谢。作者谨向 Josh Alman、Jan van den Brand、Beidi Chen、Yeqi Gao、Zhihang Li、Junze Yin、Lichen Zhang 和 Tianyi Zhou 表示感谢，感谢他们提供了非常有帮助的讨论。

2 预备知识

在第 2.1 节中，我们给出了全文使用的记号。在第 2.2 节中，我们陈述了一些众所周知的事实。在第 2.3 节中，我们介绍了一些草图矩阵。

2.1 记号

对于任意正整数，我们用 $[n]$ 表示集合 $\{1, 2, \cdots, n\}$。对于矩阵 $A \in \mathbb{R}^{n \times n}$，若 $\forall x \in \mathbb{R}^n$，则称 A 为半正定矩阵（PSD），此时有 $x^\top A x \geq 0$。对于任意矩阵 $A \in \mathbb{R}^{n \times n}$，我们用 $\|A\|_F$ 表示其弗罗贝尼乌斯范数，用 $\|A\|_1$ 表示其逐元素 ℓ_1 范数，即 $\|A\|_F := (\sum_{i \in [n]} \sum_{j \in [n]} |a_{i,j}|^2)^{1/2}$，以及 $\|A\|_1 := \sum_{i \in [n]} \sum_{j \in [n]} |a_{i,j}|$。我们用 $\|A\|_\infty$ 表示 $\max_{i,j} |A_{i,j}|$。

对于矩阵 A，我们用 $\mathrm{nnz}(A)$ 表示矩阵 A 中非零元素的数量。对于矩阵 A，我们用 $\|A\|$ 表示其算子范数/谱范数，即 $\|A\| := \max_x \|Ax\|_2 / \|x\|_2$。注意，$\|A\|_2$ 毫无意义，这种记号不应出现在草稿中。对于半正定矩阵 A，令 $U \Sigma U^\top$ 表示其奇异值分解（SVD）。我们用 $A^{1/2}$ 表示 $U \Sigma^{1/2} U^\top$。

$\exp(x)$ 的泰勒展开为 $\sum_{i=0}^\infty \frac{x^i}{i!} = 1 + x + \frac{x^2}{2!} + \cdots$。

我们用 ω 表示矩阵乘法的指数，目前$\omega \approx 2.373$ [Wil12、LG14、AW21]。

2.2 基本事实

事实 2.1. • 对于任意 $x \in [-0.1, 0.1]$, $|1 - \exp(x)| \leq 2|x|$

证明。根据泰勒展开，对于 $x \in [-0.1, 0.1]$，我们有

$$\begin{aligned} |1 - \exp(x)| &= |\sum_{i=1}^\infty \frac{x^i}{i!}| \\ &\leq |x| + |\sum_{i=2}^\infty \frac{x^i}{i!}| \\ &\leq |x| + |\sum_{i=2}^\infty (\frac{x^2}{2})^i| \\ &= |x| + |\frac{x}{2 - 2x}| \\ &\leq 2|x|, \end{aligned}$$

其中，第 1 步由 $\exp(x)$ 的泰勒展开得出，第 2 步由三角不等式得出，第 3 步由利用 $i > 2$ 得到 $i!$ 的下界 2^i 得出，第 4 步由对第二项求和得出，最后一步由 $|x| \leq 0.1$ 得出。

□

[{"id":0,"output":"事实 2.2。设 $B \in \mathbb{R}^{n \times n}$ 表示一个半正定矩阵。则有"}]

$$B_{i,i} B_{j,j} \geq B_{i,j}^2, \quad \forall i, j \in [n] \times [n]$$

证明。我们选择 $x = a \cdot e_i + b \cdot e_j$。由于对于任何 a、$b \in \mathbb{R}$，因此我们有

$$\begin{aligned} 0 &\leq x^\top B x \\ &= \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} B_{i,i} & B_{i,j} \\ B_{j,i} & B_{j,j} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \end{aligned}$$

由于对于任意 a, b，上述等式均成立，因此

$$\det(\begin{bmatrix} B_{i,i} & B_{i,j} \\ B_{j,i} & B_{j,j} \end{bmatrix}) \geq 0$$

这意味着 $B_{i,\ i}B_{j,\ j} \geq B_{i,\ j}^2$。

□

事实2.3. 对于任意 $A \in \mathbb{R}^{n \times n}$ 和 $B \in \mathbb{R}^{n \times n}$，如果满足以下条件

- $(1-\epsilon)A \preceq B \preceq (1+\epsilon)A$

那么我们有

$$(1-\epsilon)A_{i,i} \preceq B_{i,i} \preceq (1+\epsilon)A_{i,i}, \quad \forall i \in [n].$$

证明。对于每个 $i \in [n]$，我们用 $e_i \in \mathbb{R}^n$ 表示长度为n 的向量，其中第 i 个条目为 1，其他所有条目均为 0。

我们选择 $x = e_i$。然后根据

$$(1-\epsilon)A \preceq B \preceq (1+\epsilon)A,$$

我们有

$$(1-\epsilon)e_i^\top A e_i \preceq e_i^\top B e_i \preceq (1+\epsilon)e_i^\top A e_i,$$

由此完成证明。

□

2.3 草图矩阵

在此我们正式定义几种草图矩阵。

定义 2.4（稀疏嵌入矩阵 I [NN13]）。我们称 $R \in \mathbb{R}^{b \times n}$ 是参数为 s的稀疏嵌入矩阵，如果每一列恰好有 s个非零元素，这些元素 $\pm 1/\sqrt{s}$ 均匀随机选取，其位置均匀随机地无放回选取（且各列之间相互独立）。

定义 2.5（随机高斯矩阵或高斯变换，民间常识）。设 $S = \sigma \cdot G \in \mathbb{R}^{s \times m}$，其中 σ 是一个标量，且 $G \in \mathbb{R}^{s \times m}$ 的每个元素均独立地取自标准高斯分布。对于任意矩阵 $A \in \mathbb{R}^{m \times n}$，$SA$ 可在 $O(s \cdot \mathrm{nnz}(A))$ 时间内计算。

定义 2.6（AMS [AMS96]）。设 h_1、h_2、...、h_b 是从 4-独立哈希族 $\mathcal{H} = \{h : [n] \to \{-\frac{1}{\sqrt{b}}, +\frac{1}{\sqrt{b}}\}\}$ 中选取的 b随机哈希函数。那么，如果我们设置 $R_{i,j} = h_i(j)$，则 $R \in \mathbb{R}^{b \times n}$ 是一个 AMS 草图矩阵。

3 Analysis

我们考虑过参数化版本的注意力矩阵近似问题。假设 $d \gg n$。在本节中，我们分析这两类算法。在第3.1 节中，我们证明一些将要用到的半正定矩阵的性质。在第3.2 节中，我们证明对矩阵逐元素取指数运算的扰动。在第3.3 节中，我们证明对角归一化矩阵的扰动。在第3.4 节中，我们证明注意力矩阵的扰动。在第3.5 节和第3.6节中，我们分别给出随机化算法和确定性算法。

3.1 半正定矩阵的性质

引理 3.1. 令 $A \in \mathbb{R}^{n\times n}$ 和 $B \in \mathbb{R}^{n\times n}$ 表示两个半正定矩阵。

如果满足以下条件

- 条件1. $(1-\epsilon)B \preceq A \preceq (1+\epsilon)B$;
- 条件2. $-r \le A_{i,j} \le r$, 对所有 $i, j \in [n] \times [n]$.

接下来，我们有

$$B_{i,j} \in [-(1+\epsilon)r, (1+\epsilon)r].$$

证明。 利用事实 2.3 和 条件1，我们知道

$$B_{i,i} \le (1+\epsilon)A_{i,i} \tag{1}$$

注意到矩阵 $B - (1-\epsilon)\cdot A$ 是一个半正定矩阵，因此，

$$\begin{aligned} |B_{i,j} - (1-\epsilon)\cdot A_{i,j}| &\le \sqrt{(B_{i,i} - (1-\epsilon)\cdot A_{i,i})(B_{j,j} - (1-\epsilon)\cdot A_{j,j})} \\ &\le 2\epsilon\sqrt{A_{i,i}A_{j,j}} \\ &\le 2\epsilon r \end{aligned}$$

其中，第一步由引理 2.2 得出，第二步由 $B_{i,\ i} \le (1+\epsilon)A_{i,\ i}$ 得出（见公式（1）），第三步则由引理陈述中的 $A_{i,\ j}$ 的定义得出。

由引理条件2.,

$$(1-\epsilon)\cdot A_{i,j} \in [-(1-\epsilon)\cdot r, (1-\epsilon)\cdot r],$$

结合上述两个等式，我们得到 $B_{i,\ j}$ 的如下取值范围:

$$B_{i,j} \in [-(1+\epsilon)r, (1+\epsilon)r]$$

利用 $(1+\epsilon)\ \cdot A - B$ 是一个半正定矩阵这一事实，我们可以进行对称的论证:

$$B_{i,j} \in [-(1+3\epsilon)r, (1+3\epsilon)r].$$

两个取值范围的交集给出

$$B_{i,j} \in [-(1+\epsilon)r, (1+\epsilon)r]$$

证明至此完成。 □

3.2 逐元素矩阵指数运算的扰动

引理 3.2. 当以下条件成立时，

- 条件1。 $r \in (0, 0.1)$;
- 条件2. $A_{i,j} \in [-r, r]$;
- 条件3. $B_{i,j} \in [-(1+\epsilon)r, (1+\epsilon)r]$.

然后，我们有

- 第1部分。对于所有 $i \in [n],\ j \in [n]$

$$|\exp(A_{i,j}) - \exp(B_{i,j})| \leq \exp(A_{i,j}) \cdot 6r.$$

- 第2部分。对于所有 $i \in [n], j \in [n]$

$$|\exp(A_{i,j}) - \exp(B_{i,j})| \leq \exp(B_{i,j}) \cdot 6r$$

证明。由条件2和条件3，我们有

$$|A_{i,j} - B_{i,j}| \leq (2+\epsilon)r \leq 3r. \tag{2}$$

其中第 2 步由 $\epsilon \in (0, 1)$ 得出。

第1部分的证明。我们有

$$\begin{aligned} |\exp(A_{i,j}) - \exp(B_{i,j})| &= \exp(A_{i,j}) \cdot |1 - \exp(B_{i,j} - A_{i,j})| \\ &\leq \exp(A_{i,j}) \cdot \max\{|1-\exp(3r)|, |1-\exp(-3r)|\} \\ &\leq \exp(A_{i,j}) \cdot 6r. \end{aligned}$$

其中第2步由公式（2）得出，最后一步由条件1 引理陈述和事实2.1得出。

第2部分的证明。

我们有

$$\begin{aligned} |\exp(A_{i,j}) - \exp(B_{i,j})| &= \exp(B_{i,j}) \cdot |1 - \exp(A_{i,j} - B_{i,j})| \\ &\leq \exp(B_{i,j}) \cdot \max\{|1-\exp(3r)|, |1-\exp(-3r)|\} \\ &\leq \exp(B_{i,j}) \cdot 6r. \end{aligned}$$

其中第二步由式 (2) 得出，最后一步则由引理陈述中的条件1和事实2.1得出。

□

3.3 对角归一化矩阵的扰动

引理 3.3（对　角归一化矩阵的扰动）。若满足以下条件

- 条件1。 $|\exp(A_{i,\ j}) - \exp(B_{i,\ j})| \leq \exp(A_{i,\ j}) \cdot 6r$ 对所有 $i \in [n], j \in [n]$
- 条件2。 $|\exp(A_{i,j}) - \exp(B_{i,j})| \leq \exp(B_{i,j}) \cdot 6r$ 对所有 $i \in [n], j \in [n]$

于是，我们有，

- 第1部分。对于所有 $i \in [n]$

$$|(\exp(A_{i,*}) - \exp(B_{i,*})) \cdot \mathbf{1}_n| \leq \exp(A_{i,*})\mathbf{1}_n \cdot 6r$$

- 第2部分。对所有 $i \in [n]$

$$|(\exp(A_{i,*}) - \exp(B_{i,*})) \cdot \mathbf{1}_n| \leq \exp(B_{i,*})\mathbf{1}_n \cdot 6r$$

证明。第1部分的证明。我们可以证明

$$\begin{aligned} |(\exp(A_{i,*}) - \exp(B_{i,*})) \cdot \mathbf{1}_n| &= |\sum_{j=1}^n (\exp(A_{i,j}) - \exp(B_{i,j}))| \\ &\leq \sum_{j=1}^n |\exp(A_{i,j}) - \exp(B_{i,j})| \\ &\leq \sum_{j=1}^n \exp(A_{i,j}) \cdot 6r \\ &= \exp(A_{i,*}) \mathbf{1}_n \cdot 6r \end{aligned}$$

其中第1步由简单代数得出，第2步由三角不等式得出，第3步由引理陈述中的条件1得出，最后一步由简单代数得出。

第2部分的证明。我们可以证明，

$$\begin{aligned} |(\exp(A_{i,*}) - \exp(B_{i,*})) \cdot \mathbf{1}_n| &= |\sum_{j=1}^n (\exp(A_{i,j}) - \exp(B_{i,j}))| \\ &\leq \sum_{j=1}^n |\exp(A_{i,j}) - \exp(B_{i,j})| \\ &\leq \sum_{j=1}^n \exp(B_{i,j}) \cdot 6r \\ &= \exp(B_{i,*}) \mathbf{1}_n \cdot 6r \end{aligned}$$

其中，第1步由简单代数得出，第2步由三角不等式得出，第3步由引理陈述中的条件2得出，最后一步由简单代数得出。

□

3.4 AttentionMatrix 的扰动

我们将注意力矩阵的扰动引理表述如下。

引理 3.4。令 $c_1 > 0$ 和 $c_2 > 0$ 表示两个固定常数。如果以下条件成立

- 条件1。对于所有 $i \in [n]$,

$$|\mathsf{D}(X)_{i,i} - \mathsf{D}(Y)_{i,i}| \leq c_1 \cdot r \cdot \min\{\mathsf{D}(X)_{i,i}, \mathsf{D}(Y)_{i,i}\}$$

- 条件2。对于所有 $i \in [n], j \in [n]$

$$|\exp((XX^\top)_{i,j}) - \exp((YY^\top)_{i,j})| \leq c_2 \cdot r \cdot \min\{\exp((XX^\top)_{i,j}), \exp((YY^\top)_{i,j})\}$$

于是，我们有

$$\|\mathsf{D}^{-1}(X) \exp(XX^\top) - \mathsf{D}^{-1}(Y) \exp(YY^\top)\|_\infty \leq (c_1 + c_2) \cdot r.$$

证明。我们首先将差值分解为

$$\begin{aligned}
&\| \mathsf{D}^{-1}(X)\exp(XX^\top) - \mathsf{D}^{-1}(Y)\exp(YY^\top)\|_\infty \\
&\le \| \mathsf{D}^{-1}(X)\exp(YY^\top) - \mathsf{D}^{-1}(Y)\exp(YY^\top)\|_\infty + \| \mathsf{D}^{-1}(X)\exp(XX^\top) - \mathsf{D}^{-1}(X)\exp(YY^\top)\|_\infty \\
&= B_1 + B_2
\end{aligned}$$

其中最后一步由.....得出

$$B_1 := \| \mathsf{D}^{-1}(X)\exp(YY^\top) - \mathsf{D}^{-1}(Y)\exp(YY^\top)\|_\infty,$$

and

$$B_2 := \| \mathsf{D}^{-1}(X)\exp(XX^\top) - \mathsf{D}^{-1}(X)\exp(YY^\top)\|_\infty.$$

我们将证明分为两项。

第一项。对于每个索引对 $(i,j) \in [n]\times[n]$，我们有

$$\begin{aligned}
B_1 &= |(\mathsf{D}^{-1}(X)\exp(YY^\top) - \mathsf{D}^{-1}(Y)\exp(YY^\top))_{i,j}| \\
&= |(\mathsf{D}^{-1}(X)_{i,i} - \mathsf{D}^{-1}(Y)_{i,i})\exp(YY^\top)_{i,j}| \\
&= |\frac{\mathsf{D}(X)_{i,i} - \mathsf{D}(Y)_{i,i}}{\mathsf{D}(X)_{i,i}\,\mathsf{D}(Y)_{i,i}}\exp(YY^\top)_{i,j}| \\
&\le |\frac{\mathsf{D}(X)_{i,i} - \mathsf{D}(Y)_{i,i}}{\mathsf{D}(X)_{i,i}\,\mathsf{D}(Y)_{i,i}}| \cdot |\exp(YY^\top)_{i,j}| \\
&\le |\frac{c_1 r\,\mathsf{D}(X)_{i,i}}{\mathsf{D}(X)_{i,i}\,\mathsf{D}(Y)_{i,i}}| \cdot |\exp(YY^\top)_{i,j}| \\
&= c_1 r \cdot |\,\mathsf{D}(Y)_{i,i}^{-1}| \cdot |\exp(YY^\top)_{i,j}| \\
&= c_1 r \cdot |\,\mathsf{D}(Y)_{i,i}^{-1}\exp(YY^\top)_{i,j}| \\
&= c_1 r \cdot (\mathsf{D}(Y)^{-1}\exp(YY^\top))_{i,j} \\
&\le c_1 r
\end{aligned}$$

其中第一步由定义得出，第二步由简单代数得出，第三步由三角不等式得出，第四步由引理陈述中的条件1得出，第五步由简单代数得出，第六步由所有元素均为正得出，最后一步由 D 的定义得出。

第二项。对于每个索引对 $(i,j) \in [n]\times[n]$，我们有

$$\begin{aligned}
B_2 &= |(\mathsf{D}^{-1}(X)\exp(XX^\top) - \mathsf{D}^{-1}(X)\exp(YY^\top))_{i,j}| \\
&= |\,\mathsf{D}^{-1}(X)_{i,i} \cdot (\exp(XX^\top)_{i,j} - \exp(YY^\top)_{i,j})| \\
&\le \,\mathsf{D}^{-1}(X)_{i,i} \cdot |(\exp(XX^\top)_{i,j} - \exp(YY^\top)_{i,j})| \\
&\le \,\mathsf{D}^{-1}(X)_{i,i} \cdot c_2 \cdot r \cdot (\exp(XX^\top)_{i,j} \\
&\le c_2 r \cdot (\mathsf{D}^{-1}(X)\exp(XX^\top))_{i,j} \\
&\le c_2 r,
\end{aligned}$$

其中，第一步由简单代数得出，第二步由三角不等式得出，第三步由引理陈述中的条件2得出，第四步由简单代数得出，最后一步由 D 的定义得出。

最后，我们有

$$\begin{aligned}\|\mathsf{D}^{-1}(X)\exp(XX^\top) - \mathsf{D}^{-1}(Y)\exp(YY^\top)\|_\infty &\leq B_1 + B_2 \\ &\leq (c_1+c_2)r.\end{aligned}$$

由此，我们完成了证明。 □

3.5 随机对称注意力近似算法

算法 1 我们的随机化算法

过程 RANDOMIZED($X \in \mathbb{R}^{n\times d}$, δ) ▷ 定理3.5
  $\epsilon_\sigma \leftarrow \Theta(1)$
  $\widetilde{\sigma} \leftarrow$ LEVERAGESCORE($X, \epsilon_\sigma, \delta/2$) ▷ $\widetilde{\sigma} \in \mathbb{R}^n$，算法 2
  $\epsilon \leftarrow \Theta(1)$
  $m \leftarrow \Theta(\epsilon^{-2} n \log(n/\delta))$
  从 X 中抽取 m 行的子集，并根据杠杆分数进行适当的重新加权 $\widetilde{\sigma}$
  生成对角采样矩阵 $\widetilde{D} \in \mathbb{R}^{n\times n}$ 关于上述行的选择
  返回 $Y = \widetilde{D}X$ ▷ $Y \in \mathbb{R}^{n\times d}$
结束过程

定理 3.5 (随机化算法，定理 1.3 的正式版本。 设 $r \in (0,0.1)$。给定一个矩阵 $X \in \mathbb{R}^{n\times d}$，满足 $d \gg n$ 且 $\|XX^\top\|_\infty \leq r$，存在一个运行时间为

$$(\mathrm{nnz}(X) + n^\omega) \cdot \mathrm{poly}(\log(n/\delta))$$

并输出一个矩阵 $Y \in \mathbb{R}^{n\times m}$，其具有 $m = O(n\log(n/\delta))$，使得

$$\|\mathsf{D}(Y)^{-1}\exp(YY^\top) - \mathsf{D}^{-1}(X)\exp(XX^\top)\|_\infty \leq O(r)$$

以概率 $1-\delta$ 成立。
这里 $\mathsf{D}(X) = \mathrm{diag}(\exp(XX^\top)\mathbf{1}_n)$。

证明。运行时间证明。运行时间由引理 5.2得出。
正确性证明。正确性由引理 4.1、引理3.1、引理 3.2、引理3.3以及引理 3.4得出。

□

3.6 确定性对称注意力近似算法

定理 3.6 （确定性算法，定理 1.4 的正式版本）。 设 $r \in (0,0.1)$。给定一个矩阵 $X \in \mathbb{R}^{n\times d}$，满足 $d \gg n$ 且 $\|XX^\top\|_\infty \leq r$，存在一种确定性算法，其运行时间为

$$\widetilde{O}(\min\{\sum_{i\in[d]} \mathrm{nnz}(X_i)^2, dn^{\omega-1}\} + n^{\omega+1})$$

并输出一个矩阵 $Y \in \mathbb{R}^{n \times m}$，满足 $m = O(n)$，使得

$$\| \mathsf{D}(Y)^{-1} \exp(YY^\top) - \mathsf{D}^{-1}(X) \exp(XX^\top) \|_\infty \leq O(r).$$

这里 $\mathsf{D}(X)$ =diag(exp($XX^\top$)$\mathbf{1}_n$).

证明。运行时间证明。运行时间由定理 4.2
.
正确性证明。正确性由定理 4.2、引理3.1、引理 3.2、引理3.3和引理 3.4。

□

4 稀疏化

本节将介绍两类稀疏化工具。在第 4.1 节中，我们将介绍随机稀疏化算法；在第 4.2节中，我们将介绍确定性稀疏化算法。

4.1 随机稀疏化算法

我们已有如下结果。

引理 4.1。给定 $X \in \mathbb{R}^{n \times d}$。令 $m = \Theta(\epsilon^{-2} n \log(n/\delta))$。令 $D \in \mathbb{R}^{m \times n}$ 为算法 *1*输出的采样矩阵。令 $Y = DX$ 则有

$$(1-\epsilon) XX^\top \preceq YY^\top \preceq (1+\epsilon) XX^\top$$

以概率 $1 - \delta$ 成立。

证明。 证明可由引理 5.2得出。 □

4.2 确定性稀疏化算法

我们有以下先前的结果。

定理4.2 （定理 3.1，见 [SXZ22]）。 设 $X = \{x_1, \ldots, x_d\} \subseteq \mathbb{R}^n$ 是一组满足$\sum_{i=1}^d x_i x_i^\top = I_n$的向量。则存在一种确定性算法，可以找到一组权重$\{w_i\}_{i=1}^d \subseteq \mathbb{R}$ 满足

$$(1-\epsilon) \preceq \sum_{i=1}^d w_i x_i x_i^\top \preceq (1+\epsilon) I \quad and \quad |\{i \in [d] : w_i \neq 0\}| = \Theta(\epsilon^{-2} n).$$

该算法的运行时间为

$$\widetilde{O}(\min\{\sum_{i \in [d]} \mathrm{nnz}(x_i)^2, d n^{\omega-1}\} + \epsilon^{-2} n^{\omega+1}).$$

5 杠杆分数采样

在本节中，我们提供一种基于 杠杆分数的采样算法。在第 5.1 节中，我们提供快速杠杆分数计算算法。在第 5.2 节中，我们提供随机化稀疏化算法的运行时间和正确性证明。在第 5.3 节中，我们提供采样过程及其分析。

5.1 快速杠杆分数计算

在本节中，我们提供了一种能够快速近似输入矩阵的杠杆分数的算法。

算法 2 杠杆分数计算

过程 LeverageScore($X \in \mathbb{R}^{n \times d}, \epsilon_\sigma = \Theta(1)\ \delta_\sigma$) ▷引理 5.1
  ▷ 该算法旨在生成杠杆分数的 $O(1)$近似
  $s_1 \leftarrow \widetilde{O}(\epsilon_\sigma^{-2} d)$
  生成 $S_1 \in \mathbb{R}^{s_1 \times n}$ 作为稀疏嵌入矩阵 ▷定义 2.4
  $M \leftarrow (S_1 \cdot X)$ ▷ 此步骤耗时 $\widetilde{O}(\epsilon_\sigma^{-1}\,\mathrm{nnz}(X))$.
  令 $R \in \mathbb{R}^{d \times d}$ 表示 M 的 QR分解中的 R】【。
  $s_2 \leftarrow \Theta(\log(n/\delta_\sigma))$
  生成 $S_2 \in \mathbb{R}^{d \times s_2}$ 作为 JL矩阵 ▷ 定义 2.5 或定义 2.6
  计算 $N \leftarrow R^{-1} S_2$ ▷ $N \in \mathbb{R}^{d \times s_2}$
  对 $j = 1 \to n$ 执行 ▷ 此步骤所需时间为 $\widetilde{O}$（nnz(X)）。
    $\widetilde{\sigma}_j \leftarrow \|(e_j^\top X) N\|_2^2$
  结束循环
  返回 $\widetilde{\sigma}$ ▷ $\widetilde{\sigma}_j = \Theta(\sigma_j)\ \forall j \in [n]$
结束过程

先前工作 [DSW22] 研究了更难的杠杆分数采样算法版本，该算法直接推出了以下结果。

引理5.1 (杠杆分数, [DSW22])。给定一个矩阵 $X \in \mathbb{R}^{n \times d}$ 存在一个算法（过程 LeverageScore 算法 *2*），其运行时间为 $\widetilde{O}(\epsilon_\sigma^{-2}(\mathrm{nnz}(X) + n^\omega)$，并输出一个向量 $\widetilde{\sigma} \in \mathbb{R}^n$，使得 $\widetilde{\sigma}$ 是矩阵 X的杠杆分数的近似值，即

$$\widetilde{\sigma} \in (1 \pm \epsilon_\sigma) \cdot \sigma(X),$$

~概率至少为 $1 - \delta_\sigma$。 O 隐藏了 $\log(d/\delta_\sigma)$ 因子。

5.2 正确性和运行时间

这里我们介绍基于杠杆分数近似算法的随机化稀疏化算法的正确性和运行时间。

引理 5.2. 存在一个算法（算法 *1*）以 $X \in \mathbb{R}^{n \times d}$ 为输入，并输出一个对角采样矩阵 $\widetilde{D}$，其满足 $\|\widetilde{D}\|_0 = O(\epsilon^{-2} d \log(d/\delta)$，且运行时间为

$$\widetilde{O}(\mathrm{nnz}(X) + d^\omega),$$

并且满足的是

$$\Pr[X^\top \widetilde{D}^\top \widetilde{D} X \in (1 \pm \epsilon) \cdot X^\top X] \geq 1 - \delta.$$

~这里 O 隐藏了 $\log(d/\delta)$ 因子。这里我们用 ω 表示矩阵乘法的指数。对于当前的 FMM 算法，我们有 $\omega \approx 2.373$。

证明。 我们首先近似计算杠杆分数，即给出一个 $O(1)$-近似的所有杠杆分数，利用引理 5.1。然后根据杠杆分数对若干行进行采样，利用引理 5.4，我们可以证明其正确性。

□

5.3 采样一批秩一项

本节中，我们介绍形状为 $A^\top A$ 的对称矩阵稀疏化的采样过程。首先，我们定义如下采样过程，

定义5.3（对称矩阵的采样过程）。令 $H = A^\top A$。令 $p_j \geq \beta \cdot \sigma_j(A)/n$，对于某个 $\beta \geq 1$。我们根据概率 p_j，从矩阵 A 中放回采样 T 行。我们使用 j_t 表示在第 t 次试验中采样的行的索引。采样得到的矩阵定义为

$$\widetilde{H} := \frac{1}{T} \sum_{t=1}^{T} \frac{1}{p_{j_t}} a_{j_t} a_{j_t}^\top,$$

其中 T 表示试验次数。

对于上述采样过程，先前的工作给出了以下保证。

引理 5.4 (通过矩阵 Chernoff 进行采样，参见 [DSW22] 中的引理 5.2，作为示例)。令$\epsilon_0 \in (0,1)$ 表示精度参数，且 $\delta_0 \in (0, 0.1)$ 表示失败概率。对于按照定义 $\widetilde{}$采样的矩阵 H 5.3，以至少 $1 - \delta_0$ 的概率，有

$$(1 - \epsilon_0) \cdot H \preceq \widetilde{H} \preceq (1 + \epsilon_0) \cdot H.$$

试验次数为

$$T = \Theta(\epsilon_0^{-2} n \log(n/\delta_0)).$$

参考文献

[ALS+18] Alexandr Andoni、Chengyu Lin、Ying Sheng、Peilin Zhong 和 Ruiqi Zhong。基于 Orlicz 范数的子空间嵌入与线性回归。载于《国际机器学习会议》，第 224–233 页。PMLR，2018年。[AMS96] Noga Alon、Yossi Matias 和 Mario Szegedy。近似频率矩的空间复杂度。载于《第二十八届 ACM 计算理论年度研讨会论文集》，第 20–29 页，1996年。[AS23] Josh Alman 和 Zhao Song。快速注意力需要有界元素。abs/2302.13214，2023年。[AW21] Josh Alman 和 Virginia Vassilevska Williams。一种改进的激光法与更快的矩阵乘法。载于《2021年 ACM-SIAM 离散算法研讨会（SODA）论文集》，第 522–539 页。SIAM，2021年。[BCJ+RR, BMR+20] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习者。《神经信息处理系统进展》，33：1877–1901，2020年。[BNS19] Jan van den Brand、Danupon Nanongkai 和 Thatchaphol Saranurak。动态矩阵求逆：改进算法与匹配的条件下界。载于《2019年 IEEE 第60届计算机科学基础年度研讨会（FOCS）》，第 456–480 页。IEEE，2019年。

[Bra20] Jan van den Brand。一个具有当前矩阵乘法时间复杂度的确定性线性规划求解器。载于《第十四届 ACM-SIAM 离散算法年度研讨会（SODA）论文集》，第 259–278 页。SIAM，2020年。[BSZ23] Jan van den Brand、Zhao Song 和 Tianyi Zhou。大型语言模型中动态注意力维护的算法与困难性。arXiv 预印本 arXiv:2304.02207，2023年。[CAMZ22] Vincent Cohen-Addad、Vahab Mirrokni 和 Peilin Zhong。针对扰动鲁棒实例的大规模并行 k-均值聚类。载于《国际机器学习会议》，第 4180–4201 页。PMLR，2022年。[CDL+21] Beidi Chen、Tri Dao、Kaizhao Liang、Jiaming Yang、Zhao Song、Atri Rudra 和 Christopher Re。像素化蝴蝶：用于神经网络模型的简洁高效稀疏训练。arXiv 预印本 arXiv:2112.00029，2021年。[CDW+21] Beidi Chen、Tri Dao、Eric Winsor、Zhao Song、Atri Rudra 和 Christopher R´e。Scat- terbrain：统一稀疏注意力与低秩注意力。《神经信息处理系统进展》，34：17413–17426，2021年。[CGRS19] Rewon Child、Scott Gray、Alec Radford 和 Ilya Sutskever。使用稀疏 Transformer 生成长序列。arXiv 预印本 arXiv:1904.10509，2019年。[Cha22] ChatGPT。为对话优化语言模型。OpenAI 博客，2022年11月。[CKL18] Diptarka Chakraborty、Lior Kamma 和 Kasper Green Larsen。简洁布尔矩阵向量乘法的紧致 cell-probe 下界。载于《第50届 ACM SIGACT 计算理论研讨会（STOC）论文集》，第 1297–1306 页，2018年。[CLS19] Michael B Cohen、Yin Tat Lee 和 Zhao Song。在当前矩阵乘法时间内求解线性规划。载于 STOC，2019年。[CND+22]Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann 等。PaLM：借助 pathways 扩展语言建模。arXiv 预印本 arXiv:2204.02311，2022年。[CW13] Kenneth L Clarkson 和 David P Woodruff。输入稀疏时间内的低秩近似与回归。载于《ACM 数学杂志》（JACM）；本文的初步版本发表于 STOC，2013年。[DCLT18] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向 Transformer 预训练。arXiv 预印本 arXiv:1810.04805，2018年。[DJS+22] Yichuan Deng、Wenyu Jin、Zhao Song、Xiaorui Sun 和 Omri Weinstein。动态核稀疏化器。arXiv 预印本 arXiv:2211.14825，2022年。[DKOD20] Giannis Daras、Nikita Kitaev、Augustus Odena 和 Alexandros G Dimakis。SMYRF：使用非对称聚类实现高效注意力。《神经信息处理系统进展》，33：6476–6489，2020年。

[DLS23] Yichuan Deng、Zhihang Li 和 Zhao Song。秩一矩阵感知的改进样本复杂度。arXiv 预印本 arXiv:2303.06895，2023年。[DLY21] Sally Dong、Yin Tat Lee 和 Guanghao Ye。针对小树宽线性规划的近线性时间算法：鲁棒中心路径的多尺度表示。载于第53届 ACM SIGACT计算理论研讨会论文集，第1784–1797页，2021年。[DSW22] Yichuan Deng、Zhao Song 和 Omri Weinstein。输入稀疏时间内的差异最小化。arXiv 预印本 arXiv:2210.12468，2022年。[EMM+23] Alessandro Epasto、Jieming Mao、Andres Munoz Medina、Vahab Mirrokni、Sergei Vassilvitskii 和 Peilin Zhong。流式频率矩估计的差分隐私持续发布。arXiv 预印本 arXiv:2301.05605，2023年。[EMMZ22] Alessandro Epasto、Mohammad Mahdian、Vahab Mirrokni 和 Peilin Zhong。用于最小规模聚类的大规模并行动态算法。载于2022年ACM-SIAM离散算法年度研讨会论文集（SODA），第1613–1660页。SIAM，2022年。[EMZ19] Hossein Esfandiari、Vahab Mirrokni 和 Peilin Zhong。流式平衡聚类。arXiv 预印本 arXiv:1910.00788，2019年。[GS22] Yuzhou Gu 和 Zhao Song。更快的小树宽半正定规划求解器。arXiv 预印本 arXiv:2211.06033，2022年。[GSYZ23] Yuzhou Gu、Zhao Song、Junze Yin 和 Lichen Zhang。通过鲁棒交替最小化，以近线性时间实现低秩矩阵补全。arXiv 预印本 arXiv:2302.11068，2023年。[HKNS15] Monika Henzinger、Sebastian Krinninger、Danupon Nanongkai 和 Thatchaphol Saranurak。通过在线矩阵向量乘法猜想统一并强化动态问题的困难性。载于第四十七届 ACM SIGACT 计算理论研讨会论文集（STOC），第21–30页，2015年。[HWL21] Weihua He、Yongyun Wu 和 Xiaohua Li。神经机器翻译中的注意力机制：综述。载于2021年 IEEE 第五届信息技术、网络、电子与自动化控制会议（ITNEC），第5卷，第1485–1489页。IEEE，2021年。[JLSW20] Haotian Jiang、Yin Tat Lee、Zhao Song 和 Sam Chiu-wai Wong。用于凸优化、凸凹博弈及其应用的改进切平面法。载于 STOC，2020年。[JLSZ23] Haotian Jiang、Yin Tat Lee、Zhao Song 和 Lichen Zhang。在 $\widetilde{O}(n^4)$时间内进行整数极小值凸优化。arXiv 预印本，2023年。[JSWZ21] Shunhua Jiang、Zhao Song、Omri Weinstein 和 Hengjie Zhang。更快地进行动态矩阵求逆，从而更快求解线性规划。载于 STOC，2021年。[KKL20] Nikita Kitaev、Lukasz Kaiser 和 Anselm Levskaya。Reformer：高效的 Transformer。arXiv 预印本 arXiv:2001.04451，2020年。

[KVPF20] Angelos Katharopoulos、Apoorv Vyas、Nikolaos Pappas 和 Fran¸cois Fleuret。Transformer 即循环神经网络：具有线性注意力的快速自回归 Transformer。载于国际机器学习会议，第5156–5165页。PMLR，2020年。[LG14] Fran¸cois Le Gall。张量幂与快速矩阵乘法。载于第39届国际符号与代数计算研讨会论文集（ISSAC '14），2014年。[LSZ19] Yin Tat Lee、Zhao Song 和 Qiuyi Zhang。以当前矩阵乘法时间求解经验风险最小化问题。载于学习理论会议（COLT），第2140–2157页。PMLR，2019年。[LSZ23] Zhihang Li、Zhao Song 和 Tianyi Zhou。求解正则化 exp、cosh 和 sinh 回归问题。arXiv 预印本 arXiv:2303.15725，2023年。[LW17] Kasper Green Larsen 和 Ryan Williams。更快的在线矩阵-向量乘法。载于第二十八届 ACM-SIAM 离散算法研讨会论文集（SODA），第2182–2189页，2017年。[MMS+19] Louis Martin、Benjamin Muller、Pedro Javier Ortiz Suarez、Yoann Dupont、Laurent Romary、Eric Villemonte de La Clergerie、Djame Seddah 和 Benoit Sagot。Ca memBERT：一种美味的法语语言模型。arXiv 预印本 arXiv:1911.03894，2019年。[NN13] Jelani Nelson 和 Huy L Nguyˆen。OSNAP：通过更稀疏的子空间嵌入实现更快的数值线性代数算法。载于第54届 IEEE 计算机科学基础研讨会（FOCS），第117–126页。IEEE，2013年。[Ope23] OpenAI。GPT-4 技术报告。arXiv 预印本 arXiv:2303.08774，2023年。[QSZZ23] Lianke Qin、Zhao Song、Lichen Zhang 和 Danyang Zhuo。一种用于投影矩阵向量乘法的在线统一算法及其在经验风险最小化中的应用。载于 AISTATS，2023年。[RNS+18] Alec Radford、Karthik Narasimhan、Tim Salimans、Ilya Sutskever 等。通过生成式预训练提升语言理解能力。2018年。[RSW16] Ilya Razenshteyn、Zhao Song 和 David P Woodruff。具有可证明保证的加权低秩近似。载于第四十八届 ACM 计算理论研讨会论文集，第250–263页，2016年。[RSZ22] Aravind Reddy、Zhao Song 和 Lichen Zhang。动态张量积回归。arXiv 预印本 arXiv:2210.03961，2022年。[RWC+19] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。语言模型是无监督多任务学习器。OpenAI 博客，1(8):9，2019年。[SWYZ22] Zhao Song、Yitan Wang、Zheng Yu 和 Lichen Zhang。一阶方法的草图化：适用于低带宽信道且具有脆弱性的高效算法。arXiv 预印本 arXiv:2210.08371，2022年。[SWZ17] Zhao Song、David P Woodruff 和 Peilin Zhong。具有逐项 l1 范数误差的低秩近似。载于第49届 ACM SIGACT计算理论研讨会论文集，第688–701页，2017年。

[SWZ19] Zhao Song、DavidPWoodruff 和 Peilin Zhong。相对误差张量低秩近似。在第三十届 ACM-SIAM 离散算法研讨会论文集，第2772–2789 页。SIAM，2019年。[SXYZ22] Zhao Song、ZhaozhuoXu、YuanyuanYang 和 Lichen Zhang。使用低维自适应数据结构加速 Frank-Wolfe 算法。arXiv 预印本arXiv:2207.09002，2022年。[SXZ22] ZhaoSong、ZhaozhuoXu 和Lichen Zhang。使用内积搜索数据结构加速稀疏化。arXiv预印本 arXiv:2204.03209，2022年。[SYYZ22a] Zhao Song、Xin Yang、Yuanyuan Yang 和 LichenZhang。草图化与差分隐私相遇：动态 Kronecker 投影维护的快速算法。arXiv预印本 arXiv:2210.11542，2022年。[SYYZ22b] Zhao Song、Xin Yang、Yuanyuan Yang和 Tianyi Zhou。结构化约翰椭球计算的更快算法。arXiv预印本arXiv:2211.14407，2022年。[SYYZ23] Zhao Song、Mingquan Ye、JunzeYin和LichenZhang。带 ℓ_∞ 保证的快速回归的近乎最优界。arXiv预印本arXiv:2302.00248，2023年。[SZZ21] Zhao Song、LichenZhang和 RuizheZhang。在次二次时间内训练多层过参数化神经网络。arXiv预印本arXiv:2112.07628，2021年。[UAS+20] Mohd Usama、Belal Ahmad、Enmin Song、M Shamim Hossain、MubarakAlrashoud 和GhulamMuhammad。使用卷积神经网络和循环神经网络的基于注意力的情感分析。《未来一代计算机系统》，113:571–578，2020年。[VSP+17] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Lukasz Kaiser 和 Illia Polosukhin。注意力机制就是你所需要的一切。《神经信息处理系统进展》，30，2017年。[Wil12] Virginia Vassilevska Williams。比 Coppersmith-Winograd 更快地进行矩阵乘法。在第四十四届 ACM计算理论研讨会（STOC）论文集，第 887–898 页。ACM，2012年。[WLK+20] SinongWang、Belinda Z Li、Madian Khabsa、HanFang和 HaoMa。Linformer：具有线性复杂度的自注意力。arXiv预印本arXiv:2006.04768，2020年。[Ye20] Guanghao Ye。求解结构化凸程序的快速算法。华盛顿大学，本科毕业论文，2020年。[Zha22] Lichen Zhang。通过数据结构加速优化：更快的搜索、采样与维护。硕士论文，卡内基梅隆大学，2022年。[ZHDK23] Amir Zandieh、InsuHan、MajidDaliri和Amin Karbasi。KDEformer：通过核密度估计加速Transformer。arXiv预印本arXiv:2302.02451，2023年。[ZRG+22] Susan Zhang、Stephen Roller、Naman Goyal、Mikel Artetxe、Moya Chen、Shuohui Chen、Christopher Dewan、Mona Diab、Xian Li、Xi Victoria Lin 等。OPT：开放的预训练Transformer语言模型。arXiv 预印本 arXiv:2205.01068，2022年。