arXiv:2601.11765v1 [cs.DS] 16 Jan 2026

通过向量相似性搜索进行和估计

Stephen Mussmann[*†] 佐治亚理工学院[‡]
Mehul Smriti Raje[*] Coactive AI
Kavya Tumkur Coactive AI
Oumayma Messoussi Desjardins[‡]
Cyprien Hachem Coactive AI
Seby Jacob Coactive AI

摘要

用于表示图像、文本和音频等对象的语义嵌入已广泛应用于机器学习，并推动了向量相似性搜索方法的发展，以检索语义相关的对象。在本文中，我们研究了一个与之相近的任务：估计集合中所有对象的和，例如核密度估计（KDE）以及 softmax 分布的归一化常数。尽管现有解决方案已证明可将和估计任务归约为获取$\mathcal{O}(\sqrt{n})$个最相似的向量，其中 n 为对象数量，但我们提出了一种新算法，只需获取 $\mathcal{O}(\log(n))$个最相似的向量。我们以指数衰减的概率将对象随机分配到各个级别，并为每个级别构建一个向量相似性搜索数据结构。利用每个级别中的前-k个对象，我们提出了和的无偏估计，并证明了一个高概率相对误差界。我们在 OpenImages 和 Amazon Reviews 上结合向量相似性搜索实现进行了实验，结果表明，与现有归约方法相比，我们的方法能够以更少的计算时间达到更低的误差。我们展示了该方法在密度估计、计算 softmax 分母以及计数球内向量数量等应用中的结果。我们的部分删节代码可在 https://github.com/CoactiveAI/ sum-estimation-public/tree/main 找到。

1 引言

在过去十年中，机器学习系统越来越多地利用语义嵌入来表示词语和词元（Devlin et al., 2019; Peters et al., 2018）、句子和文档（Reimers and Gurevych, 2019; Cer et al., 2018）、图像（Radford et al., 2021b; Zhai et al., 2023）、音频片段（Devnani et al., 2024; Turian et al., 2022）、视频（Ashutosh et al., 2023; Chang et al., 2020）以及各种其他对象（Agarwal et al., 2021）。鉴于检索语义相关对象的重要性，向量相似性搜索算法取得了显著进展，包括局部敏感哈希（LSH）（Indyk and Motwani, 1998; Charikar, 2002）、基于聚类的搜索（Johnson et al., 2019）以及分层可导航小世界图（HNSW）（Malkov and Yashunin, 2020）。由于向量相似性搜索系统实用且有效，市面上已有多种商业解决方案，它们通常基于分层可导航小世界图（HNSW），较少采用基于聚类的搜索（Sukhodolskaya and Vasnetsov, 2025; Johnson et al., 2019）。

在这项工作中，我们研究了一组相关问题，这些问题涉及计算或估计数据集中所有元素的总和。此类任务的例子包括核密度估计（KDE）（Charikar

[*]共同贡献
[†]通讯作者：mussmann@gatech.edu
[‡]在 Coactive AI 任职期间完成的工作

预印本。

等人，2020年；Langrené 和 Warin，2018年），以及计算大型 softmax 层（对数线性模型）的概率（Radford 等人，2018年、2021b）。既有研究包含两大类通用方法：基于局部敏感哈希的采样估计（Charikar 和 Siminelakis，2017年；Spring 和 Shrivastava，2018年），以及使用向量相似性搜索预言机的方法（Mussmann 等人，2017年；Karppa 等人，2022年）。由于基于局部敏感哈希的搜索方法目前不如分层可导航小世界图（HNSW）（Malkov 和 Yashunin，2020年）等其他向量相似性搜索技术，因此我们在此重点关注后一类方法。这类方法利用贡献最大的值（通过向量相似性搜索获得）和均匀随机样本，计算总和的简单估计值。遗憾的是，这类方法需要$\mathcal{O}(\sqrt{n})$个相似向量和随机样本，其中 n 是数据集的大小。我们提出了一种方法，同样可以使用任意最大化预言机，但只需要 $\mathcal{O}(\log n)$个相似向量。

在本文中，我们假设可以访问最大化预言机：给定一个非负函数 f（例如，对于使用高斯核进行核密度估计，负欧几里得距离平方的指数函数）和一组 n对象 X，该预言机构建一种数据结构，使其能够根据查询 q快速计算 $\operatorname{argmax}_{x\in X} f(x,q)$，更一般地，还能检索排名前$k$ 的对象。借助这样的预言机，我们的算法只需从最大化预言机中检索 $\mathcal{O}(\log\ n)$个元素，即可对 $\sum_{x\in X} f(x,q)$给出可证明准确的估计。核心思想是以指数衰减的概率将每个数据点分配到某个“级别”，这与跳表（Pugh，1990年）和分层可导航小世界图（Malkov 和 Yashunin，2020年）的做法相同，然后为每个级别构建一个最大化预言机数据结构。注意，索引数据点的总数就是 $|X|$，因为每个数据点恰好出现在一个数据结构中。总共有 $\mathcal{O}(\log n)$个级别，每个级别对应一个数据结构。对于查询 q，我们从每个级别检索排名前k的对象，并使用一个简单的算法计算无偏估计（详见第 3.2 节）。我们展示了实验结果，表明我们的方法在 Open Images 和 Amazon Reviews 上的和估计任务中效果显著。

我们的贡献主要有以下几项：

- 展示一种和估计量，利用从每个指数采样层级中取得的前 k 个元素来估计总和。
- 证明我们的估计量是无偏的，并给出一个以高概率成立的误差界（相对于层级的采样）。
- 介绍一种新的和估计应用：统计球内元素的数量。
- 展示实验结果，证明我们的方法有效。

在本文余下部分中，我们首先在第 2 节介绍问题设定，并在第 3 节提出和分析我们的算法。随后，我们在第 4 节介绍并展示实验结果。最后，我们在第 5 节和第 6 节讨论相关工作、机遇与局限，并作总结。

2 设置

我们的总体目标是：对于给定的查询 $q\in\mathcal{Q}$、对象集合 $X\subset\mathcal{X}$ 以及已知的非负函数 f：$\mathcal{X}\times\mathcal{Q}\to\mathbb{R}_+$，估计 $\sum_{x\in X} f(x,q)$。一个具体示例是核密度估计（KDE），其中，对于给定的带宽， $\mathcal{Q}=\mathbb{R}^d$、$\mathcal{X}=\mathbb{R}^d$ x $2\sigma^2$ σ 和 $f(,q)$ $=$ $\frac{1}{|X|}$ $\frac{1}{2\pi\sigma^2()^{d/2}}\exp(-\|x-q\|_2^2/())$。为简化记号，定义 $f_q=f(\cdot,q)$。

虽然我们可以通过线性扫描精确计算总和，但在这里，我们利用黑盒最大化预言机，例如向量相似性搜索技术。该技术基于预先计算的数据结构，实际运行速度可能远快于线性扫描。具体而言，我们假设可以访问一个最大化预言机，它能够高效地计算 $\operatorname{Top}_k(X,f_q)=\operatorname{argmax}_{S\subset X:|S|\le k}\sum_{x\in S} f_q(x)$。对于核密度估计（KDE），请注意， f_q 在 x 与 q 之间的欧几里得距离最小时取得最大值。

我们工作的一个核心思想是：每个元素 $x\in X$ 都会被随机分配一个“级别” $\ell(x)$，并且我们会为每个级别构建一个最大化数据结构。如果将 $X_\ell=\{x\in X$ 定义为： $\ell(x)=\ell\}$ 处于级别 ℓ 的数据元素，那么我们的最大化预言机就能高效地计算 $\operatorname{Top}_k(X_\ell,\ f_q)$。为数据点建立索引时，我们将每个数据点的级别独立采样为一个参数为 $p=1/2$ 的几何随机变量，因此 $\Pr(\ell(x)=\ell)=2^{-\ell}$。

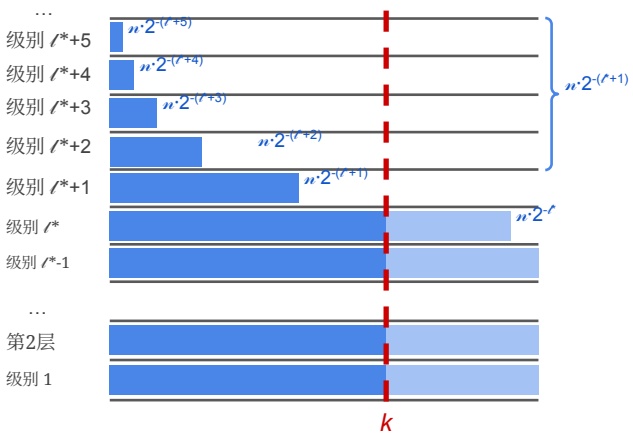


图1：U 大小上界证明的示意图。蓝色表示每个级别上元素数量的期望值，深蓝色表示与 k 的最小值。期望元素数量超过 k 的级别有 ℓ^* 个，其余级别的期望元素数量则少于 $2k$。

3 方法与分析

为便于分析，我们固定一个特定的查询 q，并将 x_i 定义为元素 $x \in X$，其 i^{th} 值最大；若出现并列，则随机打破。因此，$x_1 = \text{argmax}_{x \in X} f_q(x)$ 和 $x_n = \text{argmin}_{x \in X} f_q(x)$。为方便起见，定义 $f_i = f_q(x_i)$ 和 $\ell_i = \ell(x_i)$。因此，我们的目标是在能够访问 $\{\text{Top}_k(X_\ell,\ f_q)\}_\ell$ 的情况下估计 $\sum_{i=1}^n f_i$。

3.1 前 {style id='1'>k</style} 的并集

将所有层级中的前 k 个元素的并集定义为 $U = \bigcup_\ell \text{Top}_k(X_\ell, f_q)$。

我们的估计将是 U 的函数。直观地说，U 的大小将与 k（每个级别中的最大元素数）以及 $\log n$（非空级别的数量）成正比。

命题 3.1。 $\mathbb{E}[|U|] \le \mathcal{O}(k \log\ n)$

证明。

$$\mathbb{E}[|U|] = \mathbb{E}\left[\sum_\ell \min(|X_\ell|, k)\right] \le \sum_\ell \min(2^{-\ell} n, k) \tag{1}$$

将 ℓ^* 定义为满足以下条件的最高级别：$2^{-\ell} n \ge k$：$\ell^* = \lfloor \log_2(n/k) \rfloor$。于是，$\mathbb{E}[|U|] \le \ell^* k + \sum_{\ell > \ell^*} 2^{-\ell} n \le (\ell^* + 2)k$ □

有关证明中不同级别贡献的可视化，请参见图1。

3.2 构造无偏估计

对于给定的 i，$x_i \in U$的概率是多少？如果以 $\{\ell_j\}_{j<i}$为条件，这个概率就很容易计算。定义$C_{\ell,\ i} = \sum_{j=1}^{i} \mathbf{1}[\ell_j = \ell]$ 为被分配到级别 ℓ的前i 个元素的数量。如果$C_{\ell,\ i-1} < k$，那么级别 ℓ 上对于第 i^{th} 大的元素仍有“可用性”。定义 I_i为指示变量 $\mathbf{1}\left[C_{\ell_i, i-1} < k\right]$，这等价于 $x_i \in U$。定义 p_i 为 ℓ_i 被抽样为此类级别的概率：$p_i = \Pr_\ell[C_{\ell, i-1} < k] = \sum_{\ell:\ C_{\ell,\ i-1} < k} 2^{-\ell}$。

我们可以看到，在给定 $\{\ell_j\}_{j<i}$ 的条件下，$\mathbb{E}[I_i/p_i] = 1$。这促使我们使用无偏估计 $E = \sum_{i=1}^n \frac{I_i}{p_i} f_i$ 来估计 $F = \sum_{i=1}^n f_i$。

算法 1 快速估计**输入：** 每个级别的前 k 个元素的并集： U，待估计的函数 $f_q E = 0 \forall \ell : C_\ell = 0 p = 1$**对于** x **中** U，按 f_q 值降序排列**执行** $E \leftarrow E + f_q(x)/p C_{\ell(x)} \leftarrow C_{\ell(x)} + 1$**如果** $C_{\ell(x)} = k$ **则** $p \leftarrow p - 2^{-\ell(x)}$**结束 if，结束 for，返回** E

由于 $I_i = 1$ 当且仅当 $x_i \in U$，因此 U 之外的元素不会产生任何贡献。此外，或许令人惊讶的是，我们仅根据 U 中的元素，就能计算 U 中元素的 p_i。具体而言，如果我们按照 f_q 值递减的顺序遍历 U，那么每当某个级别填满时（$C_{\ell(x)} = k$），我们就可以降低概率 p。简单的算法见算法 1，其运行时间为 $\mathcal{O}(|U|)$。

3.3 高概率界

本节给出我们的主要分析结果：我们的方法的相对误差高概率界。我们将逐步讲解证明，不过有几个引理的证明放在附录 A 中。

定理 3.2。 对于任意 n、 k、 δ，只要 $k \geq 8 \log(3(\ell^* + 3)/\delta)$ *set* $b = k - \left\lceil \sqrt{2k \log(3(\ell^* + 3)/\delta)} \right\rceil$。那么，以概率 $1 - \delta$，

$$\frac{|E - F|}{F} \leq \sqrt{\frac{3 \log(3/\delta)}{4b}} + \frac{2 \log(3/\delta)}{3b} = \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{k}}\right) \tag{2}$$

例如，对于一个满足 $n = 10^7$、 $k = 200$和 $\delta = 0.05$的数据集，则有 $\ell^* = 15$ 和 $b = 147$，因此，我们的估计值以 95% 的置信度具有小于 0.1631 的相对误差。

证明。 我们先展示一个以高概率成立的事件。

引理 3.3。 对于任意 δ 和 $k \geq 8 \log(1/\delta)$，令 $b = k - \lceil \sqrt{2k \log(1/\delta)} \rceil$。那么， $b \geq k/2 - 1$，且以概率 $1 - (\ell^* + 3)\delta$，

$$\forall \ell : C_{\ell, \min(2^\ell b, n)} < k \tag{3}$$

证明见附录 A.1，推导过程与 Chernoff 界类似，因为 $C_{\ell,i} \sim$ 二项分布$(i, 2^{-\ell})$

直观而言，公式 3 意味着没有任何层级会“过快”填满。这将使我们之后能够为 p_i 给出一个下界，

我们的关键步骤是应用 Bernstein 不等式的鞅版本。下面给出该不等式，它是对标准 Bernstein 不等式的鞅版本（Dzhaparidze 和 Van Zanten，2001）所作的改编。更多详情见附录 A.3。

定理 3.4。 设 $\{Z_i\}_{i=1}^n$ 是一个带滤过 $\{\mathcal{F}_i\}_{i=1}^n$ 的鞅，因此 Z_i 是 $\mathcal{F}_i$ 和 $\mathbb{E}[Z_i | \mathcal{F}_{i-1}] = 0$ 的函数。假设以概率 1 有 $|Z_i| \leq M$。那么，对于任意 δ，令 V 为和的条件方差 $V = \sum_{i=1}^n \mathbb{E}[Z_i^2 | \mathcal{F}_{i-1}]$，

$$\Pr\left(\left|\sum_i Z_i\right| \geq \sqrt{2V \log(2/\delta)} + \frac{2M}{3} \log(2/\delta)\right) \leq \delta \tag{4}$$

注意到 $E-F=\sum_{i=1}^{n}\left(\frac{I_i}{p_i}-1\right)f_i$ 并定义 $Z_i=\underbrace{\left(\frac{I_i}{p_i}-1\right)}f_i$，我们的目标是得到 $|E-F|=|\sum_i Z_i|$ 的高概率界。

接下来，我们证明两个引理，以界定 p_i 和 f_i，从而帮助界定 Z_i 及其方差。

引理 3.5. 若公式 *3* 成立，则对于 $i\le 2b$、$p_i=1$，以及对于 $i>2b$、$p_i\ge b/i$

证明。若 $i\le 2b$，对于任意 $\ell\ge 1$ $i\le 2^\ell b$，因此由公式 3 可得 $C_{\ell,\ i}<k$。因此，$I_i=1$ 和$p_i=\sum_{\ell\ge 1}2^{-\ell}=1$。
若 $i>2b$，对于任意 $\ell\ge\lceil\log_2(i/b)\rceil$ $i\le 2^\ell b$，因此由公式 3 可得 $C_{\ell,\ i}<k$。$p_i=\sum_{\ell:\ C_{\ell,\ i-1}<k}2^{-\ell}\ge\sum_{\ell\ge\lceil\log_2(i/b)\rceil}2^{-\ell}=2^{1-\lceil\log_2(i/b)\rceil}\ge b/i$。

□

引理 3.6。 $f_i\le F/i$

证明。由于 $\{f_i\}_{i\in[n]}$ 为正且单调不减，因此 $F\ge\sum_{j=1}^{i}f_j\ge if_i$。所以，$f_i\le F/i$。

□

为求 $|Z_i|$ 的上界，我们在公式 3 成立的前提下考察三种情况：

- If $i\le 2b$, $p_i=1$ and $I_i=1$, so $I_i/p_i-1=0$ and $Z_i=0$.
- If $i>2b$ and $I_i=1$, $|(I_i/p_i-1)f_i|=(1/p_i-1)f_i\le\frac{f_i}{p_i}\le\frac{F/i}{b/i}=F/b$.
- If $i>2b$ and $I_i=0$, $|(I_i/p_i-1)f_i|=f_i\le\frac{F}{i}\frac{i}{2b}\le F/b$.

在任何情况下，若公式 3 成立，则 $|Z_i|\le F/b$。

通过方差计算可得和的条件方差为：

$$V=\sum_{i=1}^{n}\mathbb{E}\left[\left(\frac{I_i}{p_i}-1\right)^2 f_i^2\middle|\{\ell_j\}_{j<i}\right]=\sum_{i=1}^{n}(1/p_i-1)f_i^2 \tag{5}$$

根据公式 3，引理 3.5 表明，对于 $i\le 2b$，$p_i=1$，因此求和中的前 $2b$ 项均为零。接着，利用引理 3.5 和引理 3.6，和的条件方差为

$$V=\sum_{i=b+1}^{n}(1/p_i-1)f_i^2\le\sum_{i=b+1}^{n}(i/b)(F/i)f_i\le F^2/b \tag{6}$$

我们在附录 A.2 中证明，经过更仔细的论证可得 $V\le\frac{3}{8}\frac{F^2}{b}$。

应用引理 3.3（失败概率为 3^{δ} (ℓ^*+3)）和定理 3.4（失败概率为 $2\delta/3$），可知以概率 $1-\delta$

$$|E-F|\le\sqrt{2\frac{3}{8}\frac{F^2}{b}\log(3/\delta)}+\frac{2F}{3b}\log(3/\delta) \tag{7}$$

将等式两边同时除以 F，即可得出结果。 □

3.4 控制变量项

请注意，当所有 $\{f_i\}_i$ 都相等时，我们可以将任意 f_i 乘以已知量 n，从而得到完美估计。不过，由于 $\sum_{i=1}^{n}\frac{I_i}{p_i}$ 的平均值为 n，我们的算法并不完美，但通常会高估或低估。这里我们引入一个修正项。

由于 $F=\sum_i f_i=cn+\sum_i(f_i-c)$，我们可以构造一个无偏估计 $E_c=cn+\sum_i\frac{I_i}{p_i}(f_i-c)=c(n-\sum_i I_i/p_i)+E$。这是一种用于方差缩减的控制变量方法。在附录 B 中，我们说明了为何将 c 选为 f_i 值中较小一半的平均值：$\{f_i\}_{i\ge n/2}$；该值可以通过均匀随机样本估计（未填充层级上的元素并集是一个均匀随机样本）。

4 实验

在本节中，我们将所提出的方法与基线方法进行比较，并对 f、X 和 q 采用若干种具体实现。一个关键考量是，方法需要在 f 呈“尖峰”分布（少数几个 $f(x,\ q)$值占据了 F的大部分）以及 f呈“平坦”分布（ $f(x,\ q)$的变化很小）这两种情况下都表现良好。
除非另有说明，所有实验均使用 Qdrant（Apache 2.0 许可证）作为向量相似性搜索实现，并在一台配备 300 GB RAM 和 16 个核心的 r5.4xlarge 亚马逊云科技实例上运行。所有其他作业均在配备 488 GB RAM 和 64 个核心的 i3.16xlarge 亚马逊云科技实例上运行。一次性编码耗时约一周，最后一组实验耗时数小时。为协助加载数据集和模型，我们大量使用 Hugging Face（Apache 2.0 许可证）。

4.1 任务

核密度估计核密度估计（KDE）（Parzen，1962年）是一种从样本估计概率密度函数（pdf）的非参数方法。它广泛应用于异常值检测、聚类以及各种基于密度的数据选择策略等领域。KDE 通过累加数据集中所有向量 x 的概率贡献，来估计点 q 处的密度。对于常见的高斯核，核密度估计为

$$\frac{1}{|X|}\sum_{x\in X}(2\pi\sigma^2)^{-d/2}\exp\left(-\frac{\|x-q\|_2^2}{2\sigma^2}\right) \tag{8}$$

其中，d 是向量维度，σ 是控制概率密度函数（pdf）平滑程度的带宽参数。$f(x,\ q) = \frac{1}{|X|}(2\pi\sigma^2)^{-d/2}\exp\left(-\frac{\|x-q\|_2^2}{2\sigma^2}\right)$，它关于欧几里得距离单调递减，因此适合通过向量相似性搜索进行优化。这里，f 在 σ 较小时呈“尖峰状”，在 σ较大时则较为“平坦”。

Softmax 归一化常数softmax 函数用于对离散集合进行概率建模，例如数据集或批次中的数据点、用于分类的标签，以及用于语言建模的词元。
一般形式为 $\Pr(x) \propto \exp((q\cdot x)/T)$，其中 q 用于参数化分布。为了计算概率，我们需要知道归一化常数，也称为配分函数，$\sum_{x\in X}\exp((q\cdot x)/T)$。对于大型集合，计算这一量可能会非常耗费计算资源$X$，例如分类和语言建模中的大型输出空间，以及对比学习中的大型数据点集合。这里，$f(x,\ q) = \exp((q\cdot x)/T)$，它关于点积单调递增，因此可以通过向量相似性搜索进行优化。这里，f 在 T 较小时呈“尖峰状”，在 T较大时则较为“平坦”。

计数 计数问题涉及估算数据集中距离向量 q不超过半径 r 的向量数量。计数等于 $F = \sum_{x\in X} f(x,\ q)$，其中 $f(x,\ q) = \mathbf{1}[||x-q||_2 \le r]$。$f$ 随欧几里得距离增大而单调递减。在这里，当 r 较小时，f 呈“尖峰”状；当 r较大时，则较为“平坦”。

4.2 模型

ResNet-50 对于图像嵌入，我们使用在 ImageNet 上预训练的 ResNet-50 模型的最后一层（He 等，2016年），该模型通过 torchvision（BSD 三条款许可证）获取。由于经过预训练，这些嵌入包含丰富的通用视觉特征。

CLIP ViT-L/14@336px 对于 softmax 实验中的多模态嵌入，我们使用 OpenAI 的 CLIP ViT-L/14 @336px 图像编码器（Radford 等，2021年a）（MIT 许可证）。该编码器采用视觉变换器（ViT），以 336×336 图像作为输入。此模型基于大量且多样化的图文对数据集进行训练，从而为视觉和文本模态学习共享的嵌入空间。按照 CLIP 损失函数中的做法，我们对实验所用的这些嵌入进行归一化。

DistilBERT-base-uncased 在基于文本的实验中，我们使用 distilbert-base-uncased 模型（Sanh 等，2019年）的最终隐藏状态（Apache 2.0 许可证）。DistilBERT 是 BERT 的一个更小（参数量减少 40%）、速度更快（处理的文本多 60%）的变体，通过从 BERT 进行知识蒸馏得到，

能够保留 BERT 97% 的语言理解性能。该模型在与 BERT 相同的语料库（英文维基百科和 BookCorpus）上训练。

4.3 数据集

Open Images Open Images 数据集 v7（Kuznetsova 等人，2020年）（标注：CC BY 4.0，图像：CC BY 2.0）包含超过 900 万张带有图像级标签的自然图像。在我们的实验中，我们使用训练集中 674 万张图像。每张图像都包含由人工或机器标注的多个类别标签，训练数据集约有 20,000 个类别。

Flickr30k 数据集 Flickr30k 数据集（Young 等人，2014年）（CC0 1.0）包含约 3 万张 Flickr 图像，每张图像配有 5 条众包生成的描述。在本研究中，我们不使用图像，只使用描述。

亚马逊评论（Kindle 商店） McAuley-Lab/Amazon-Reviews-2023 数据集（Hou 等人，2024年）（MIT 许可证）是一个大型产品评论集合，涵盖亚马逊上的众多类别。在我们的实验中，我们从原始_评论_Kindle_商店子集中随机抽取 1,000 万条评论（通过 Hugging Face 下载）。这些评论针对在 Kindle 商店购买的电子书。我们的实验使用评论正文进行编码。

4.4 基线方法

前 k 个 前 k 个方法仅对贡献最大的 k 个元素求和，以近似总和。对于查询 q，总和估计值为

$$E_{\text{top-}k} = \sum_{x \in \text{Top}_k(X, f_q)} f(x, q), \tag{9}$$

where 前$_k(X,\ f_q)$定义于第 2 节。当 $f(x,\ q)$中的大部分 $f(\cdot)$ 贡献了大部分 $f(\cdot)$ 具有“尖峰”特征）时，该方法表现良好。

随机采样 随机采样方法从 X中均匀抽取大小为 m 的子集 S，不进行放回，并使用该子集估计总和。对于查询 q，估计值为

$$E_{\text{random}} = \frac{n}{|S|} \sum_{x \in S} f(x, q). \tag{10}$$

当 $f(x,\ q)$ 的变化较小时，此方法效果良好（f 是“平坦的”）。

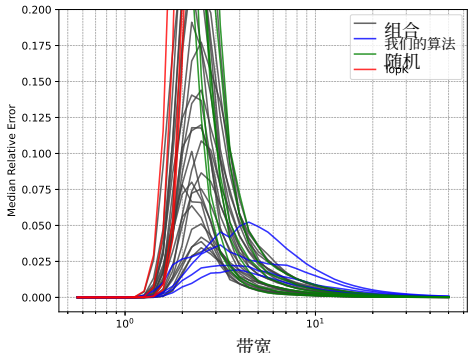
组合 top-k 与随机采样 注意到，当 f“平坦”时，top-k 方法会失效；而当 f“尖峰”时，随机采样也会失效。（Mussmann 等人，2017年；Karppa 等人，2022年）提出了一种组合方法，试图兼得两者的优点。对于查询 q 和大小为 m 的随机样本 S，我们可以定义 $T = S \setminus \text{Top}_k(X,\ f_q)$，并使用无偏估计

$$E_{\text{combined}} = \sum_{x \in \text{Top}_k(X, f_q)} f(x, q) + \frac{N - k}{|T|} \sum_{x \in T} f(x, q) \tag{11}$$

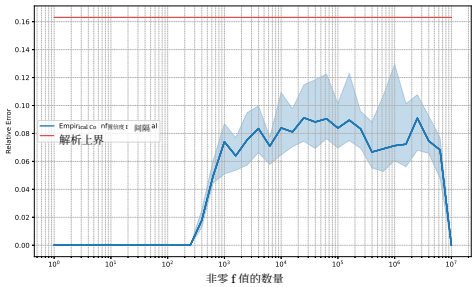
4.5 实验设置

为了评估我们的算法，我们选择 $k \in \{25, 50, 100, 200\}$；对于其他方法，我们选择 $k \in \{250, 500, 1000, 2000\}$ 和 $m \in \{1000, 2000, 5000, 10000, 20000\}$。对于下面的每项任务，我们选择范围较广的任务超参数，以研究在“尖峰” f、“平坦” f及介于两者之间的谱下的性能。结果报告为 30 次查询的中位数。为防止 Qdrant 缓存查询，我们为每种方法抽取一组新的 q。

核密度估计 我们提供了 Open Images 的 ResNet-50 嵌入结果，以及 Amazon Reviews 数据集所选子集的 DistilBERT 嵌入结果。在这两种情况下，我们都会随机抽取一个嵌入作为查询，并在该位置计算核密度估计。



(a) 核密度估计：Open I mages



(b) 计数：Open Images

图 2：(a) 核密度估计在不同带宽下的相对误差中位数，展示了从平坦到尖峰的谱。(b) 我们的解析上界与经验 95% 置信区间的比较，该置信区间使用二项分布 Binomial(100，0。95) 的尾概率计算得出。

Softmax 归一化常数 我们提供了 Open Images 的 CLIP 嵌入结果，其中查询向量是从 Flickr30k 数据集中随机抽取的图像人工标注描述所对应的 CLIP 文本嵌入。根据 CLIP 损失的形式，我们可以将图像的 softmax 概率解释为文本与该图像配对的概率。

Counting 与核密度估计类似，我们给出了 Open Images 的 ResNet-50 嵌入和 d DistAmazon Reviews 的 SentenceBERT 嵌入结果。

4.6 实验结果

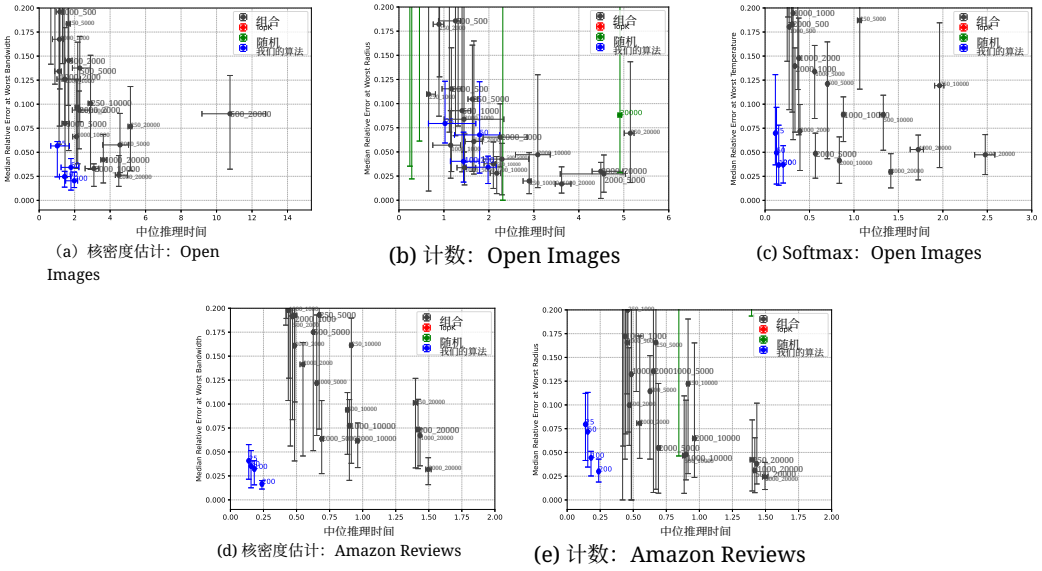
为展示各方法在平坦度/尖锐度方面的性能，我们在图 2(a) 中展示了针对 Open Images 上的核密度估计，四种方法的相对误差中位数随带宽变化的情况。可以看到，随机采样在带宽较大（“平坦”的 f）时表现良好，但在带宽较小（“尖锐”的 f）时效果不佳；top-k 基线则恰恰相反。当 f“平坦”或“尖锐”时，组合方法和我们的方法误差都较低，但在两者之间时误差较高。其他任务的图见附录 C。

为展示我们方法的实用性，我们在图 3 中绘制了 5 个任务/数据集对的相对误差与总运行时间之间的权衡图。由于我们展示的是最差任务参数（平坦度与尖锐度）下的性能，随机采样和 top-k 的相对误差始终约为 1。比较我们的方法与组合方法可以发现，除一种情况外，我们的方法在相同运行时间下误差更低；在该例中，两者误差大致相等（图 3(b)）。请注意，我们方法的运行时间不会随 k线性增长；第 6 节将讨论一种相对直接的方案，将 HNSW 方法与我们提出的方法结合以改善运行时间，并利用 HNSW 同样采用按指数衰减概率生成随机层级这一特点。

最后，为了说明我们的上界，我们展示了合成设置下的结果（见图 2(b)）：其中$f(x、\ q)$要么为 0，要么为 1，并改变产生 1 的 x 数量。由于形式简单，我们可以解析地计算具有完美召回率的 top-k，并且在本实验中不使用 Qdrant。我们使用 $k = 200$ 运行我们的算法 100 次，并针对 95^{th}百分位的相对误差，除了给出第 3.3 节开头的解析上界外，还给出了经验置信区间。与大多数理论上界不同，我们的上界至多只会松弛 2 倍。

5 相关工作

结合 top-k 与随机采样 Mussmann 等（2017）提出了一种结合向量相似性搜索算法与随机采样的简单方法：对最大的 k 个值求和，并通过随机采样估算其余的 $n - k$ 个元素。Mussmann 等（2017）为在给定 $\mathcal{O}\left(\frac{\sqrt{n}}{\epsilon}\right)$ 个 top 元素和 $\mathcal{O}\left(\frac{\sqrt{n}}{\epsilon}\right)$ 个随机样本时的 ϵ 相对误差提供了理论保证。这一思路被应用于深度神经网络中 softmax 函数的学习与推理



（a）核密度估计：Open Images　(b) 计数：Open Images　(c) Softmax：Open Images

(d) 核密度估计：Amazon Reviews　(e) 计数：Amazon Reviews

图 3：比较我们提出的算法与基线方法的权衡图。对于纵轴，我们报告在使相对误差中位数达到最大值的任务参数取值（例如带宽、温度）下的相对误差中位数。随机方法和 TopK 的相对误差中位数接近 1，因此未显示在图中。我们使用二项分布(30，0。5)的尾概率，给出中位数的 95% 置信区间（横轴和纵轴均包括）。

（Levy 等人，2018）提出，近期又被重新发现（Karppa 等人，2022）。相比之下，我们提出的算法只需要$\mathcal{O}\left(\frac{\log n}{\epsilon^2}\right)$个 top 元素。

LSH 采样和基于哈希的估计器 局部敏感哈希（LSH）（Indyk 和 Motwani，1998年；Charikar，2002；Lv 等，2007年；Andoni 和 Indyk，2008年；Andoni 等，2015年）是一类向量相似性搜索技术，或许是唯一具有理论保证（且避免对维度呈指数依赖）的方法，尽管基于聚类的方法和基于邻近图的方法在实证表现上要好得多。对于本文研究的和估计任务，有两篇论文针对两种不同设置提出了结合重要性采样与 LSH 的思路，分别称为 LSH 采样（Spring 和 Shrivastava，2018年）和基于哈希的估计器（Charikar 和 Siminelakis，2017年）。后续关于 LSH 采样的研究将该方法扩展到了更多应用（Spring 和 Shrivastava，2020年；Coleman 和 Shrivastava，2020年）。对于基于哈希的估计器，这一思路已有实现（Siminelakis 等，2019年）和改进（Backurs 等，2019年；Charikar 和 Siminelakis，2019年；Charikar 等，2020年；Qin 等，2022年），包括采用几何采样层级（Charikar 等，2020年）；本文也采用了这一思路。本文的不同之处在于，我们可以使用任意向量相似性搜索方法，不过无法获得任何完整的（非预言机）理论保证。

6 讨论

黑盒最大化预言机方法的局限性 我们在算法设计中使用黑盒最大化预言机的一个局限是，性能保证依据的是检索到的相似向量数量，而非总运行时间。我们选择这一角度，是因为启发式向量相似性搜索技术与具备理论保证的技术之间存在显著的实证差距。此外，向量相似性搜索技术通常是近似而非精确的，其精确率和召回率都低于 100%（我们的实验召回率见附录 C）。由于检索误差取决于许多因素，包括数据的属性和向量相似性搜索技术，为简化起见，我们方法的理论动机和分析假设完美召回率。未来工作可以探索如何对检索误差提出现实的假设，同时仍能为我们的方法提供相对误差保证。

HNSW 与我们方法的整合机会 本文介绍了如何为每个按指数方式采样的层级构建单独的数据结构，并并行执行每个层级的最大化查询。由于 HNSW 方法采用随机分配的层级，且各层级出现的概率按指数递减，并在每次查询执行中迭代地找出每个层级中（近似）最相似的向量，因此稍作修改的分层可导航小世界图算法很可能能够针对单次查询返回每个层级中的相似向量，从而大幅缩短我们方法的总运行时间。

参考文献

Rishabh Agarwal、Max Schwarzer、Aaron Courville 和 Marc G. Bellemare。用于强化学习泛化的对比式行为相似性嵌入。发表于国际学习表征会议 *(ICLR)*，2021年。

Alexandr Andoni 和 Piotr Indyk。用于高维空间近似最近邻的近乎最优哈希算法。*ACM*通信，51(1):117–122，2008年。

Alexandr Andoni、Piotr Indyk、Thijs Laarhoven、Ilya Razenshteyn 和 Ludwig Schmidt。用于角距离的实用且最优的 LSH。神经信息处理系统进展，28，2015年。

Kumar Ashutosh、Rohit Girdhar、Lorenzo Torresani 和 Kristen Grauman。Hiervl：学习分层视频语言嵌入。*arXiv* 预印本 *arXiv:2301.02311*，2023。

Arturs Backurs、Piotr Indyk 和 Tal Wagner。高维空间中高效利用空间和时间的核密度估计。神经信息处理系统进展，32，2019年。

Daniel Cer、Yinfei Yang、Sheng-yi Kong、Nan Hua、Nicole Limtiaco、Rhomni St. John、Noah Constant、Mario Guajardo-Cespedes、Steve Yuan、Chris Tar 等。《通用句子编码器》。载于*2018*年自然语言处理实证方法会议论文集：系统演示，第169–174页。计算语言学协会，2018年。

Matthew Chang、Arjun Gupta 和 Saurabh Gupta。《通过观看 YouTube 视频进行语义视觉导航》。载于神经信息处理系统进展，第2923–2934页，2020年。

Moses Charikar 和 Paris Siminelakis。《基于哈希的估计量用于高维核密度估计》。载于*2017*年*IEEE*第*58*届计算机科学基础年会（*FOCS*），第1032–1043页。IEEE，2017年。

Moses Charikar 和 Paris Siminelakis。《用于快速成对求和的多分辨率哈希》。载于*2019*年*IEEE*第*60*届计算机科学基础年会（*FOCS*），第769–792页。IEEE，2019年。

Moses Charikar、Michael Kapralov、Navid Nouri 和 Paris Siminelakis。通过密度约束近邻搜索进行核密度估计。载于*2020*年*IEEE*第*61*届计算机科学基础年会 *(FOCS)*，第 172–183 页。IEEE，2020年。

Moses S Charikar。基于舍入算法的相似度估计技术。载于第三十四届 *ACM*计算理论研讨会年会论文集，第 380–388 页，2002年。

Benjamin Coleman 和 Anshumali Shrivastava。用于流式数据上近似核密度估计的次线性竞赛草图。载于网络会议 *(WWW)* 论文集，2020年。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：面向语言理解的深度双向变换器预训练。*arXiv* 预印本 *arXiv:1810.04805*，2019年。

Bhavika Devnani、Skyler Seto、Zakaria Aldeneh、Alessandro Toso、Elena Menyaylenko、Barry-John Theobald、Jonathan Sheaffer 和 Miguel Sarabia。学习空间感知语言和音频嵌入。载于神经信息处理系统进展，2024年。

Kacha Dzhaparidze 和 JH Van Zanten。关于鞅的伯恩斯坦型不等式。随机过程及其应用，93(1):109–117，2001。

何恺明、张祥宇、任少卿和孙剑。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，第770–778页，2016年。侯宇鹏、李嘉诚、何展魁、Yan An、陈修思和Julian McAuley。连接语言与项目，以用于检索和推荐。*arXiv*预印本 *arXiv:2403.03952*，2024年。Piotr Indyk 和 Rajeev Motwani。近似最近邻：迈向消除维数灾难。载于第三十届年度*ACM*计算理论研讨会论文集，第604–613页，1998年。Jeff Johnson、Matthijs Douze 和 Hervé Jégou。利用GPU进行十亿规模相似性搜索。*IEEE*大数据汇刊，7(3):535–547，2019年。Matti Karppa、Martin Aumüller 和 Rasmus Pagh。Deann：利用近似最近邻搜索加速核密度估计。载于人工智能与统计国际会议，第3108–3137页。PMLR，2022年。Alina Kuznetsova、Hassan Rom、Neil Alldrin、Jasper Uijlings、Ivan Krasin、Jordi Pont-Tuset、Shahab Kamali、Stefan Popov、Matteo Malloci、Alexander Kolesnikov、Tom Duerig 和 Vittorio Ferrari。Open Images数据集V4：大规模统一图像分类、目标检测和视觉关系检测。国际计算机视觉杂志（*IJCV*），2020年。Nicolas Langrené 和 Xavier Warin。通过快速求和更新实现快速稳定的多变量核密度估计。计算与图形统计学杂志，27(3):468–479，2018年。Daniel Levy、Danlu Chan 和 Stefano Ermon。LSH softmax：深度架构中softmax层的亚线性学习与推理。载于*ICLR*，2018年。吕钦、William Josephson、王哲、Moses Charikar 和李凯。多探针LSH：高维相似性搜索的高效索引。载于第*33*届超大规模数据库国际会议论文集，第950–961页，2007年。Yu. A. Malkov 和 D. A. Yashunin。使用分层可导航小世界图进行高效稳健的近似最近邻搜索。*IEEE*模式分析与机器智能汇刊，42(4):824–836，2020年。Stephen Mussmann、Daniel Levy 和 Stefano Ermon。利用随机扰动最近邻搜索，实现对数线性模型的快速摊销推理与学习。人工智能不确定性会议（*UAI*）论文集，2017年。Emanuel Parzen。概率密度函数与众数的估计。数学统计年刊，33(3):1065–1076，1962年。Matthew E Peters、Mark Neumann、Mohit Iyyer、Matt Gardner、Christopher Clark、Kenton Lee 和 Luke Zettlemoyer。深度上下文化词表示。载于*2018*年北美计算语言学协会分会会议：人类语言技术论文集，第*1*卷（长篇论文），第2227–2237页。计算语言学协会，2018年。William Pugh。跳表：平衡树的一种概率替代方案。*ACM*通信，33(6):668–676，1990年。秦连科、Aravind Reddy、Zhao Song、Zhaozhuo Xu 和 Danyang Zhuo。用于成对求和的自适应动态多分辨率哈希。载于*2022 IEEE*大数据国际会议（大数据），第115–120页。IEEE，2022年。Alec Radford、Karthik Narasimhan、Tim Salimans 和 Ilya Sutskever。通过生成式预训练提升语言理解能力。*OpenAI*，2018年。Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin 等。通过自然语言监督学习可迁移视觉模型。载于第*38*届国际机器学习会议论文集，第8748–8763页。PMLR，2021a。

Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal 等。从自然语言监督中学习可迁移的视觉模型。载于第*38*届国际机器学习会议论文集，第8748–8763页，2021年b。

Nils Reimers 和 Iryna Gurevych。Sentence-BERT：使用孪生 BERT 网络的句子嵌入。载于*2019*年自然语言处理实证方法会议和第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，第3982–3992页。计算语言学协会，2019年。

Victor Sanh、Lysandre Debut、Julien Chaumond 和 Thomas Wolf。DistilBERT：BERT 的蒸馏版本，更小、更快、更便宜、更轻。*arXiv*，abs/1910.01108，2019年。

Paris Siminelakis、Kexin Rong、Peter Bailis、Moses Charikar 和 Philip Levis。重新探讨高维空间中的核评估。载于国际机器学习会议，第5789–5798页。PMLR，2019年。

RyanSpring 和 Anshumali Shrivastava。通过 LSH 采样器（LSS）进行可扩展估计。发表于*ICLR w*研讨会分会，2018年。

Ryan Spring 和 Anshumali Shrivastava。使用 LSH 采样进行互信息估计。发表于*IJCAI*，2020年。

Evgeniya Sukhodolskaya 和 Andrey Vasnetsov。专为向量搜索打造。*Qdrant* 技术博客，2025年。

Joseph Turian、Jordie Shier、Humair Raj Khan、Bhiksha Raj、Björn W. Schuller、Christian J. Steinmetz、Colin Malloy、George Tzanetakis、Gissel Velarde、Kirk McNally 等。HEAR：音频表征的整体评估。发表于神经信息处理系统进展，2022年。

Peter Young、Alice Lai、Micah Hodosh 和 Julia Hockenmaier。从图像描述到视觉指称：用于事件描述语义推理的新型相似度指标。计算语言学协会会刊，2：67–78，2014年。doi：10.1162/tacl_a_00166。网址：https://aclanthology.org/Q14-1006/。

Xiaohua Zhai、Basil Mustafa、Alexander Kolesnikov 和 Lucas Beyer。用于语言图像预训练的 Sigmoid 损失。*arXiv* 预印本 *arXiv:2303.15343*，2023。

A 若干引理的证明

在本附录中，我们证明了定理 3.2 的证明中用到的几个引理。

A.1 高概率界的证明

引理 3.3. 对于任意 δ 和 $k \geq 8\log(1/\delta)$，令 $b = k - \lceil\sqrt{2k\log(1/\delta)}\rceil$。那么，$b \geq k/2 - 1$，且以 $1 - (\ell^* + 3)\delta$ 的概率，

$$\forall \ell : C(\ell, \min(2^\ell b, n)) < k \tag{12}$$

证明。 首先，对于 b 的上界，注意到

$$\begin{aligned}
b - (k/2 - 1) &= k/2 - \lceil\sqrt{2k\log(1/\delta)}\rceil + 1 && (13)\\
&\geq k/2 - \sqrt{2k\log(1/\delta)} && (14)\\
&\geq 4\log(1/\delta) - \sqrt{16\log^2(1/\delta)} && (15)\\
&\geq 0 && (16)\\
& && (17)
\end{aligned}$$

对于高概率界，我们先注意到二项随机变量的一个标准尾界（二项分布的 Chernoff 界，使用 KL 散度的二次下界）：对于 $k \geq np$，$\Pr(\text{二项分布}(n,p) \geq k) \leq \exp(-(k-np)^2/(2k))$。

对于任意 ℓ,

$$\begin{aligned}
\Pr(C_{\ell,\min(2^\ell b,n)} \geq k) &\leq \Pr(C_{\ell,2^\ell b} \geq k) && (18)\\
&\leq \Pr(\text{Binomial}(2^\ell b, 2^{-\ell}) \geq k) && (19)\\
&\leq \exp(-(k-b)^2/2k) && (20)\\
&\leq \exp(-(2k\log(1/\delta))/2k) && (21)\\
&\leq \delta && (22)
\end{aligned}$$

因此，我们可以对前 $\ell^* + 2$ 个层级应用并集界。对于其余层级：

$$\begin{aligned}
\Pr(\exists \ell \geq \ell^* + 3 : C_{\ell,\min(2^\ell b,n)} \geq k) &\leq \Pr\left(\sum_{\ell \geq \ell^*+3} C_{\ell,\min(2^\ell b,n)} \geq k\right) && (23)\\
&\leq \Pr\left(\sum_{\ell \geq \ell^*+3} C_{\ell,n} \geq k\right) && (24)\\
&= \Pr\left(\sum_{\ell \geq \ell^*+3} \text{Binomial}(n, 2^{-\ell}) \geq k\right) && (25)\\
&= \Pr(\text{Binomial}(n, 2^{1-(\ell^*+3)}) \geq k) && (26)\\
&\leq \Pr(\text{Binomial}(n, k/(2n)) \geq k) && (27)\\
&\leq \exp(-(k/2-k)^2/(2k)) && (28)\\
&= \exp(-k/8) && (29)\\
&\leq \exp(-8\log(1/\delta)/8) && (30)\\
&= \delta && (31)
\end{aligned}$$

对前 $\ell^* + 2$ 个层级及其余所有层级应用并集界，我们可以得出 $\Pr(\exists \ell : C(\ell, \min(2^\ell b, n)) \geq k) \leq (\ell^* + 3)\delta$。 □

A.2 关于总条件方差的更紧上界

引理 A.1. $V \leq \frac{3}{8} \frac{F^2}{b}$—

证明。定义

$$z_i = i(f_i - f_{i+1}) \geq 0 \tag{32}$$

对于 $i < n$ 和 $z_n = nf_n$。

首先，我们证明一些有用的引理:

引理 A.2. $f_i = \sum_{j=i}^{n} z_j/j$

证明。 为方便起见，若定义 $f_{n+1} = 0$，则有 $z_i/i = f_i - f_{i+1}$。

$$f_i = \sum_{j=i}^{n} (f_j - f_{j+1}) \tag{33}$$

$$= \sum_{j=i}^{n} z_j/j \tag{34}$$

□

引理 A.3. $F = \sum_{j=1}^{n} z_j$

证明。

$$F = \sum_{i=1}^{n} f_i \tag{35}$$

$$= \sum_{i=1}^{n} \sum_{j=i}^{n} z_j/j \tag{36}$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{j} z_j/j \tag{37}$$

$$= \sum_{j=1}^{n} j z_j/j \tag{38}$$

$$= \sum_{j=1}^{n} z_j \tag{39}$$

□

引理 A.4. 对于任意 α， $\alpha - \frac{2}{3} \leq \frac{3}{8}\alpha^2$

证明。

$$0 \leq \left(\alpha - \frac{4}{3}\right)^2 \tag{40}$$

$$= \alpha^2 - \frac{8}{3}\alpha + \frac{16}{9} \tag{41}$$

$$\frac{8}{3}\left(\alpha - \frac{2}{3}\right) \leq \alpha^2 \tag{42}$$

□

引理 A.5。 如果公式 *3* 成立，那么对于任意 $m \leq n$

$$\sum_{i=1}^{m} 1/p_i - 1 \leq \frac{3}{8}\frac{m^2}{b} \tag{43}$$

证明。 首先我们回顾一下，引理 3.5 表明 $1/p_i - 1 = 0$ 对于 $i \leq 2b$ 和 $1/p_i - 1 \leq 2^{\lceil \log_2(i/2b) \rceil}$ 若$i > 2b$。

如果 $m \leq 2b$，总和为 0，问题就解决了。否则，令 $m = 2^g b\alpha$ 表示 $g = \lfloor \log_2(m/b) \rfloor$ ，并且$1 \leq \alpha < 2$。

$$\begin{aligned}
\sum_{i=1}^{m}(1/p_i - 1) &\leq \sum_{i=2b+1}^{m} 2^{\lceil \log_2(i/2b) \rceil} && (44)\\
&= \sum_{g'=1}^{g-1} \sum_{i=2^{g'}b+1}^{2^{g'+1}b} 2^{\lceil \log_2(i/2b) \rceil} + \sum_{i=2^g b+1}^{2^g b\alpha} 2^{\lceil \log_2(i/2b) \rceil} && (45)\\
&= \sum_{g'=1}^{g-1} \sum_{i=2^{g'}b+1}^{2^{g'+1}} 2^{g'} + \sum_{i=2^g b+1}^{2^g b\alpha} 2^g && (46)\\
&= \sum_{g'=1}^{g-1} \left[2^{g'}b\right] 2^{g'} + \left[(\alpha-1)2^g b\right] 2^g && (47)\\
&= \sum_{g'=1}^{g-1} 4^{g'}b + (\alpha-1)4^g b && (48)\\
&= \frac{4^g - 4}{4-1}b + (\alpha-1)4^g b && (49)\\
&\leq 4^g b\left(\alpha - \frac{2}{3}\right) && (50)\\
&\leq 4^g b\frac{3}{8}\alpha^2 && (51)\\
&= \frac{3}{8}\frac{(2^g b\alpha)^2}{b} && (52)\\
&= \frac{3}{8}\frac{m^2}{b} && (53)
\end{aligned}$$

□

最后，我们可以将所有结果结合起来，

$$V = \sum_{i=1}^{n} (1/p_i - 1) f_i^2 \tag{54}$$

$$= \sum_{i=1}^{n} \sum_{j=i}^{n} \sum_{k=i}^{n} (1/p_i - 1) \frac{1}{jk} z_j z_k \tag{55}$$

$$= \sum_{j=1}^{n} \sum_{k=1}^{n} z_j z_k \frac{1}{jk} \sum_{i=1}^{\min(j,k)} (1/p_i - 1) \tag{56}$$

$$\leq \sum_{j=1}^{n} \sum_{k=1}^{n} z_j z_k \frac{1}{jk} \frac{3}{8} \frac{\min(j,k)^2}{b} \tag{57}$$

$$\leq \frac{3}{8} \frac{1}{b} \sum_{j=1}^{n} \sum_{k=1}^{n} z_j z_k \tag{58}$$

$$= \frac{3}{8} \frac{F^2}{b} \tag{59}$$

□

A.3 伯恩斯坦不等式的鞅版本重述

Dzhaparidze 和 Van Zanten（2001）引用了已有研究，并陈述了标准的 Bernstein 不等式鞅版本：

定理 A.6（伯恩斯坦不等式的鞅版本）。 设 $\{X_i\}_{i=1}^n$ 是关于滤过 $\{\mathcal{F}_i\}_{i=1}^n$ 的鞅差序列，因此 $\mathbb{E}[X_i|\mathcal{F}_{i-1}] = 0$。
若 $|X_i| < M$ 以概率 1 成立，且 $V = \sum_{i=1}^n \mathbb{E}[X_i^2|\mathcal{F}_{i-1}]$ 有限，则，

$$\Pr\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{1}{2}\frac{t^2}{V + Mt/3}\right) \tag{60}$$

由此，我们可以证明以下重述：

定理 3.4。 设 $\{X_i\}_{i=1}^n$ 是关于滤过 $\{\mathcal{F}_i\}_{i=1}^n$ 的鞅差序列，因此$\mathbb{E}[X_i|\mathcal{F}_{i-1}] = 0$。

假设 $|X_i| < M$ 以概率 1 成立，且 $V = \sum_{i=1}^n \mathbb{E}[X_i^2|\mathcal{F}_{i-1}]$ 有限，
那么，对于任意 $\delta > 0$,

$$\Pr\left(\left|\sum_{i=1}^{n} X_i\right| \geq \sqrt{2V\log(2/\delta)} + \frac{2M}{3}\log(2/\delta)\right) \leq \delta \tag{61}$$

证明。 如果我们使用定理 A.6，并令 $t = \sqrt{2V\log(2/\delta)} + \frac{2M}{3}\log(2/\delta)$,

$$\Pr\left(\left|\sum_{i=1}^{n} X_i\right| \geq \sqrt{2V\log(2/\delta)} + \frac{2M}{3}\log(2/\delta)\right) \tag{62}$$

$$\leq 2\exp\left(-\frac{1}{2}\frac{2V\log(2/\delta) + 2\sqrt{2V\log(2/\delta)}\frac{2M}{3}\log(2/\delta) + \frac{4M^2}{9}\log^2(2/\delta)}{V + \frac{M}{3}\sqrt{2V\log(2/\delta)} + \frac{2M^2}{9}\log(2/\delta)}\right) \tag{63}$$

$$\leq 2\exp\left(-\frac{1}{2}\frac{2\log(2/\delta)V + 2\log(2/\delta)\frac{M}{3}\sqrt{2V\log(2/\delta)} + 2\log(2/\delta)\frac{2M^2}{9}\log(2/\delta)}{V + \frac{M}{3}\sqrt{2V\log(2/\delta)} + \frac{2M^2}{9}\log(2/\delta)}\right) \tag{64}$$

$$= 2\exp\left(-\frac{1}{2}2\log(2/\delta)\right) \tag{65}$$

$$= \delta \tag{66}$$

□

B 修正项的超参数选择

首先，我们证明，对伯恩斯坦不等式（定理 3.4）而言，对最小的函数值取平均会得到相同的 M 值。

命题 B.1。 对于任意整数 $a \leq n$，若公式 *3* 成立且 $c = \frac{\sum_{i=(n-a)+1}^{n} f_i}{a}$，则 $\left|\left(\frac{I_i}{p_i} - 1\right)(f_i - c)\right| \leq F/b.$

证明。 若 $f_i \geq c$，则 $|f_i - c| \leq f_i$ 且 $(f_i - c)$ 为正。因此，$\left|\left(\frac{I_i}{p_i} - 1\right)(f_i - c)\right| \leq |Z_i| \leq F/b$（最后一个不等式来自正文）。

如果 $f_i \leq c$，那么 $|f_i - c| \leq c \leq F/n \leq F/i$，我们可以在正文中重复这一论证（证明 $|Z_i| \leq F/b$）。由于 f_i 单调递减，我们知道 $c \leq F/n$。 □

将 V 作为原始和的条件方差（公式 5），将 V_c 作为带有 c 时和的条件方差，

$$V_c = \sum_{i=1}^{n}(1/p_i - 1)(f_i - c)^2 \tag{67}$$

我们找出关于 c 的一个条件，使得 $V_c \leq V$，

命题 B.2. 如果 $c \leq 2\frac{\sum_{i=1}^{n}(1/p_i - 1\ f_i)()}{\sum_{i=1}^{n} 1/p_i - 1}$，那么 $V_c \leq V$。

证明。

$$V - V_c = \sum_{i=1}^{n}(1/p_i - 1)\left[f_i^2 - (f_i - c)^2\right] \tag{68}$$

$$= 2c\sum_{i=1}^{n}(1/p_i - 1)f_i - c^2\sum_{i=1}^{n}(1/p_i - 1) \tag{69}$$

关于 c 的条件意味着

$$c\sum_{i=1}^{n}(1/p_i - 1) \leq 2\sum_{i=1}^{n}(1/p_i - 1)f_i \tag{70}$$

将其乘以 c 后，即可得出结论。 □

表 1：TopK 的召回率（无层级）

Task	数据集	$k = 250$	$k = 500$	$k = 1000$	$k = 2000$
KDE	Open Images	0.9920 ± 0.0047	0.9951 ± 0.0031	0.9972 ± 0.0017	0.9988 ± 0.0010
计数	Open Images	0.9843 ± 0.0020	0.9848 ± 0.0019	0.9846 ± 0.0016	0.9850 ± 0.0018
Softmax	Open Images	0.9223 ± 0.0206	0.9468 ± 0.0142	0.9638 ± 0.0098	0.9764 ± 0.0059
KDE	Amazon Reviews	0.9208 ± 0.0089	0.9231 ± 0.0088	0.9356 ± 0.0086	0.9238 ± 0.0079
计数	Amazon Reviews	0.9684 ± 0.0066	0.9702 ± 0.0058	0.9652 ± 0.0063	0.9596 ± 0.0071

最后，我们可以结合命题 B.1 中 c 的形式和命题 B.2 中的约束，求解 a：

$$\frac{\sum_{i=(n-a)+1}^{n} f_i}{a} \leq 2\frac{\sum_{i=1}^{n}(1/p_i - 1)f_i}{\sum_{i=1}^{n}(1/p_i - 1)} \tag{71}$$

$(1/p_i - 1)$ 大致按 i 的比例缩放，因此我们可以求解以下方程：

$$\frac{\sum_{i=(n-a)+1}^{n} f_i}{a} \leq 2\frac{\sum_{i=1}^{n} i f_i}{\sum_{i=1}^{n} i} \tag{72}$$

$$\frac{n(n+1)}{4a}\sum_{i=1}^{n}\mathbf{1}[i \geq (n-a)+1]f_i \leq \sum_{i=1}^{n} i f_i \tag{73}$$

当且仅当对所有可能的 f_i 都成立时，

$$\frac{n(n+1)}{4a} \leq (n-a)+1 \tag{74}$$

$$n(n+1) \leq 4a(n+1) - 4a^2 \tag{75}$$

$$n(n+1) \leq -(2a-(n+1))^2 + (n+1)^2 \tag{76}$$

$$(2a-(n+1))^2 \leq n+1 \tag{77}$$

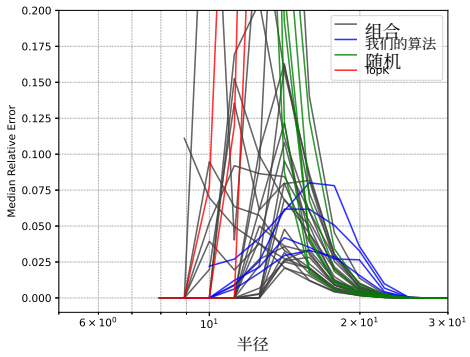
如果 $a = \left\lfloor \frac{n+1}{2} \right\rfloor = \left\lceil \frac{n}{2} \right\rceil$，我们就能将左侧最小化并满足约束。

因此，我们将 c 设为较小一半 $\{f_i\}_{i=1}^{n}$ 值的平均值。

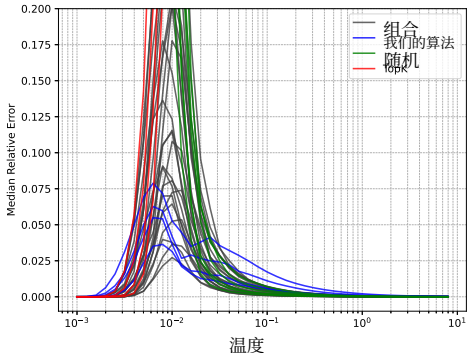
C 完整实验结果

这里，我们展示其余的实验结果。除图像核密度估计（图 2(a)）之外的设置中，相对误差中位数随设置参数变化的情况如图 4 所示。请注意，随机方法在平坦情况下表现最佳，而 TopK 在尖峰情况下表现最佳。

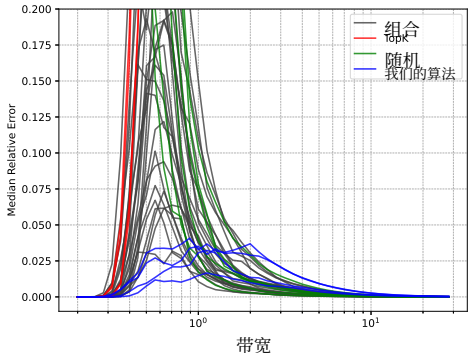
图 5 展示了 k 在不同层级下若干取值对应的召回率。我们仅绘制该层级点数大于 k 的层级。误差棒表示两个标准误。图 1 展示了不分层级时 TopK 的召回率，并同样标出了正负两个标准误。



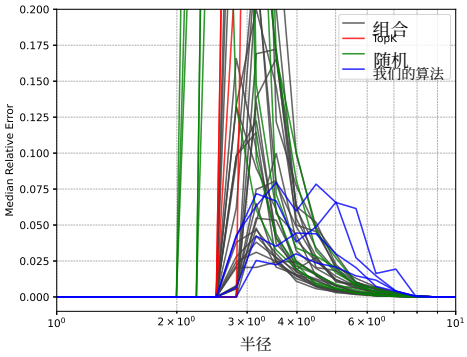
(a) 计数：Open Images



(b) Softmax：Open Images



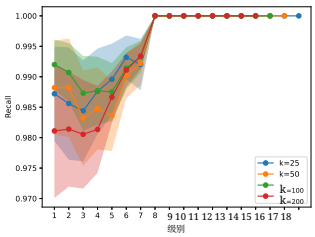
(c) 核密度估计：Amazon Reviews



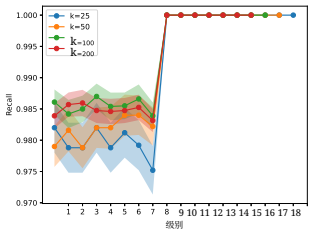
(d) 计数：Amazon Reviews

图 4：不同设置下，相对误差中位数随设置参数（带宽、半径、温度）的变化，用以展示从平坦到尖峰的谱。对于计数任务，当超过 75% 的运行中出现 $F = 0$，导致相对误差无法确定时，不绘制相应曲线。其他情况下，中位数仅根据出现 $F \neq 0$ 的运行计算。

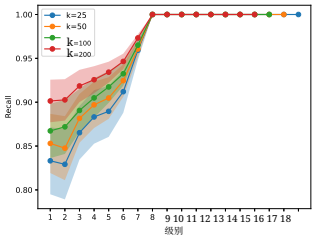
.



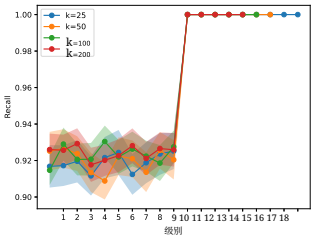
(a) 核密度估计：Open I
mages



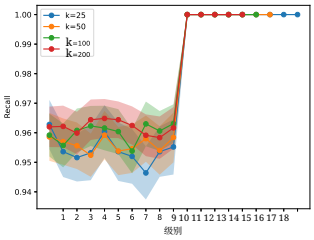
(b) 计数：Open Images



(c) Softmax：Open Images



(d) 核密度估计：Amazon Reviews



(e) 计数：Amazon Reviews

图 5：不同设置下，不同 k 值对应的召回率随级别变化的函数图。我们仅绘制该级别上的点数大于 k 的层级。误差棒表示两个标准误。