针对一种基本计算问题的神经算法

姓名：王迎旭
学号：16340226
班级：16级软工教务三班

相似性搜索是大型信息检索系统需要处理的一个最基本的计算问题，例如，在数据库中找出相似的信息或辨识出web服务器端相似的文献。我们在搜索中发现果蝇嗅觉回路可以通过使用一个计算机科学算法中的变体（也被称为局部敏感哈希）来处理这个问题。果蝇网络将相似的神经活动模式归类到与其类似的气味上，以便于以后在遇到相似气味的时候，可以应用某种已经从气味中学习到的行为活动。然而，苍蝇算法使用的三种计算策略却与传统算法有些许差别。这些策略可用于提高计算相似度搜索的性能。以上所涉及的观点有助于理清逻辑进而支持下文所要论述的重要感官功能，同时以上观点也为解决一个计算问题提供一个了概念上的新算法。

许多神经回路的基本任务就是去给输入的刺激分配对应的神经活动模式，从而尽可能准确的识别不同输入。我们研究了果蝇嗅觉系统中用来处理气味的回路，并就此设计了一种解决基本机器学习问题的最新计算方法（也即:近邻相似性(或最近邻)搜索）。

果蝇嗅觉回路会为每一种气味分配一类标签，每一类标签又分别对应于一组神经元，当这种气味出现时，与标签对应的神经元就会被激活。这个标签在果蝇接受不同气味信息，并作出应有行为反应方面是至关重要的。例如，一种激励（例如糖水）或者一种惩罚（例如电击）与某一种气味相关联，这种气味就会变得对果蝇具有吸引性（果蝇会接近这种气味）或者具有排斥性（果蝇会避开这个气味）。被分配到气味的标签可以被认为是稀疏的——接收到气味信息的神经元中只有一小部分会对众多的气味做出相应的响应，同时作出反应的神经元之间也并不重叠——也即任意两种不同气味的标签几乎不会去共享（如果存在）相同活跃程度的神经元，因此不同的气味可以很容易被区分开。

气味的标签是通过一个被分为三个步骤的程序生成的(如图1A所示)。第一步是建立从果蝇鼻中气味受体神经元到肾小球结构投射神经元的前馈连接。这个前馈连接之中有50种不同的气味受体神经元，每种气味受体对不同的气味有独立的灵敏性和选择性。于此，每一种输入气味都在50维空间中有一个位置，这个空间是由50个气味受体神经元的激活率构建的。每一种气味的受体神经元激活率分布在50种气味受体神经元中，神经元激活率的平均值取决于气味的浓度并呈指数型增长。然而，投射神经元比较特殊，这种神经元并没有依赖特性，也就是说，这50种投射神经元的激活率分布呈指数型，且投射神经元可接受气味的浓度都接近一个相同的平均值。因此，构建回路的第一步本质上是使用“分裂化归一”的方法，将待处理的数据平均值居中，这也是繁杂的流水线计算问题中的一个规范化预处理方法。这一步保证了果蝇不会把气味的强度和气味的类型混淆在一起。

第二步，算法主框架开端。这一步主要是对神经元数目上进行40倍的扩展:也即将50个投射神经元投影到2000个凯尼恩细胞上，并通过一个稀疏的二进制随机矩阵将彼此之间联系起来。每一个凯尼恩细胞都接收并对随机选择的六个气味受体神经元的激活率进行求和。第三步是建立APL单一抑制神经元强烈抑制反馈WTA回路。这就导致所有的凯尼恩细胞中激活率最高的5%都被灭活了。余下的5%拥有较高激活率的凯尼恩细胞则会与输入气味的标签相对应。

从计算机科学的角度来看，我们可以把果蝇的回路看作是一个哈希函数，这个函数的输入是一种气味，函数的输出是与某个气味相对应的标签(也称为哈希)。尽管这个标签已经可

以用于区分气味，但果蝇可以进一步发挥自身优势把相似的气味与之对应的标签相串联(图1B)，这也就使得对于某一种气味的学习可以应用于与之非常相似的气味或者容易受到干扰的气味之上。上面描述的现象客观上引导我们进行一定的推测，果蝇电路产生的标签是局部且敏感的，两种气味越相似（由50个气味感官神经元对该气味的激活率定义），这两种气味被分配的标签就越相似。局部灵敏哈希(LSH)是解决计算机科学中众多相似搜索问题的基础。基于以上的分析，我们将已经掌握的对果蝇回路的见解应用于开发LSH算法，并通过使用这个算法去高效地找出高维度点的近似点。

假设，现在为你提供了一张大象的图片，并让你试图从网上的数十亿张图片中找到100张与之十分相似的照片。这种搜索相似图片的行为被称为最近邻搜索问题，这种行为在信息检索、数据压缩和机器学习方面具有重要的意义。每幅图像通常被表示为d维向量的特征值（每一只果蝇可以识别的气味都是一个50维特征向量的激活率）。使用距离度量来计算两个图像之间的相似性(特征向量),这样做的目的也是去有效地找到任何与被查询图像最近接的图像。如果Web只包含很少的图像，那么使用暴力线性搜索就可以很容易地找到与目标图像相似性最大的图像。如果Web含有许多图像,但是每个图像都由一个低维度的向量表示(例如只有10或20个特征)，那么空间分区方法也就足够适用于解决问题。然而，在处理某个有高维度数据的大型数据库时候，这两种方法都并不适用。

在许多应用程序中，为了能够尽可能快的完成查询，系统会选择返回接近于查询结果的一个邻近集合。这种特性研发出一种新方法，也即使用局部敏感哈希(LSH)技术来寻找最近邻查询的集合。对于果蝇来说，气味的标签(或哈希)与该气味的凯尼恩细胞激活率向量相对应。局部敏感特性表明,两种相似(如甲醇和乙醇)的气味将由两个自身相似的标记来表示(如图1B)。同样地，针对图像搜索而言，大象图像的标记与另一个大象图像的标记有许多共同点，而不是摩天大楼图像的标记有许多共同点。

传统的(非LSH)哈希函数的输入点在定义域范围内呈随机且均匀的分布，LSH函数与传统哈希函数不同，LSH函数提供了一个从d维空间到m维空间(后者对应的标签)的点距离维护。因此，在输入空间中距离较近的点被分配相同或相似的标记的概率要比相隔很远的点的概率高。

为了设计一个LSH函数,一个通用的方法是计算输入数据的随机投影也即将输入的特征向量乘上一个随机矩阵。The Johnson-Lindenstrauss引理和它的许多变形推导为我们使用各种各样随机投影将数据从d维映射到m维时,可以极大程度保存局域相对结构提供了坚实的理论支持。

比较新奇的地方是：果蝇能够通过随机投影（例子1A中的步骤2，将50投射神经元映射到2000个凯尼恩细胞)为气味分配标签,这也就为实现这部分回路的功能提供一个关键支持。如上文所言，果蝇算法与传统的LSH算法有三个不同之处。第一，果蝇使用的是稀疏、二进制的随机投影，而LSH函数通常使用预测成本比较高昂的密集型高斯随机投影。第二，果蝇在完成投影之后扩展了输入的维度，而LSH算法则收缩了输入的维度。第三，果蝇回路通过使用“WTA”机制使高维度表示变得稀疏，而LSH则仍旧保持着稠密表示法。

在补充材料中,我们通过分析发现果蝇嗅觉回路中的稀疏二元随机投影被分配对应的标签，并且使用这些标签来保留输入点的邻域结构。这就表明，果蝇回路是局部哈希族中我们尚未发现的一种。

我们以每个算法如何精确识别给定查询点的最近邻点为依据,进而对果蝇回路算法与传统LSH进行评估。为了保证两种算法比较的公平性，我们将两种算法的计算复杂度固定为相同（图1C）。也就是说，两种方法固定使用相同数量的数学运算来为每个输入生成相同长度k的哈希（即带有k个非零值的向量）。

我们在三个基准数据集（SIFT（d = 128），GLOVE（d = 300）和MNIST（d = 784））

中比较了两种寻找最近邻算法，SIFT 和 MNIST 包含用于图像相似性搜索的图像的矢量表示，而 GLOVE 则包含用于语义相似性搜索的关键词的矢量表示。我们收集了每个数据集的一个子集，每个数据集都有 10000 个输入，其中每个输入都可以表示为 d 维空间中的一个特征向量。为了测试性能，我们从 10000 输入中选择了 1000 个随机输入，并将真实值和预期值的最近邻进行了比较。也就是说，进行每一次查询时候，我们根据特征向量之间的欧几里得距离，在输入空间中找到了与其真实近邻的前 2%(200)个体。我们对哈希(k)的长度进行改变，并利用中间平均精度计算出了排序表上真值和预测值最近邻之间的重叠部分。我们在 50 次试验中都先平均了中间平均精度，因此每一次试验中的随机投影矩阵和查询输入都会被改变。我们首先区分出果蝇回路算法和哈希算法之间的三个不同点，进而测试这两种算法对最近邻检索性能的个体影响。

用稀疏的二进制随机来代替 LSH 的稠密的高斯随机投影并不影响所搜寻到的最近邻的精确性。这些结果支持我们的理论计算，即果蝇的随机投影确实是对局部敏感的。此外，稀疏的二元随机投影比稠密的高斯随机投影的计算量节省了将近 20 倍。

在扩展了维度的过程中，用 WTA 方法来简化标签比用随机选择标签的成效要好。WTA 选择了前 k 个激活性能比较好的凯尼恩细胞作为标签，而不是像另一种随机选择了 k 个凯尼恩细胞作为标签。对于两者，我们都用了 20k 个随机投影，将两个算法所使用的数学运算的数量等同起来。例如：对于哈希长度 k=4 的 SIFT 数据集来说，随机选择产生了 17.7%的平均精度，而使用 WTA(32.4%)的平均精度大约是随机选择的两倍。因此，选择激活性能靠前的神经元可以很好保持输入之间的相对距离，同时维度的增加也使得分离不同的输入变得更加容易。对于随机标签的筛选，我们选择了 k 个随机的(但所有输入都是固定的)凯尼恩细胞；因此，它的性能与只做 k 个随机投影是完全相同的，就像在 LSH 中一样。

在更进一步的维度扩展中（从 20k 到 10d 凯尼恩细胞，更接近于实际的果蝇电路），我们在识别数据集和最近邻哈希长度方面获得了比 LSH 更大的收益。在非常小的哈希长度中的收获最为重大，平均精准度几乎提高了三倍。(例如：对于 MNIST 来说，k=4，LSH 是 16.0%，而果蝇回路则是 44.8%）。

当测试更高维的果蝇算法和二进制 LSH 时，我们也发现了性能有类似的提升。因此，果蝇算法是可拓展的，并且可能比其他哈希族算法更加有效。

我们的工作确立了大脑相似性匹配策略和大型信息检索系统中最近邻搜索方法之间的协同作用。我们的工作也可以应用于重复检测、聚类和高效的深度学习。LSH 有大量的拓展，例如：使用多个哈希表来提高精度(上文中，我们的两个算法仅使用一种哈希表)，使用多探针将类似的标签划分到一组(因为标签是稀疏的，对于果蝇来说可能更容易实现)，使用各种应用于离散哈希的技巧。同时，还有一些方法可以加速随机投影乘法，例如，应用于 LSH，使用快速的 Johnson-lindenstrauss 变换；应用于果蝇，使用快速稀疏矩阵乘法。我们的目标是比较两个概念上不同的方法来解决最近邻的搜索问题；在实际应用中，所有这些扩展都需要移植到果蝇算法中。

果蝇算法中的许多策略已经被应用过。例如，尽管不建议使用者扩展维度，但是 MinHash 和 winner-take-all 哈希都使用了类似的 WTA 组件。在 LSH 算法大家族中，类似的许多算法也是用到了随机投影，但据我们所知，没有一个 LSH 算法使用到稀疏二进制投影。果蝇回路似乎已经进化到可以把不同的计算部分结合起来。有证据表明，用于果蝇回路的三种特性也出现在果蝇大脑的其他区域和其他物种身上。因此，局部灵敏哈希可能是一个在大脑中使用的一般计算原则。

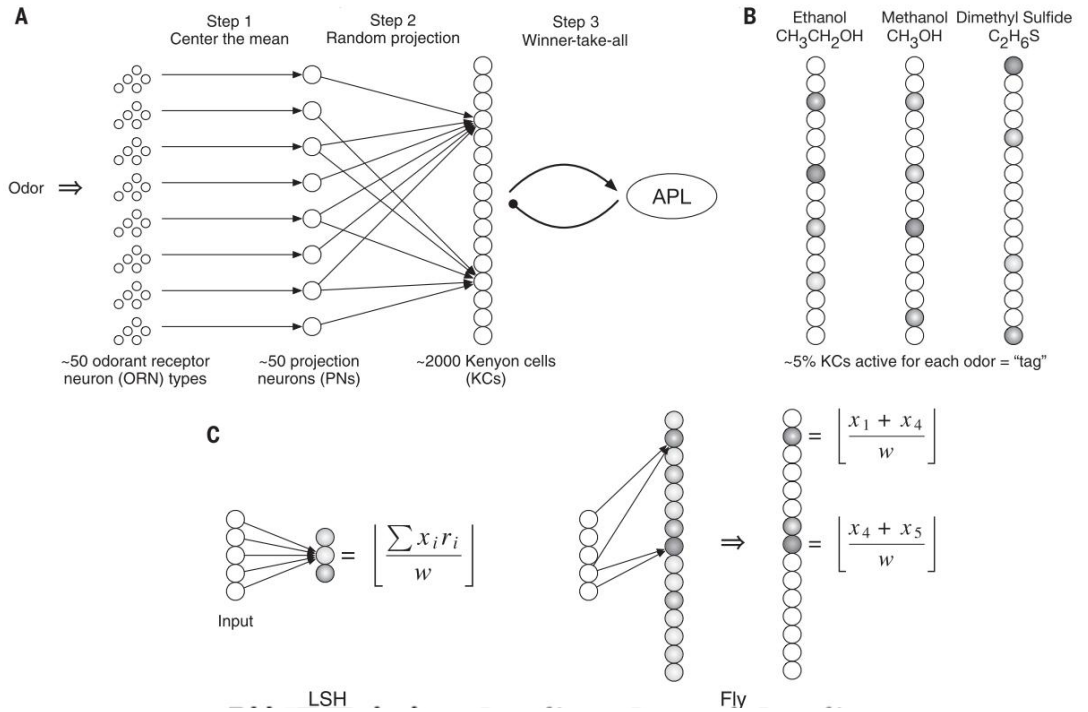

图 1：果蝇嗅觉回路和局部敏感哈希之间的映射。

A）：果蝇嗅觉回路原理图。第一步，果蝇鼻子里的 50 个气味受体神经元把轴突送到肾小球里的 50 个投射神经元上；这样投影以后，每一种气味被表示为激活率的指数分布，所有气味和所有气味浓度有相同的均值。第二步，投射神经元进行了维度扩展，即通过连接一个稀疏二进制随机投影矩阵投影到了 2000 个凯尼恩细胞上；第三步，凯尼恩细胞从前对侧(APL)神经元中获得反馈抑制，这之后只会剩下前 5%的凯尼恩细胞继续保持对气味的应激机制。这 5%凯尼恩细胞对应于气味的标签(哈希)。

B）：气味反应说明。类似的气味(如甲醇和乙醇)的标签比不同的气味更相似。较暗的阴影表示较高的活性。

C）：传统 LSH 和果蝇算法之间的区别。在这个例子中，LSH 和苍蝇的计算复杂度是一样的。输入维数 d=5，随后使用 LSH 计算 m=3 随机投影，每一次投影都需要 10 次操作(5 次乘法+5 次法)。果蝇算法则计算 m=15 随机投影，每一次投影需要 2 次加法运算。因此，两者都需要 30 次操作。

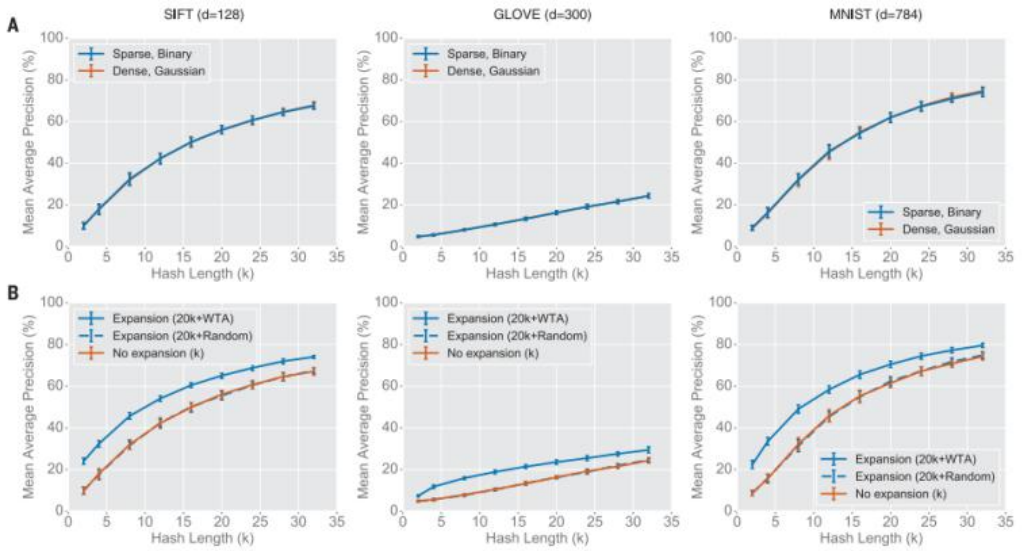

图 2：不同随机投影类型和标签选择方法的实验比较。在所有的图中，x 轴是哈希的长度，而 y 轴是中间平均精度(越高越好)。

A)：稀疏的二元随机投影提供了与稠密的高斯随机投影几乎相同的性能，但是前者节约了大量的计算。

B)：维度扩展(从 k 到 20k)加上 WTA 稀疏化，相比没有扩展的来说，这种做法进一步提高了性能。所有三个基准数据集的结果是一致的。误差条表示标准偏差超过 50 次试验。

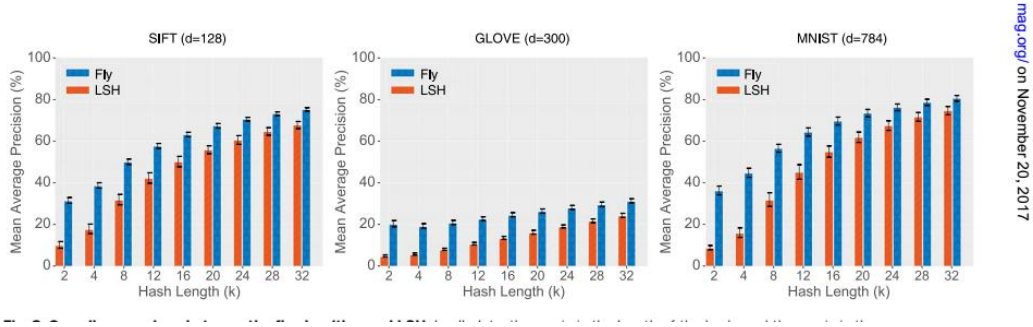

图 3：果蝇算法和局部敏感哈希算法的总体比较。在所有的图中，x 轴是哈希的长度，而 y 轴是中间平均精度(越高越好)。10d 的扩展用于果蝇。在所有三个数据集中，果蝇算法性能优于 LSH，最显著的是较短的哈希代码。误差条表示标准偏差超过 50 次试验。

表一：大脑中局部敏感哈希的普遍性。下面所展示的是果蝇嗅觉回路中所使用的步骤以及在脊椎动物中的深层潜在模拟。

	Step 1	随机投影	Step 2	Step 3
果蝇嗅觉	触角神经叶；50 个肾小球	稀疏的，二进制投影；采样 6 个肾小球	蕈状体；2000 个凯尼恩细胞	APL 神经元；前 5%
宠物鼠嗅觉	嗅球；1000 个肾小球	稠密的，羸弱的；采样来自所有的肾小球	梨状皮质；100000 个半月形细胞	2A 层；前 10%
老鼠嗅觉	小脑前核	稀疏的，二进制投影；采样 4 个小脑前核	颗粒细胞层；250 百万个颗粒细胞	高尔基体；前 10%~20%
大鼠海马	内嗅皮层	未知	齿状回；1.2 百万个颗粒细胞	门细胞；前 2%