基于语义-视觉相似度增强的深度离散哈希

合肥工業大學

杨宇轩

2022.11.27

Yang Z, Yang L, Huang W, et al. Enhanced Deep Discrete Hashing with semantic-visual similarity for image retrieval[J]. Information Processing & Management, 2021, 58(5): 102648.

提　纲

01 研究背景

02 研究方法

一 研究背景

最近邻（Nearest Neighbor, NN）搜索算法在信息检索领域发挥着重要作用。然而，随着收集的图像数据量迅速增加，最近邻搜索的时间复杂度呈指数型增长。为此，近似最近邻（Approximate Nearest Neighbor, ANN）搜索更适合用于大规模检索任务。

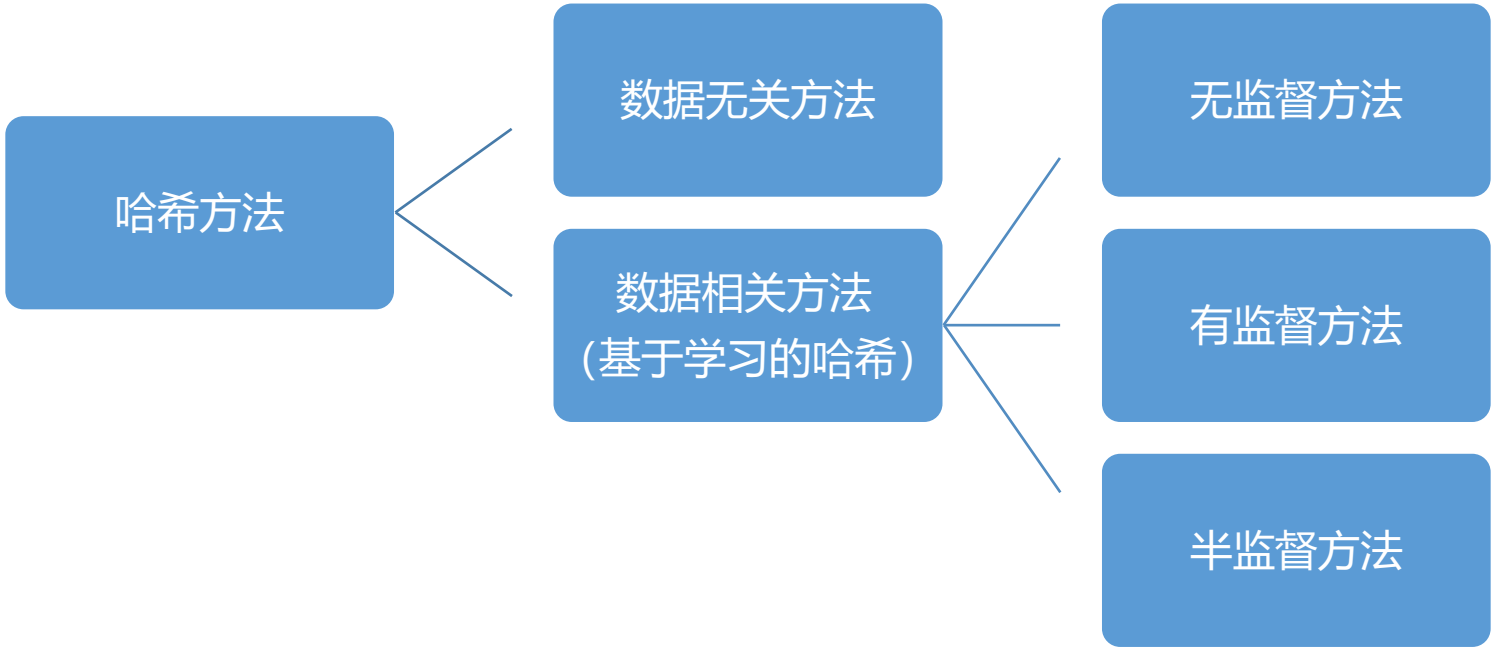
其中，哈希方法具有检索速度快、存储成本低的特点，受到了广泛研究。一般而言，哈希函数能够将高维特征映射为低维二进制编码，即：

$$\boldsymbol{b_i} = F(\boldsymbol{x_i}) \in \{-1, +1\}^k$$

其中$F(\cdot)$表示哈希函数，$\boldsymbol{x_i}$表示d维特征，$\boldsymbol{b_i}$表示k比特二进制哈希码。

一 研究背景

哈希方法主要分为数据无关的方法和数据相关的方法。其中，数据无关的方法效率非常低，需要较长哈希码（通常高于1024bits）才能达到良好的精度。数据相关的方法可以分为无监督方法、有监督方法和半监督方法。近年来，由于深度神经网络能够直接提取图像的内容语义，显著提高检索精度，深度哈希成为主流方法。



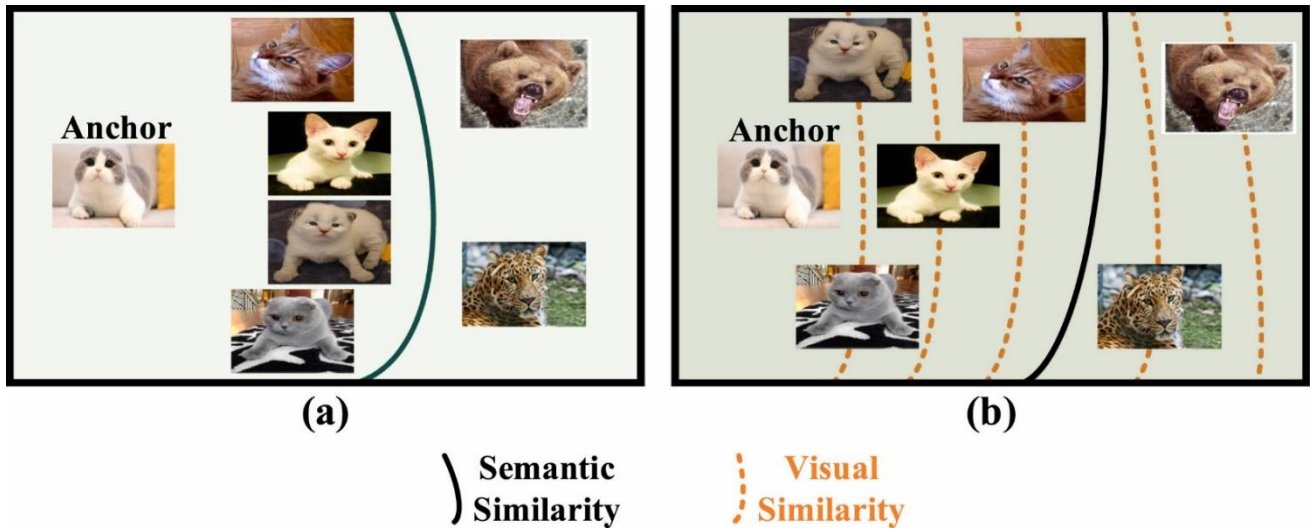
一 研究背景

然而，现有的深度哈希方法仍存在以下问题：

（1）大部分深度哈希方法只考虑图像对的二元相似信息，即当两个图像$\{x_i, x_j\}$具有同一标签类别时，相似度值$s_{i,j} = 1$，否则$s_{i,j} = 0$。然而，这种粗略的相似性无法反映不同图像之间的视觉关联，导致哈希方法存在视觉相似性缺失的问题。

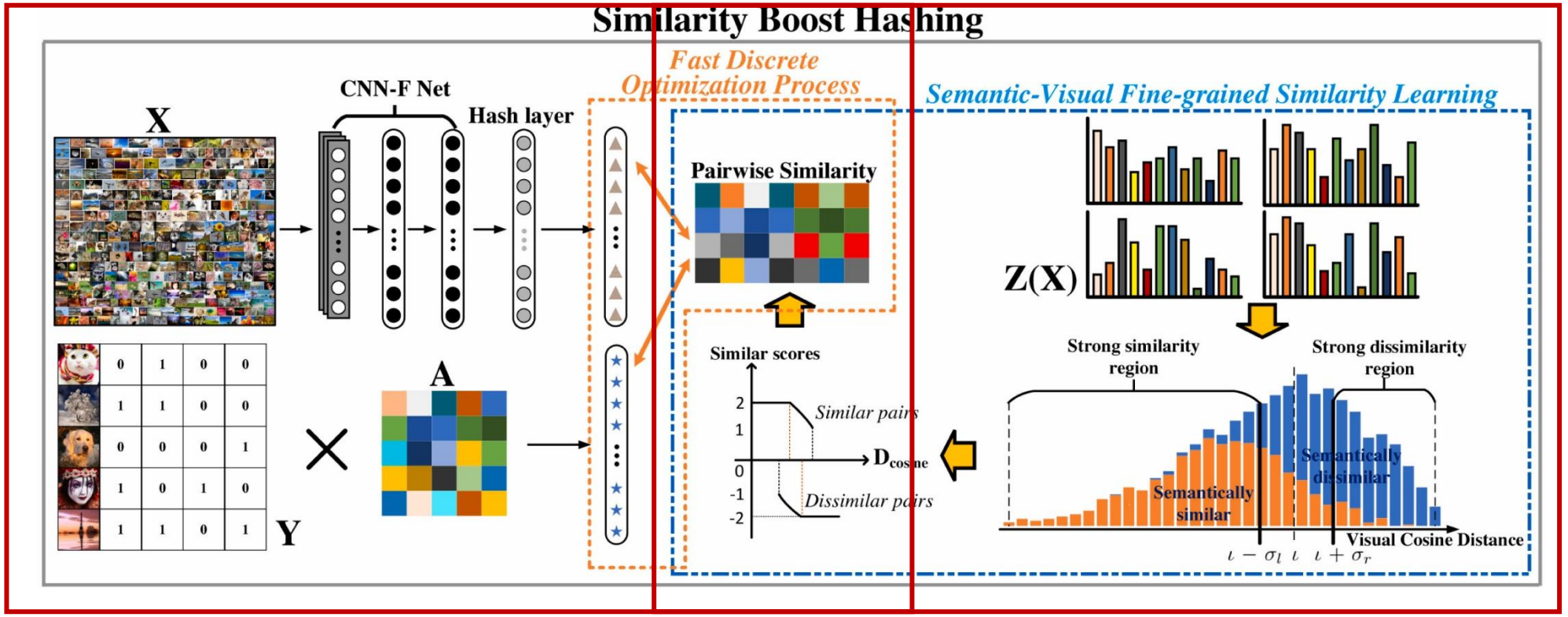
（2）部分哈希方法以对称形式处理离散约束，哈希码的学习效率低于非对称方式的学习。

（3）大多数深度哈希方法利用松弛策略（如tanh(·)函数）来解决二进制离散优化问题，然而这会引入较大的量化误差。



一 研究背景

本研究设计了一种用于图像检索的增强深度离散哈希（Enhanced Deep Discrete Hashing, EDDH）方法。具体而言，针对问题（1），EDDH考虑标签信息引导下的视觉相似性关系，为此构建了语义-视觉连续相似性。针对问题（2），EDDH基于非对称学习框架，以此逼近语义-视觉连续相似性。针对问题（3），EDDH采用离散优化方法求解。



二 研究方法

问题定义

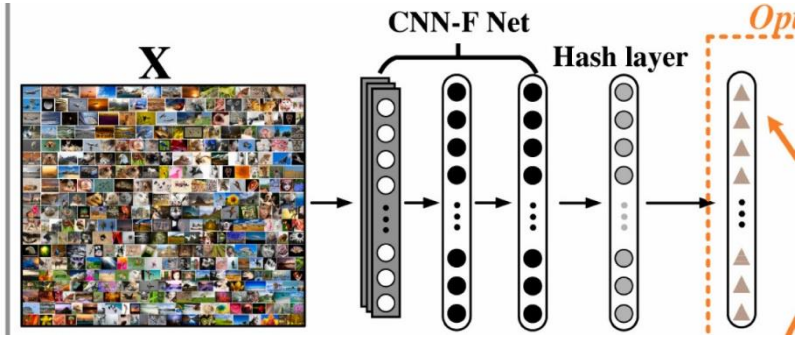
给定训练集$\boldsymbol{X} = \{\boldsymbol{x_1}, \dots, \boldsymbol{x_n}\} \in \mathbb{R}^{d \times n}$，包括n张图片，每张图片用d维向量表示。对应的类别标签矩阵$\boldsymbol{Y} = \{\boldsymbol{y_1}, \dots, \boldsymbol{y_n}\} \in \{0,1\}^{c \times n}$，$c$为类别数。成对相似性矩阵$\boldsymbol{S} \in \{-1, +1\}^{n \times n}$由标签矩阵$\boldsymbol{Y}$得到，即当$\boldsymbol{x_i}$和$\boldsymbol{x_j}$至少有一个共同标签时，$s_{ij} = 1$，否则$s_{ij} = 0$。

EDDH的目标是学习一个哈希函数$\mathcal{F}(\cdot)$来生成用于检索的哈希码$\boldsymbol{B} = [\boldsymbol{b_1}, \boldsymbol{b_2}, \dots, \boldsymbol{b_n}] \in \{-1,1\}^{k \times n}$，长度为$k$比特。网络输出的近似二值哈希码表示为$\boldsymbol{B}_*$，即$\boldsymbol{B}_* = \mathcal{F}(\boldsymbol{X})$，$\boldsymbol{B}$则可以由$\boldsymbol{B} = sign(\boldsymbol{B}_*)$得到。

图片表示学习

采用在ImageNet上预训练的CNN-F[1]网络作为特征提取器，并在网络最后添加了一个线性哈希层和tanh激活函数，使得输出的近似哈希码值在$(-1,1)$之间。

Arch.	conv1	conv2	conv3	conv4	conv5	full6	full7	full8
CNN-F	64x11x11 st. 4, pad 0 LRN, x2 pool	256x5x5 st. 1, pad 2 LRN, x2 pool	256x3x3 st. 1, pad 1 -	256x3x3 st. 1, pad 1 -	256x3x3 st. 1, pad 1 x2 pool	4096 drop- out	4096 drop- out	1000 soft- max



[1] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: Delving deep into convolutional nets[C]//Proceedings of the British Machine Vision Conference. BMVA Press, 2014.

二 研究方法

相似性保持

首先，经典的成对相似性哈希目标函数如下：

$$\min_{\boldsymbol{B}} \|\boldsymbol{B}^T\boldsymbol{B} - k\boldsymbol{S}\|_F^2 \quad s.t. \boldsymbol{B} \in \{-1,1\}^{k\times n}$$

上式最小化了基于输出哈希码的相似度和基于标签的相似度之间的差异。然而，上式的主要问题在于其采用了对称矩阵分解的形式。为此，可以用图像的深度特征代替第一个$\boldsymbol{B}$，由此可得：

$$\min_{\theta,\boldsymbol{B}} \left\| sign\left(\mathcal{F}(\boldsymbol{X};\theta)\right)^T\boldsymbol{B} - k\boldsymbol{S}\right\|_F^2 \quad s.t. \boldsymbol{B} \in \{-1,1\}^{k\times n}$$

由于二值约束难以求解，且采用实值可以显著提升相似度保持的能力，因此，本研究直接去掉$sign(\cdot)$函数，并且鼓励$\mathcal{F}(\boldsymbol{X};\theta)$近似为二值：

$$\min_{\theta,\boldsymbol{B}} \|\mathcal{F}(\boldsymbol{X};\theta)^T\boldsymbol{B} - k\boldsymbol{S}\|_F^2 + \lambda\|\mathcal{F}(\boldsymbol{X};\theta) - \boldsymbol{B}\|_F^2 \quad s.t.\ \boldsymbol{B} \in \{-1,1\}^{k\times n}$$

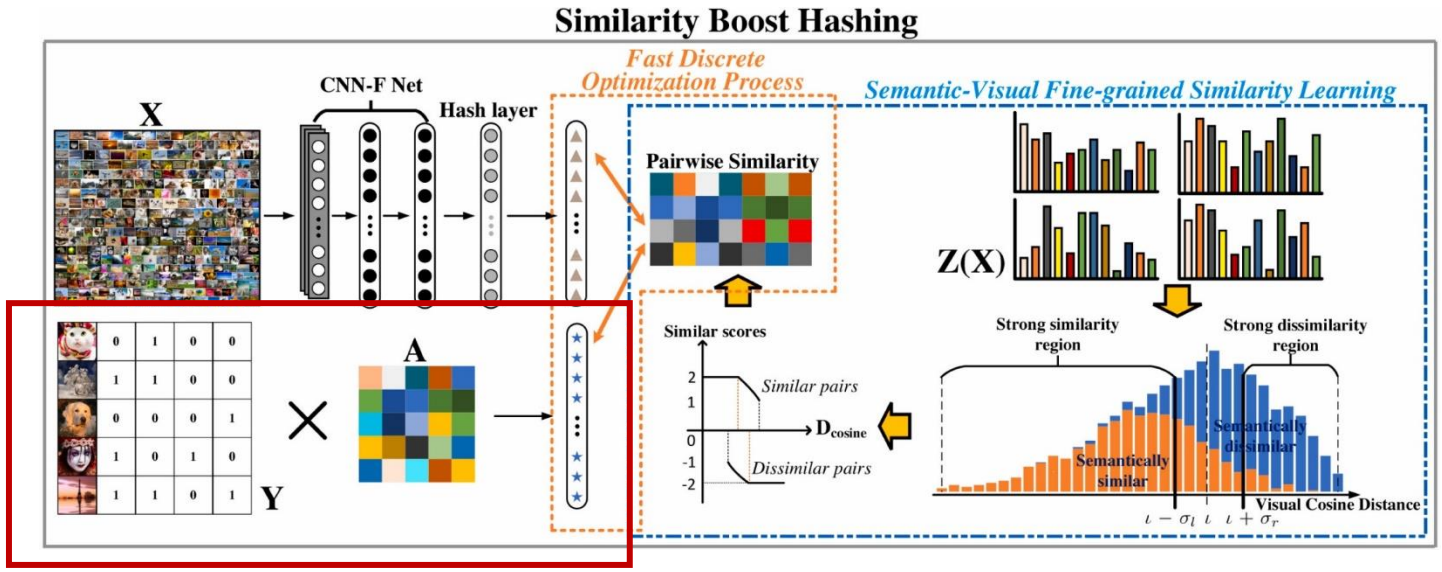
二 研究方法

语义标签迁移

学习到的哈希码不应仅包含图像的特征，还应该包含丰富的语义。已有研究表明，将标签信息直接嵌入哈希码中可以大大提高哈希码的判别能力。在本研究中，给定语义标签$\boldsymbol{Y}$，将其映射为学习的哈希码。

具体而言，引入变换矩阵$\boldsymbol{A} = [\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_k}] \in \mathbb{R}^{c \times k}$，其中$\boldsymbol{a_i} \in \mathbb{R}^{c \times l}$为哈希码第 i 比特的语义变换向量，目标是最小化哈希码与标签迁移得到的实值语义矩阵之间的差异：

$$\min_{\boldsymbol{A}} \|\boldsymbol{B} - \boldsymbol{A}^T \boldsymbol{Y}\|_F^2$$



二 研究方法

语义标签迁移

本研究进一步发展非对称学习框架，用$\boldsymbol{A}^T\boldsymbol{Y}$替代第二个$\boldsymbol{B}$，目标函数可进一步改写为：

$$\min_{\theta,\boldsymbol{B},\boldsymbol{A}} \|\mathcal{F}(\boldsymbol{X};\theta)^T\boldsymbol{A}^T\boldsymbol{Y} - k\boldsymbol{S}\|_F^2 + \lambda\|\mathcal{F}(\boldsymbol{X};\theta) - \boldsymbol{B}\|_F^2 + \gamma\|\boldsymbol{B} - \boldsymbol{A}^T\boldsymbol{Y}\|_F^2 + \rho\|\boldsymbol{A}^T\boldsymbol{Y}\|_F^2$$

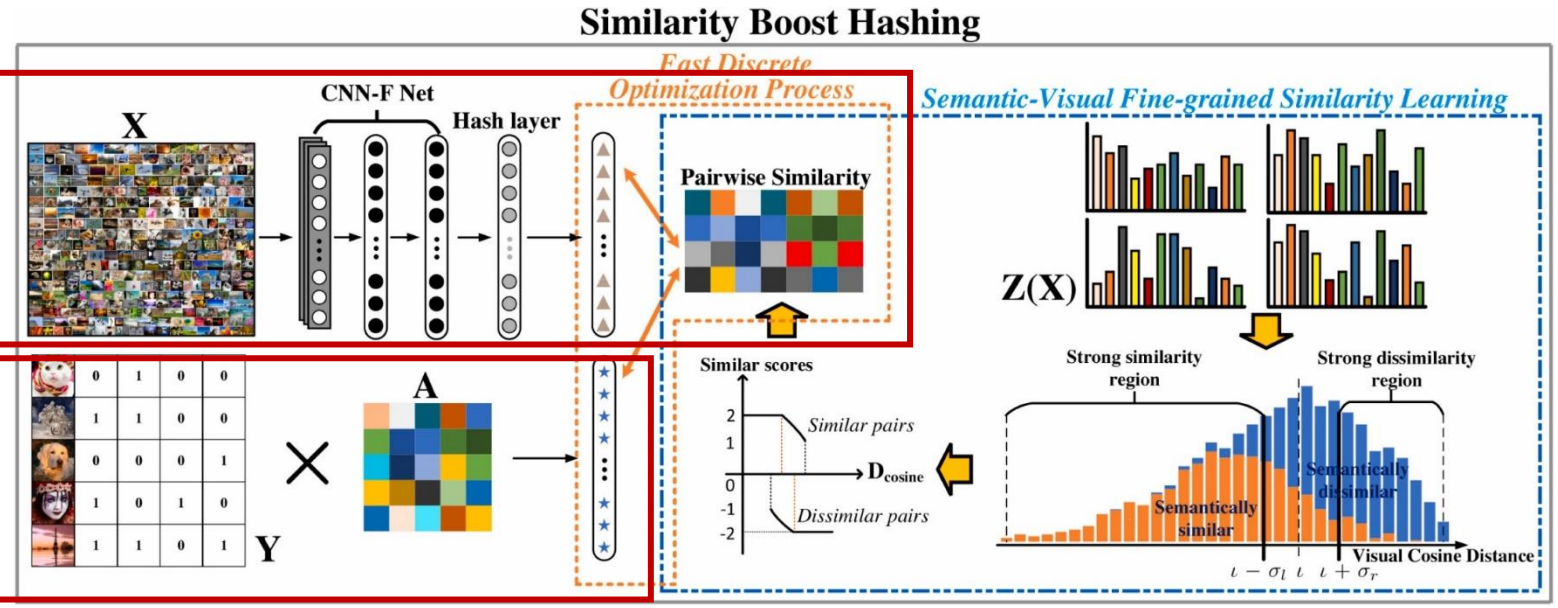
$$s.t.\ \boldsymbol{B} \in \{-1,1\}^{k\times n}$$

其中λ、γ是平衡参数，ρ是正则化参数。

$$\min_{\theta,\boldsymbol{B}} \|\mathcal{F}(\boldsymbol{X};\theta)^T\boldsymbol{B} - k\boldsymbol{S}\|_F^2 + \lambda\|\mathcal{F}(\boldsymbol{X};\theta) - \boldsymbol{B}\|_F^2$$

$$s.t.\ \boldsymbol{B} \in \{-1,1\}^{k\times n}$$

$$\min_{\boldsymbol{A}} \|\boldsymbol{B} - \boldsymbol{A}^T\boldsymbol{Y}\|_F^2$$



二 研究方法

语义-视觉连续相似性学习

如前所述，图像对的粗略二元相似信息无法反映不同图像之间的视觉关联，导致哈希方法存在视觉相似性缺失的问题。深度神经网络的中间层特征包含丰富的视觉信息，因此，本研究从预训练的CNN-F网络中提取4096维特征来表示图像视觉内容，CNN-F特征提取器用$\mathcal{Z}(\cdot)$表示。

为了考虑语义-视觉连续相似性的度量，可以将图像相似性分为四种情况：

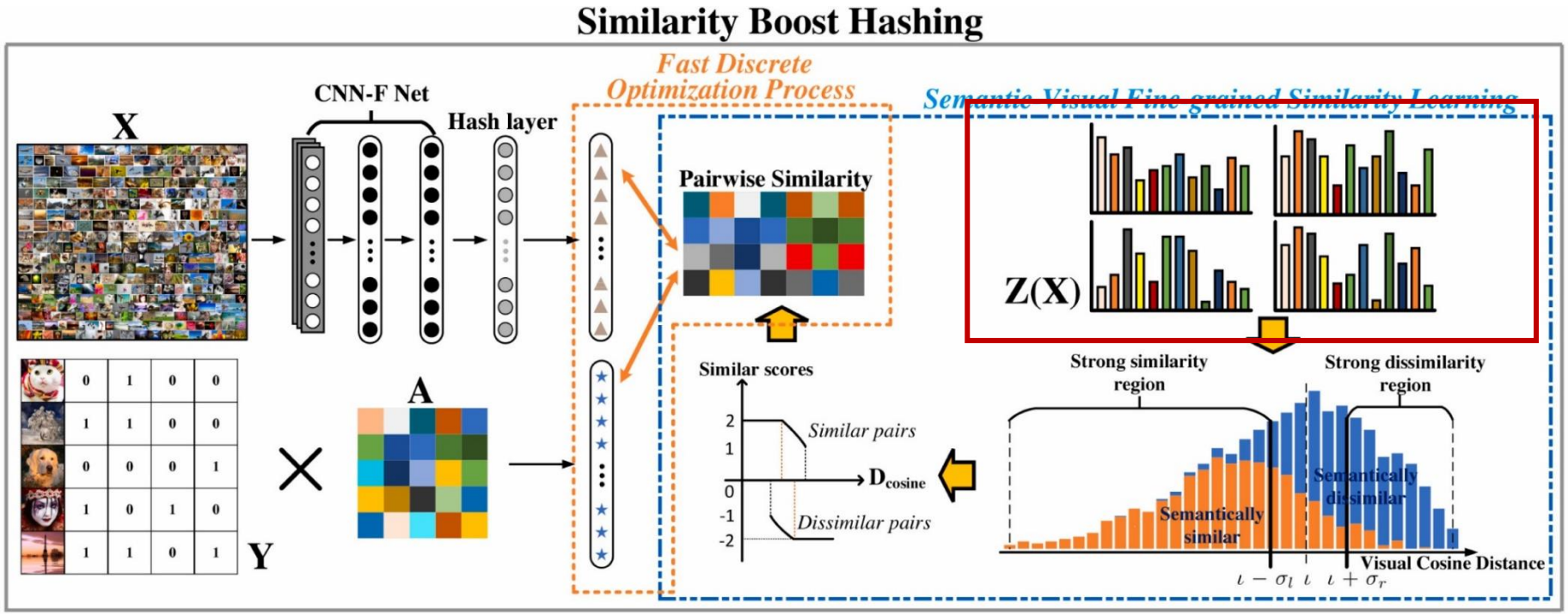
- **完全相似**：语义相似且视觉相似
- **部分相似**：语义相似且视觉不相似
- **部分不相似**：语义不相似且视觉相似
- **完全不相似**：语义不相似且视觉不相似

这样在之前常见的“完全相似”和“完全不相似”基础上，通过考虑视觉相似性进一步细化了图像之间的相似度关系。为此，需要一个新的相似度矩阵构建方式。

二 研究方法

语义-视觉连续相似性学习

首先，从ImageNet、 NUS-WIDE、 MS-COCO数据集随机抽取10000个图像对，其中包括5000个语义相似的图像对和5000个语义不同的图像对。然后，利用CNN-F提取每个图像的4096维特征，计算特征之间的余弦距离作为图像对的视觉距离。

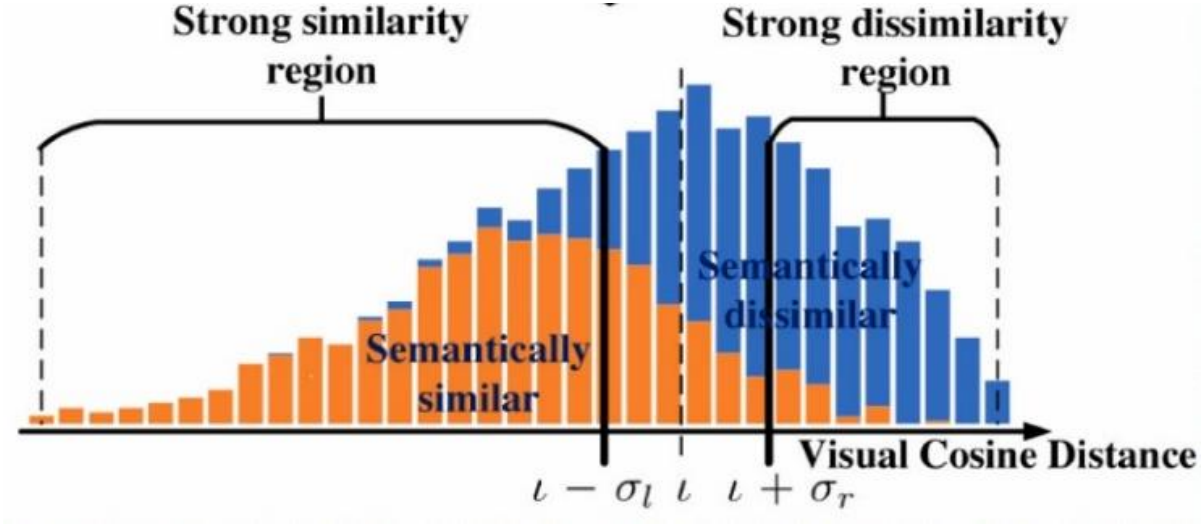


二 研究方法

语义-视觉连续相似性学习

10000个图像对的余弦距离分布如下图。接下来，用一个非对称广义正态分布来拟合该距离分布，该分布定义为：

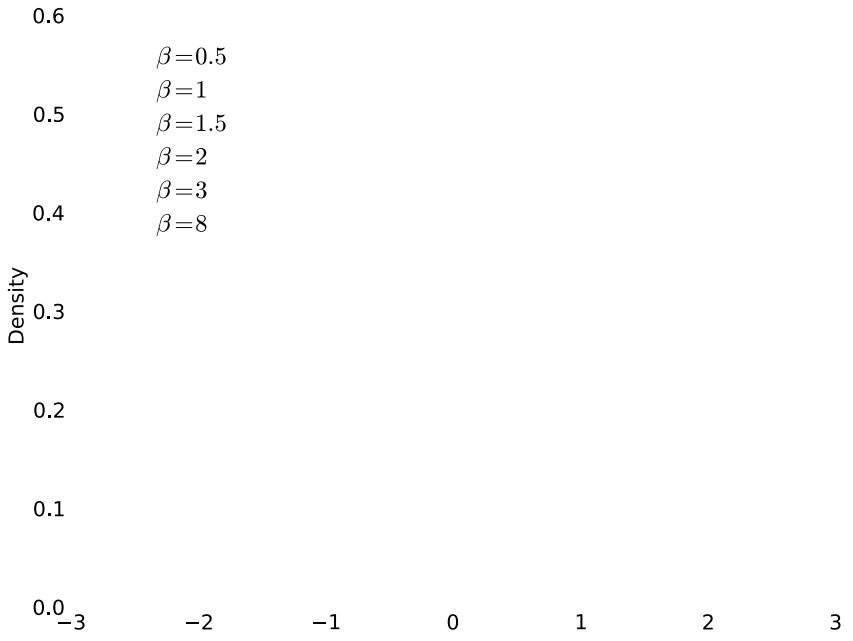
$$f(x;\zeta,\tau_l,\tau_r)=\begin{cases}\frac{\zeta}{(\tau_l+\tau_r)\Gamma(1/\zeta)}\exp-(\frac{-x+\iota}{\tau_l})^{\zeta} & x-\iota<0\\ \frac{\zeta}{(\tau_l+\tau_r)\Gamma(1/\zeta)}\exp-(\frac{x-\iota}{\tau_r})^{\zeta} & x-\iota\geq 0,\end{cases}$$



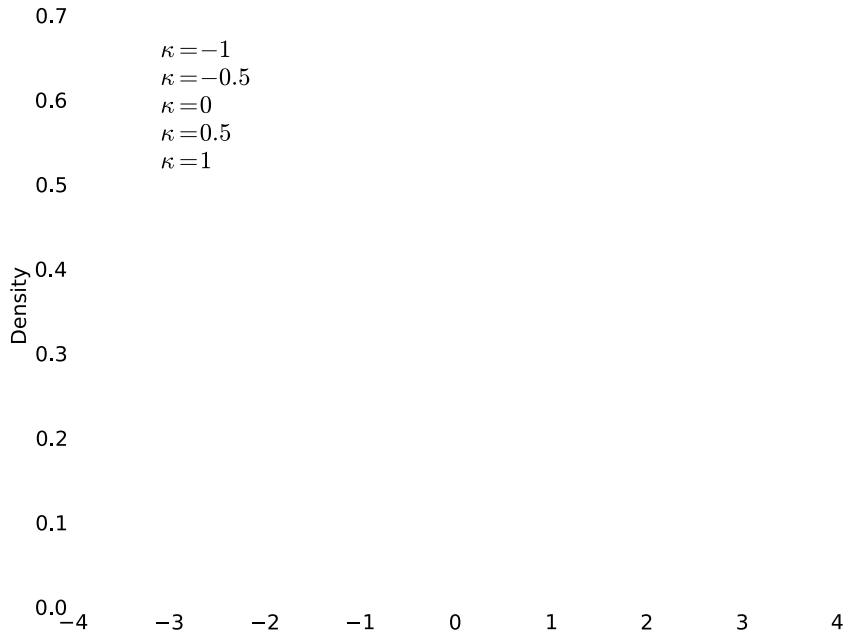
二 研究方法

语义-视觉连续相似性学习

- 对称广义正态分布：向正态分布添加形状参数（图中为β），当$\beta = 2$时，特化为正态分布；当$\beta = 1$时，特化为拉普拉斯分布；当$\beta \to \infty$时，收敛到均匀分布。
- 非对称广义正态分布：向对称广义正态分布添加形状参数，以产生一定的偏度。



对称广义正态分布



非对称广义正态分布

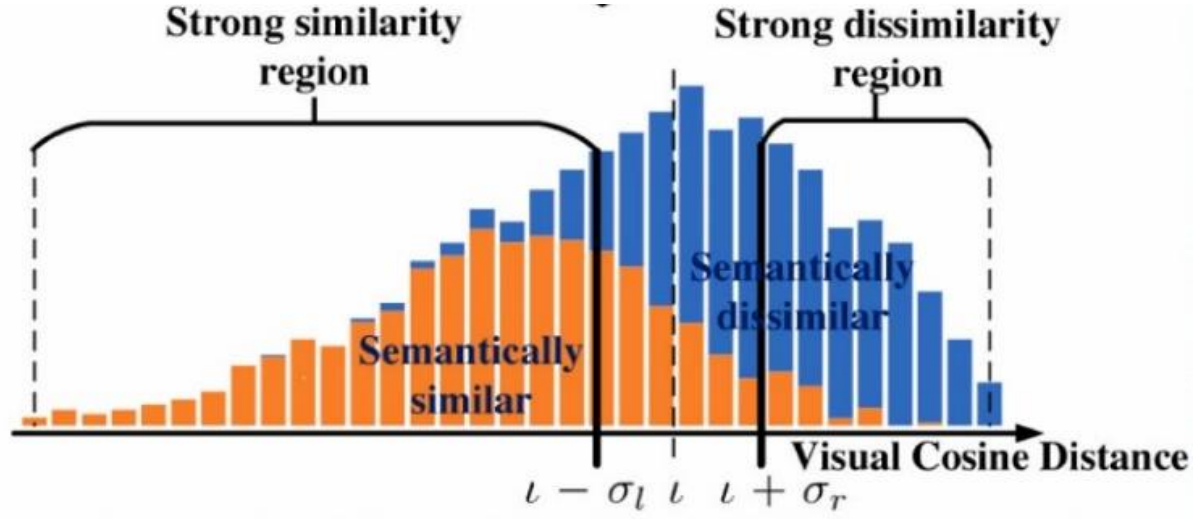
二 研究方法

语义-视觉连续相似性学习

为方便起见，本研究取$\zeta = 2$，分布可视为非对称正态分布。分别设置$\iota - \sigma_l$、$\iota + \sigma_r$两个阈值，根据正态分布的性质，阈值之外的区域占比约为31.73%。设置阈值之外区域分别为强相似区域和强不相似区域。

直观来说，可以采用以下方式构建语义-视觉连续相似性矩阵$\hat{\boldsymbol{S}}$，当语义相似图像对的余弦距离落在强相似区域时，令$\hat{s}_{ij} = 2$；当语义不相似图像对的余弦距离落在强不相似区域时，令$\hat{s}_{ij} = -2$。当余弦距离落在中间区域时，令$1 \le |\hat{s}_{ij}| \le 2$。

$$f(x;\zeta,\tau_l,\tau_r) = \begin{cases} \frac{\zeta}{(\tau_l+\tau_r)\Gamma(1/\zeta)} \exp -(\frac{-x+\iota}{\tau_l})^{\zeta} & x-\iota < 0 \\ \frac{\zeta}{(\tau_l+\tau_r)\Gamma(1/\zeta)} \exp -(\frac{x-\iota}{\tau_r})^{\zeta} & x-\iota \ge 0, \end{cases}$$



二 研究方法

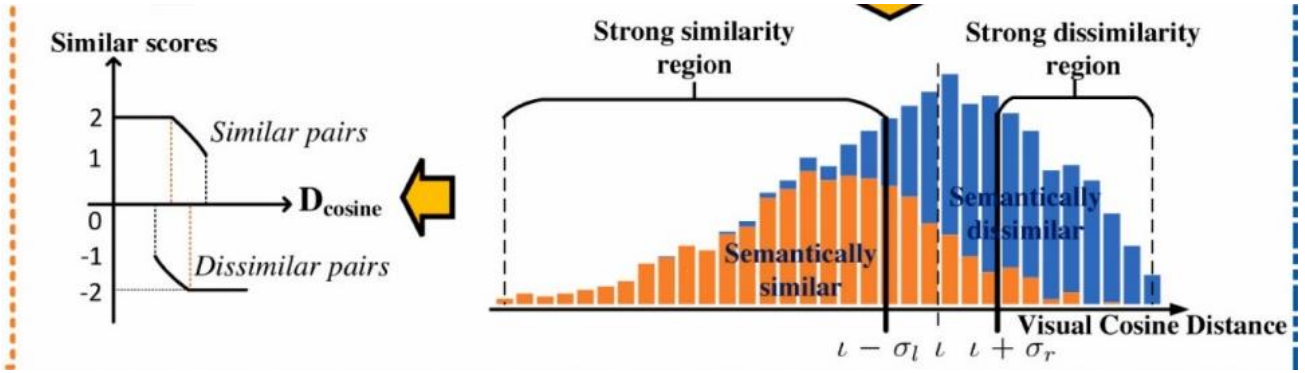
语义-视觉连续相似性学习

语义-视觉连续相似性矩阵$\hat{S}$的具体公式为：

$$\hat{s}_{ij} = \begin{cases} \ell\left(\log\left(\frac{\alpha+\frac{1}{n}}{D(\mathcal{Z}(x_i),\mathcal{Z}(x_j))+\frac{1}{n}}\right)\right)+1 & s_{ij}=1 \\ -\ell\left(\log\left(\frac{D(\mathcal{Z}(x_i),\mathcal{Z}(x_j))+\frac{1}{n}}{\beta+\frac{1}{n}}\right)\right)-1 & s_{ij}=-1, \end{cases}$$

其中，$D(x,y)=1-\frac{<x,y>}{\|x\|\|y\|}$表示余弦相似度，$\alpha=(\iota-\sigma_l)\cdot e$控制强相似区域的阈值，$\beta=(\iota+\sigma_r)/e$控制强不相似区域的阈值，$1/n$为平滑常数项。最后的常数项$\pm1$为语义相似性，$\ell(\cdot)$为阈值函数，定义为：

$$\ell(x)=\begin{cases}1 & x>1\\ x & 0\le x\le 1\\ 0 & \text{otherwise.}\end{cases}$$



二 研究方法

EDDH算法

Algorithm 1 Enhanced Deep Discrete Hashing (EDDH)

Input: Training images $\mathbf{X}$, label matrix $\mathbf{Y}$, hash code length k, parameter $\alpha, \beta, \lambda, \mu, \gamma, \rho$, maximum iteration number T.
Output: parameters θ, discriminative hash codes $\mathbf{B}$.
Procedure:
1. **Initialize**: $\mathbf{A}, \mathbf{B}, \mathbf{P}$ and θ.
2. Initialize the weights and bias of whole network.
3. $\hat{\mathbf{S}} \leftarrow$ using (10).
4. **Repeat**
   P-Step: Use Eq. (17) to solve $\mathbf{P}$.
   θ-Step: Use SGD algorithm to update θ.
   A-Step: Use Eq. (22) to solve $\mathbf{A}$.
   B-Step: Use Eq. (25) to solve $\mathbf{B}$.
   Until up to T.
5. Learn the hash codes by $\mathbf{B} = sgn(\mathcal{F}(\mathbf{X}; \theta))$.

逐步求解

二 研究方法

实验结果

Table 1
mAP results of hamming ranking for different number of bits on the three image datasets.

Method	ImageNet				NUS-WIDE				MS-COCO			
	16 bits	32 bits	48 bits	64 bits	16 bits	32 bits	48 bits	64 bits	16 bits	32 bits	48 bits	64 bits
SH	0.093	0.131	0.157	0.159	0.172	0.192	0.223	0.268	0.434	0.487	0.482	0.511
ITQ	0.259	0.333	0.356	0.411	0.352	0.427	0.459	0.498	0.521	0.584	0.621	0.632
SDH	0.092	0.138	0.162	0.188	0.302	0.407	0.428	0.464	0.484	0.523	0.546	0.559
KSH	0.083	0.108	0.131	0.152	0.210	0.277	0.311	0.337	0.489	0.521	0.524	0.536
CNNH	0.101	0.204	0.283	0.336	0.366	0.443	0.457	0.484	0.562	0.560	0.552	0.564
DHN	0.273	0.393	0.435	0.491	0.584	0.613	0.636	0.638	0.622	0.648	0.665	0.669
DDSH	0.281	0.399	0.451	0.513	0.589	0.622	0.657	0.662	0.637	0.656	0.671	0.689
DPH	0.347	0.512	0.556	0.579	0.654	0.683	0.699	0.728	0.687	0.722	0.739	0.763
SDAH	0.356	0.523	0.570	0.583	0.661	0.692	0.704	0.734	0.703	0.729	0.761	0.786
EDDH	**0.381**	**0.562**	**0.597**	**0.604**	**0.711**	**0.730**	**0.745**	**0.771**	**0.747**	**0.771**	**0.794**	**0.811**

Q&A

合肥工業大學