河海大学
HOHAI UNIVERSITY

面向图像与文本的跨模态检索

汇报人：jiangli

2020年2月17日

汇报内容

河海大学

研究背景

- 数据量逐年增长

- 数据模态形式丰富
  - 文本、图像、视频、音频……

问题定义

- **跨模态检索：**
  - 给定一种模态下的查询，从大规模数据库中快速找到相关的其它模态下的数据。

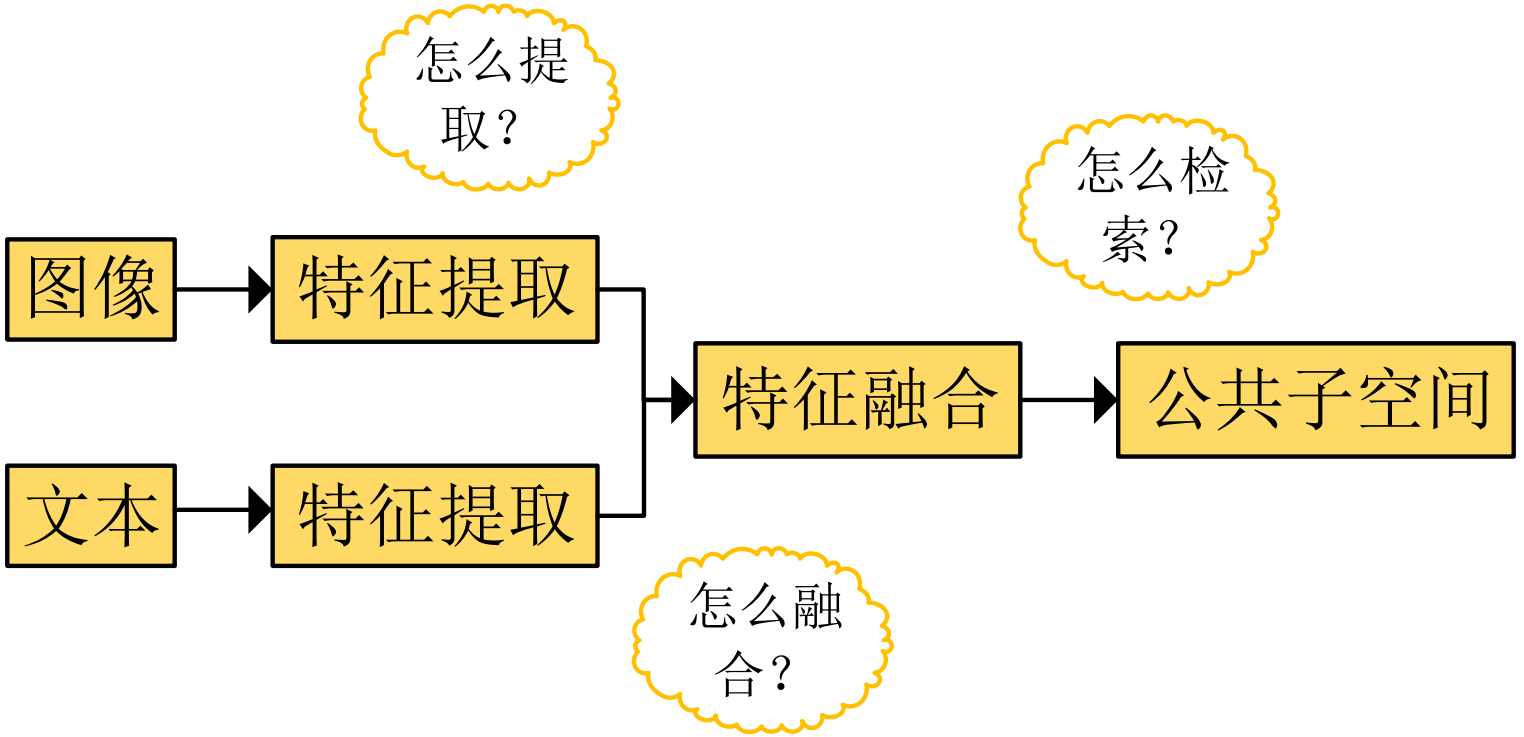
- **图像和文本之间的跨模态检索：**
  - 自然图像作为查询，检索有着相同或相似语义的文本句子（基于图像的文本检索），或者反之（基于文本的图像检索）。

检索

1. A black and white dog is running in a grassy garden surrounded by a white fence .
2. A boston terrier is running on lush green grass in front of a white fence .
3. A dog runs on the grass near a wooden fence .

问题定义

- **跨模态检索过程：**
  - 特征提取
  - 多模态特征融合
  - 向量索引方法



相关论文

论文	论文内容
Deep cross-modal hashing[C]//CVPR. 2017	提出跨模态深度哈希(DCMH)方法，将深度学习和哈希学习集成到同一框架中。
Dual attention networks for multimodal reasoning and matching[C]//CVPR. 2017.	提出双重注意力网络(DANs)，同时学习视觉和文本注意力模型，以挖掘视觉与文本之间的细粒度交互作用。
Spreading vectors for similarity search[J]. ICLR, 2019.	设计了一个深度神经网络，对数据分布重新进行调整，将其投影到均匀分布的空间，然后采用量化器进行量化。

文献一

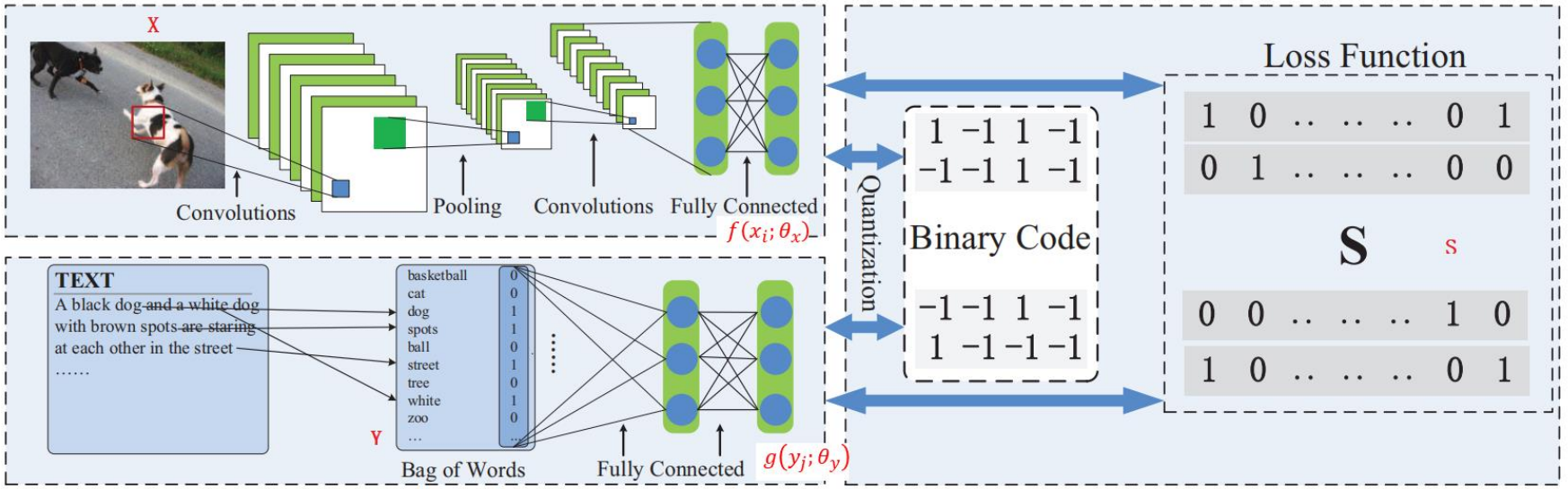


Figure 1. The end-to-end deep learning framework of our DCMH model.

$X = \{x_i\}_{i=1}^{n}$: n points of image modality.

$Y = \{y_j\}_{j=1}^{n}$: n points of text modality.

$S = \{S_{ij}\}_{n\times n}$: cross-modal similarity.

$f(x_i;\theta_x)$: the output of deep neural network for image modality.

$g(y_j;\theta_y)$: the output of deep neural network for text modality.

[1] Jiang Q Y, Li W J. Deep cross-modal hashing[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3232-3240.

文献一

1 Feature learning

- Image deep neural network input:
  raw image
- Text deep neural network input:
  BOW feature

Table 1. Configuration of the CNN for image modality.

Layer	Configuration
conv1	f. $64 \times 11 \times 11$; st. 4×4, pad 0, LRN, $\times 2$ pool
conv2	f. $265 \times 5 \times 5$; st. 1×1, pad 2, LRN, $\times 2$ pool
conv3	f. $265 \times 3 \times 3$; st. 1×1, pad 1
conv4	f. $265 \times 3 \times 3$; st. 1×1, pad 1
conv5	f. $265 \times 3 \times 3$; st. 1×1, pad 1, $\times 2$ pool
full6	4096
full7	4096
full8	Hash code length c

Table 2. Configuration of the deep neural network for text modality.

Layer	Configuration
full1	8192
full2	Hash code length c

河海大學

BoW模型

- 首先给出两个简单的文本文档如下：
  - John likes to watch movies. Mary likes too.
  - John also likes to watch football games.

- 基于上述两个文档中出现的单词，构建如下一个词典：
  - dictionary: {"John": 1, "likes": 2,"to": 3, "watch": 4, "movies": 5,"also": 6, "football": 7, "games": 8,"Mary": 9, "too": 10}

- 每个文本我们可以使用一个10维的向量来表示，该向量与原来文本中单词出现的顺序没有关系，而是词典中每个单词在文本中出现的频率：
  - [1, 2, 1, 1, 1, 0, 0, 0, 1, 1]
  - [1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

文献一

The objective function of DCMH is defined as follows:

$$
\begin{aligned}
\min_{\mathbf{B},\mathbf{B}^{(x)},\mathbf{B}^{(y)},\theta_x,\theta_y} \quad & \mathcal{J} = -\sum_{i,j=1}^{n}(S_{ij}\Theta_{ij} - \log(1+e^{\Theta_{ij}})) \text{①} \\
& + \gamma(||\mathbf{B}^{(x)} - \mathbf{F}||_F^2 + ||\mathbf{B}^{(y)} - \mathbf{G}||_F^2) \text{②} \\
& + \eta(||\mathbf{F1}||_F^2 + ||\mathbf{G1}||_F^2) \text{③} \qquad (1) \\
s.t. \quad & \mathbf{B}^{(x)} \in \{-1,+1\}^{c\times n}, \\
& \mathbf{B}^{(y)} \in \{-1,+1\}^{c\times n}, \\
& \mathbf{B} \in \{-1,+1\}^{c\times n}, \\
& \mathbf{B} = \mathbf{B}^{(x)} = \mathbf{B}^{(y)},
\end{aligned}
$$

图像特征

where $\mathbf{F} \in \mathbb{R}^{c\times n}$ with $\mathbf{F}_{*i} = f(\mathbf{x}_i;\theta_x)$, $\mathbf{G} \in \mathbb{R}^{c\times n}$ with $\mathbf{G}_{*j} = g(\mathbf{y}_j;\theta_y)$, $\Theta_{ij} = \frac{1}{2}\mathbf{F}_{*i}^T\mathbf{G}_{*j}$, $\mathbf{B}_{*i}^{(x)}$ is the binary hash code for image $\mathbf{x}_i$, $\mathbf{B}_{*j}^{(y)}$ is the binary hash code for text $\mathbf{y}_j$, γ and η are hyper-parameters.

文本特征　图文相似性　image的二值码　text的二值码

2 Hash-code learning

- 目标函数：
  - 最大化似然函数
  - 二值码和输出特征的距离尽可能小
  - 哈希码的两种状态码尽可能平均

The first term $-\sum_{i,j=1}^{n}(S_{ij}\Theta_{ij} - \log(1+e^{\Theta_{ij}}))$ in (1) is the negative log likelihood of the cross-modal similarities with the likelihood function defined as follows:

得到

似然函数 $p(S_{ij}|\mathbf{F}_{*i},\mathbf{G}_{*j}) = \begin{cases} \sigma(\Theta_{ij}) & S_{ij}=1 \\ 1-\sigma(\Theta_{ij}) & S_{ij}=0 \end{cases}$

代入

where $\Theta_{ij} = \frac{1}{2}\mathbf{F}_{*i}^T\mathbf{G}_{*j}$ and $\sigma(\Theta_{ij}) = \frac{1}{1+e^{-\Theta_{ij}}}$.

F和G之间的相似性

河海大学

文献一

Algorithm 1 The learning algorithm for DCMH.

Input: Image set $\mathbf{X}$, text set $\mathbf{Y}$, and cross-modal similarity matrix $\mathbf{S}$.

Output: Parameters θ_x and θ_y of the deep neural networks, and binary code matrix $\mathbf{B}$.

Initialization

Initialize neural network parameters θ_x and θ_y, mini-batch size $N_x = N_y = 128$, and iteration number $t_x = \lceil n/N_x \rceil, t_y = \lceil n/N_y \rceil$.

repeat

  for $iter = 1, 2, \cdots, t_x$ **do**

    Randomly sample N_x points from $\mathbf{X}$ to construct a mini-batch.

    For each sampled point $\mathbf{x}_i$ in the mini-batch, calculate $\mathbf{F}_{*i} = f(\mathbf{x}_i; \theta_x)$ by forward propagation.

    Calculate the derivative according to (3).

    Update the parameter θ_x by using back propagation.

  end for

  for $iter = 1, 2, \cdots, t_y$ **do**

    Randomly sample N_y points from $\mathbf{Y}$ to construct a mini-batch.

    For each sampled point $\mathbf{y}_j$ in the mini-batch, calculate $\mathbf{G}_{*j} = g(\mathbf{y}_j; \theta_y)$ by forward propagation.

    Calculate the derivative according to (4).

    Update the parameter θ_y by using back propagation.

  end for

  Learn $\mathbf{B}$ according to (5).

until a fixed number of iterations

3 Learning

- We adopt an alternating learning strategy to learn, and B. Each time we learn one parameter with the other parameters fixed.

河海大學

文献一

4 Experiment

- Datasets
  - MIRFLICKR-25K、IARP-TC12、NUS-WIDE
- Baselines
  - SePH、STMH、SCM、CMFH、CCA
- Task
  - Hamming ranking & Hash lookup

文献一

Table 3. MAP. The best accuracy is shown in boldface. The baselines are based on hand-crafted features.

Task	Method	MIRFLICKR-25K			IAPR TC-12			NUS-WIDE		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
$I \rightarrow T$	DCMH	**0.7410**	**0.7465**	**0.7485**	**0.4526**	**0.4732**	**0.4844**	**0.5903**	**0.6031**	**0.6093**
	SePH	0.6573	0.6603	0.6616	0.4112	0.4158	0.4203	0.4787	0.4869	0.4888
	STMH	0.5921	0.5950	0.5980	0.3580	0.3732	0.3819	0.3973	0.4082	0.4153
	SCM	0.6290	0.6404	0.6480	0.3833	0.3898	0.3878	0.4650	0.4714	0.4822
	CMFH	0.5818	0.5808	0.5805	0.3683	0.3734	0.3786	0.3568	0.3624	0.3661
	CCA	0.5695	0.5663	0.5641	0.3345	0.3254	0.3193	0.3414	0.3336	0.3282
$T \rightarrow I$	DCMH	**0.7827**	**0.7900**	**0.7932**	**0.5185**	**0.5378**	**0.5468**	**0.6389**	**0.6511**	**0.6571**
	SePH	0.6480	0.6521	0.6545	0.4024	0.4074	0.4131	0.4489	0.4539	0.4587
	STMH	0.5802	0.5846	0.5855	0.3445	0.3570	0.3690	0.3607	0.3738	0.3842
	SCM	0.6195	0.6302	0.6366	0.3698	0.3734	0.3696	0.4370	0.4428	0.4504
	CMFH	0.5787	0.5774	0.5784	0.3619	0.3687	0.3769	0.3623	0.3670	0.3723
	CCA	0.5690	0.5659	0.5639	0.3340	0.3255	0.3197	0.3392	0.3320	0.3272

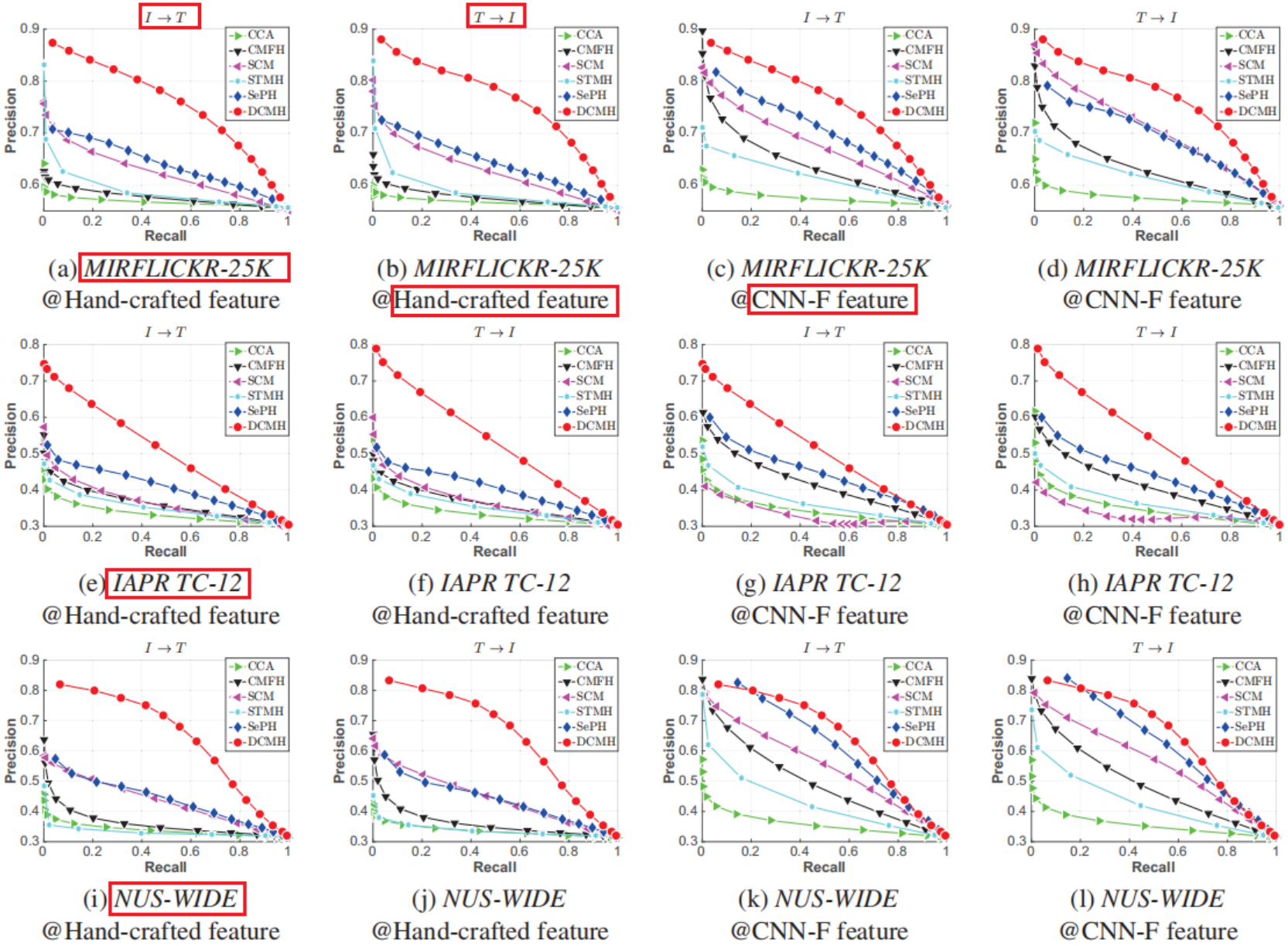
Table 4. MAP. The best accuracy is shown in boldface. The baselines are based on CNN-F features.

Task	Method	MIRFLICKR-25K			IAPR TC-12			NUS-WIDE		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
$I \rightarrow T$	DCMH	**0.7410**	**0.7465**	**0.7485**	**0.4526**	**0.4732**	**0.4844**	0.5903	0.6031	0.6093
	SePH	0.7123	0.7194	0.7232	0.4442	0.4563	0.4639	**0.6037**	**0.6136**	**0.6211**
	STMH	0.6132	0.6219	0.6274	0.3775	0.4002	0.4130	0.4710	0.4864	0.4942
	SCM	0.6851	0.6921	0.7003	0.3692	0.3666	0.3802	0.5409	0.5485	0.5553
	CMFH	0.6377	0.6418	0.6451	0.4189	0.4234	0.4251	0.4900	0.5053	0.5097
	CCA	0.5719	0.5693	0.5672	0.3422	0.3361	0.3300	0.3604	0.3485	0.3390
$T \rightarrow I$	DCMH	**0.7827**	**0.7900**	**0.7932**	**0.5185**	**0.5378**	**0.5468**	**0.6389**	**0.6511**	**0.6571**
	SePH	0.7216	0.7261	0.7319	0.4423	0.4562	0.4648	0.5983	0.6025	0.6109
	STMH	0.6074	0.6153	0.6217	0.3687	0.3897	0.4044	0.4471	0.4677	0.4780
	SCM	0.6939	0.7012	0.7060	0.3453	0.3410	0.3470	0.5344	0.5412	0.5484
	CMFH	0.6365	0.6399	0.6429	0.4168	0.4212	0.4277	0.5031	0.5187	0.5225
	CCA	0.5742	0.5713	0.5691	0.3493	0.3438	0.3378	0.3614	0.3494	0.3395

Hammimg Ranking Task(MAP):

- "$I \rightarrow T$ "denotes the case where the query is image and the database is text.
- "$T \rightarrow I$ " denotes the case where the query is text and the database is image.

文献一



(a) *MIRFLICKR-25K* @Hand-crafted feature
(b) *MIRFLICKR-25K* @Hand-crafted feature
(c) *MIRFLICKR-25K* @CNN-F feature
(d) *MIRFLICKR-25K* @CNN-F feature
(e) *IAPR TC-12* @Hand-crafted feature
(f) *IAPR TC-12* @Hand-crafted feature
(g) *IAPR TC-12* @CNN-F feature
(h) *IAPR TC-12* @CNN-F feature
(i) *NUS-WIDE* @Hand-crafted feature
(j) *NUS-WIDE* @Hand-crafted feature
(k) *NUS-WIDE* @CNN-F feature
(l) *NUS-WIDE* @CNN-F feature

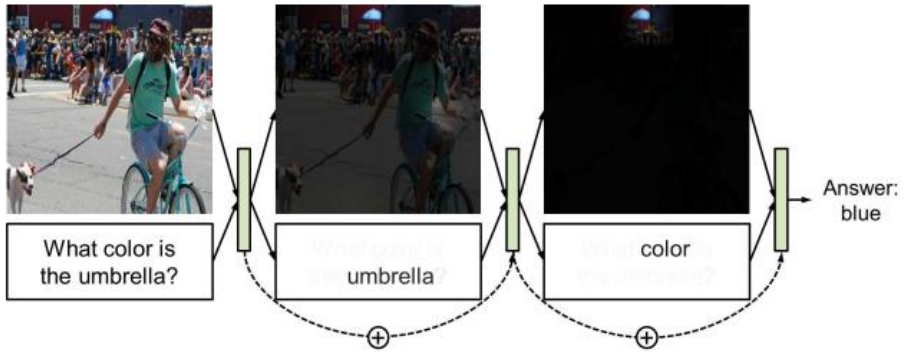
Figure 2. Precision-recall curves on three datasets. The code length is 16.

Hash Lookup Task
(Precision Recall Curve)

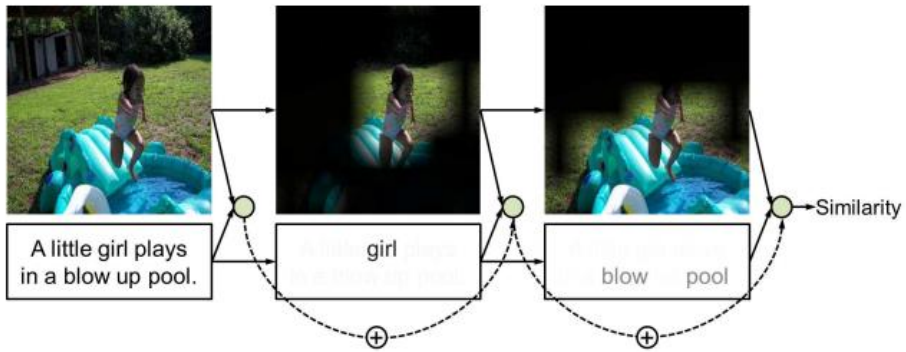
河海大学

文献二

- Contributions[2]:
  - We propose an integrated framework of visual and textual attentions, where critical regions and words are jointly located through multiple steps.
  - Two variants of the proposed framework are implemented for multimodal reasoning and matching, and applied to VQA and image-text matching.
  - Detailed visualization of the attention results validates that our models effectively focus on vital portions of visual and textual data for the given task.
  - Our framework demonstrates the state-of-the-art performance on the VQA dataset and the Flickr30K image-text matching dataset.



(a) DAN for multimodal reasoning. (r-DAN)



(b) DAN for multimodal matching. (m-DAN)

Figure 1: Overview of Dual Attention Networks (DANs) for multimodal reasoning and matching. The brightness of image regions and darkness of words indicate their attention weights predicted by DANs.

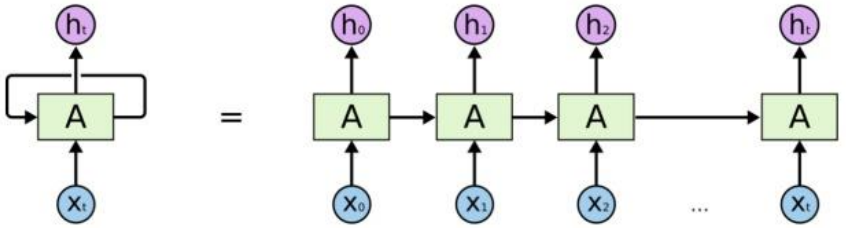
[2] Nam H, Ha J W, Kim J. Dual attention networks for multimodal reasoning and matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 299-307.

文献二

1 Input representation

- Image representation:
  - model: VGGNet or ResNet.
  - Representation: $\{v_1, \dots, v_n\}$
    N: number of image regions
    v_n: 512 (VGGNet) or 2048 (ResNet) dimensional feature vector
- Text representation:
  - model: Bidirectional LSTMs.
  - Representation: $\{u_1, \dots, u_T\}$

[3]



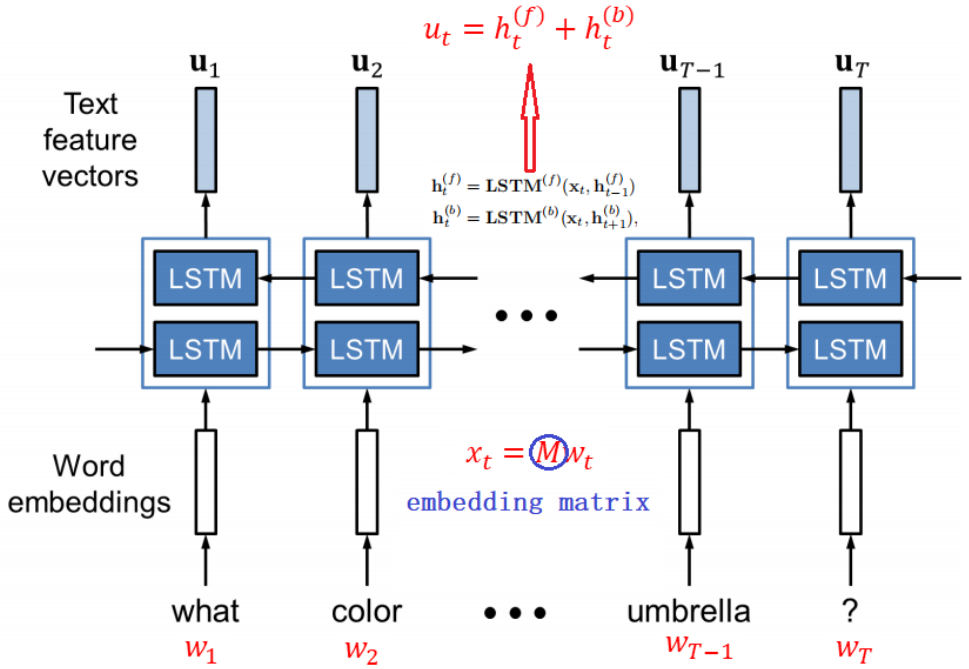


Figure 2: Bidirectional LSTMs for text encoding.

[3] Schuster M, Paliwal K K. Bidirectional recurrent neural networks[J]. IEEE transactions on Signal Processing, 1997, 45(11): 2673-2681.

文献二

2 Attention mechanisms

- Image attention:

$$v^k = V\_Att(\{v_n\}_{n=1}^N, m_v^{(k-1)})$$

$$\begin{aligned}
\mathbf{h}_{\mathbf{v},n}^{(k)} &= \tanh\left(\mathbf{W}_{\mathbf{v}}^{(k)}\ \mathbf{v}_n\right) \odot \tanh\left(\mathbf{W}_{\mathbf{v},\mathbf{m}}^{(k)}\ \mathbf{m}_{\mathbf{v}}^{(k-1)}\right), \\
\alpha_{\mathbf{v},n}^{(k)} &= \mathrm{softmax}\left(\mathbf{W}_{\mathbf{v},\mathbf{h}}^{(k)}\ \mathbf{h}_{\mathbf{v},n}^{(k)}\right), \\
\mathbf{v}^{(k)} &= \tanh\left(\mathbf{P}^{(k)} \sum_{n=1}^{N} \alpha_{\mathbf{v},n}^{(k)}\ \mathbf{v}_n\right),
\end{aligned}$$

- Textual attention:

$$u^k = V\_Att(\{u_t\}_{t=1}^T, m_u^{(k-1)})$$

$$\begin{aligned}
\mathbf{h}_{\mathbf{u},t}^{(k)} &= \tanh\left(\mathbf{W}_{\mathbf{u}}^{(k)}\ \mathbf{u}_t\right) \odot \tanh\left(\mathbf{W}_{\mathbf{u},\mathbf{m}}^{(k)}\ \mathbf{m}_{\mathbf{u}}^{(k-1)}\right), \\
\alpha_{\mathbf{u},t}^{(k)} &= \mathrm{softmax}\left(\mathbf{W}_{\mathbf{u},\mathbf{h}}^{(k)}\ \mathbf{h}_{\mathbf{u},t}^{(k)}\right), \\
\mathbf{u}^{(k)} &= \sum_{t} \alpha_{\mathbf{u},t}^{(k)}\ \mathbf{u}_t.
\end{aligned}$$

$m_v^{(k-1)}$ is a memory vector encoding the information that has been attended until step $k-1$.

文献二

3 r-DAN for visual question answering

$$\mathbf{m}^{(k)} = \mathbf{m}^{(k-1)} + \mathbf{v}^{(k)} \odot \mathbf{u}^{(k)},$$

$$\mathbf{m}^{(0)} = \mathbf{v}^{(0)} \odot \mathbf{u}^{(0)},$$

where $\mathbf{v}^{(0)} = \tanh\left(\mathbf{P}^{(0)} \frac{1}{N} \sum_n \mathbf{v}_n\right)$

$$\mathbf{u}^{(0)} = \frac{1}{T} \sum_t \mathbf{u}_t.$$

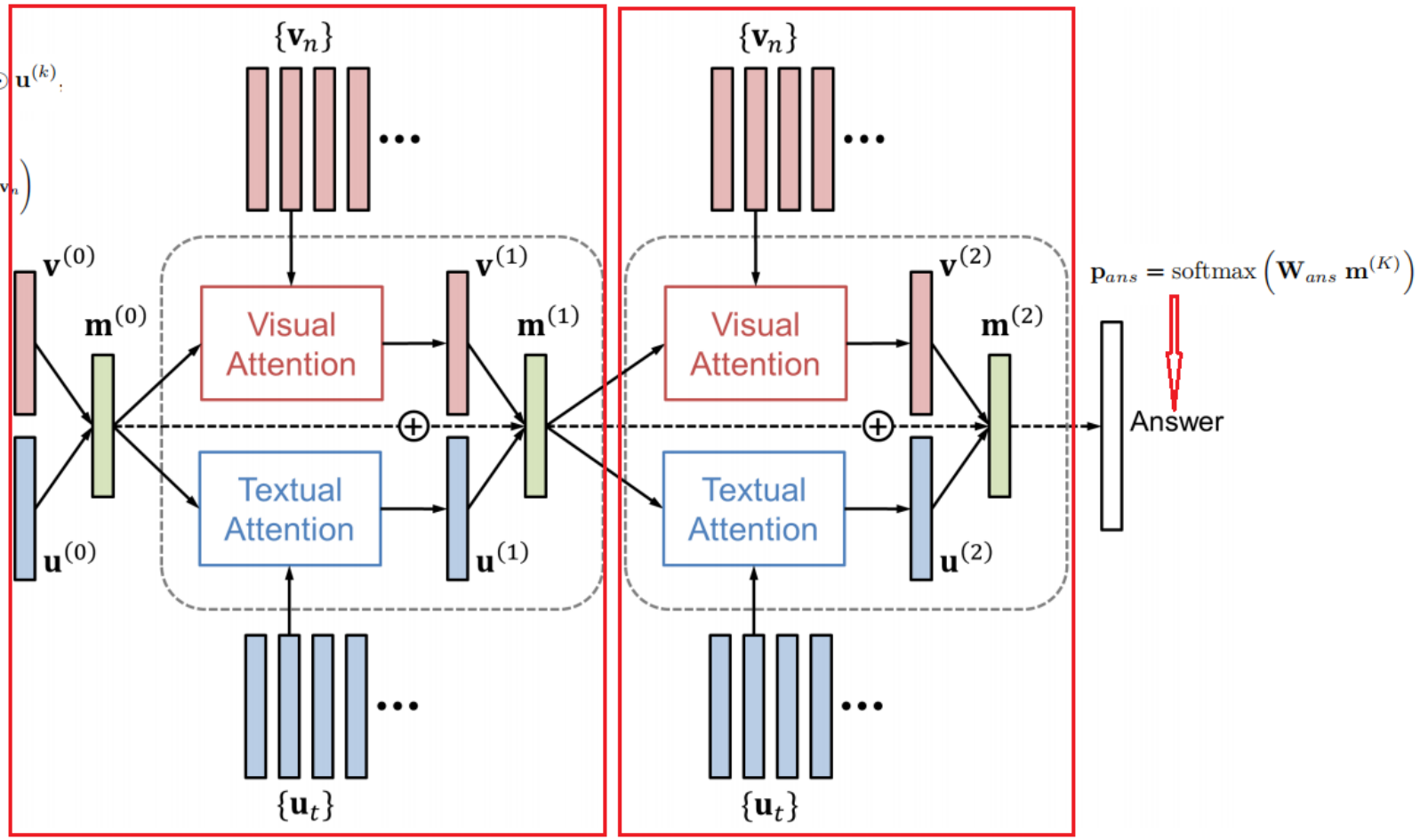


Figure 3: r-DAN in case of $K = 2$.

河海大學

文献二

4 m-DAN for image-text matching

$$m_v^{(k)} = m_v^{(k-1)} + v^{(k)}$$
$$m_u^{(k)} = m_u^{(k-1)} + u^{(k)}$$

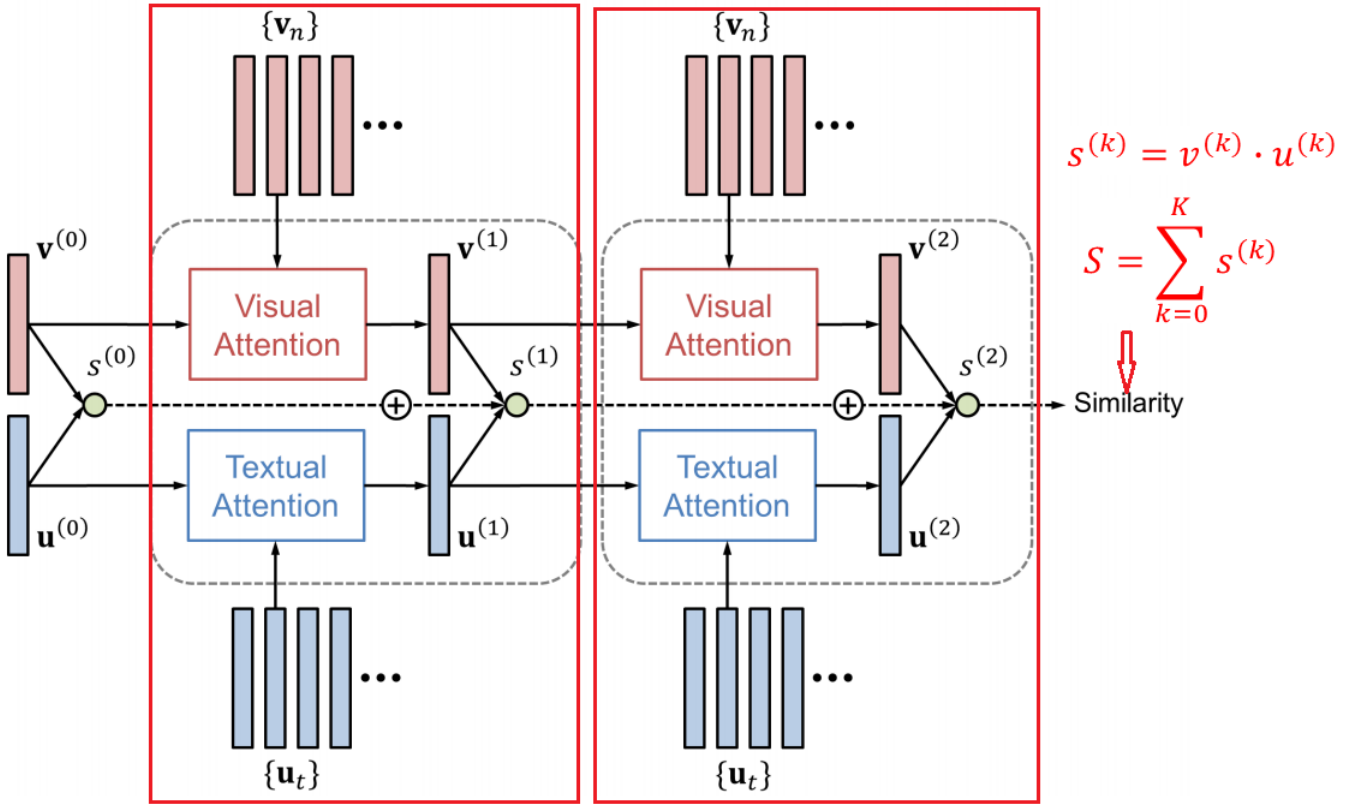


Figure 4: m-DAN in case of $K = 2$.

河海大学

文献二

5 Experiment

- Datasets
  - Flickr30K dataset
- Task
  - Image-text matching
- Input
  - <positive image, positive sentence, negative image, negative sentence>

文献二

Table 2: Bidirectional retrieval results on the Flickr30K dataset compared with state-of-the-art methods.

	Image-to-Text				Text-to-Image			
Method	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR
DCCA [34]	27.9	56.9	68.2	4	26.8	52.9	66.9	4
mCNN [19]	33.6	64.1	74.9	3	26.2	56.3	69.6	4
m-RNN-VGG [20]	35.4	63.8	73.7	3	22.8	50.7	63.1	5
GMM+HGLMM FV [14]	35.0	62.0	73.8	3	25.0	52.7	66.0	5
HGLMM FV [24]	36.5	62.2	73.3	-	24.7	53.4	66.8	-
SPE [30]	40.3	68.9	79.9	-	29.7	60.1	72.1	-
DAN (VGG)	41.4	73.5	82.5	2	31.8	61.7	72.5	3
DAN (ResNet)	**55.0**	**81.8**	**89.0**	**1**	**39.4**	**69.2**	**79.1**	**2**

文献二

- Text-to-image retrieval

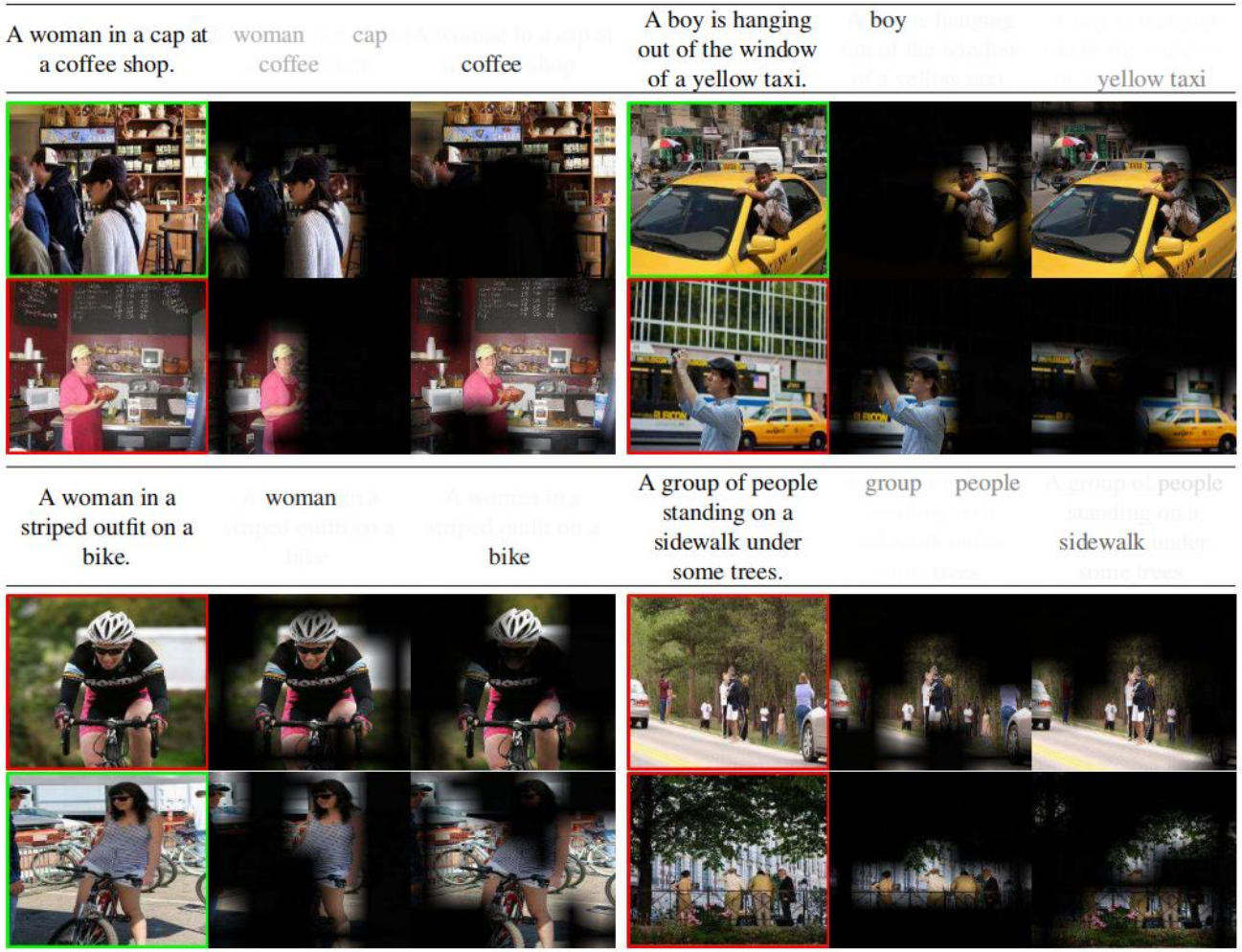


Figure 7: Qualitative results from text-to-image retrieval with attention visualization. For each example, the query sentence and the top two retrieved images are shown from top to bottom; the original sentence (image), the first and second attention maps are shown from left to right. Green and red boxes indicate ground-truth and non ground-truth images, respectively.

Figure 6: Qualitative results from image-to-text retrieval with attention visualization. For each example, the query image and the top two retrieved sentences are shown from top to bottom; the original image (sentence), the first and second attention maps are shown from left to right. (+) and (-) indicate ground-truth and non ground-truth sentences, respectively.

- Image-to-text retrieval

河海大学

向量索引

向量索引方法

基于树的方法

- KD树
- 在空间维度比较低的时候，KD树是比较高效的，当空间维度较高时，其时间复杂度会非常大。

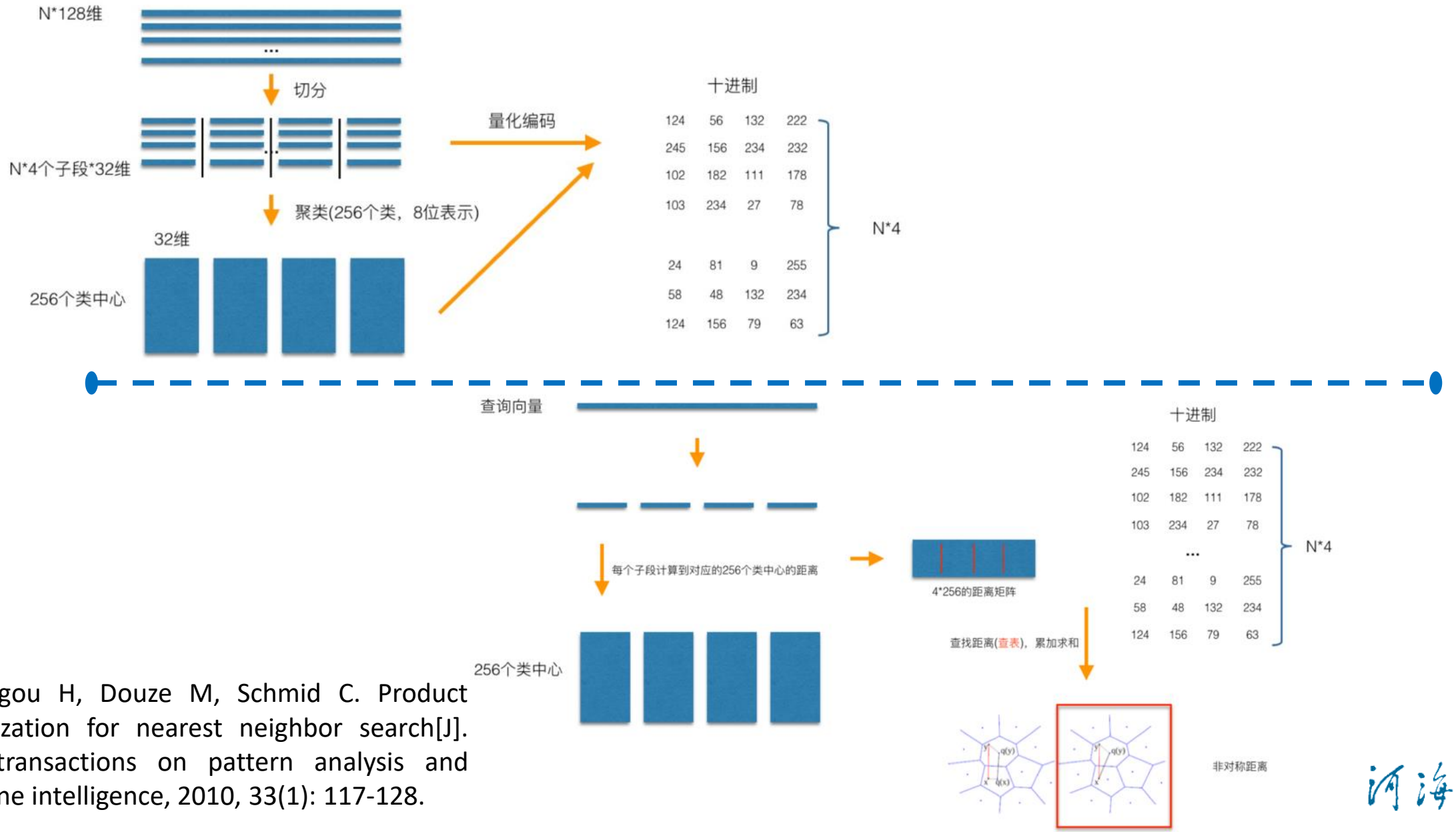
基于哈希的方法

- LSH（Local Sensitive Hashing）
- 将连续的实值散列化为0、1的离散值。

基于矢量量化的方法（vector quantization）

- PQ（Product Quantization）
- 将一个向量空间中的点用其中的一个有限子集来进行编码的过程。

乘积量化



[4] Jegou H, Douze M, Schmid C. Product quantization for nearest neighbor search[J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 33(1): 117-128.

文献三

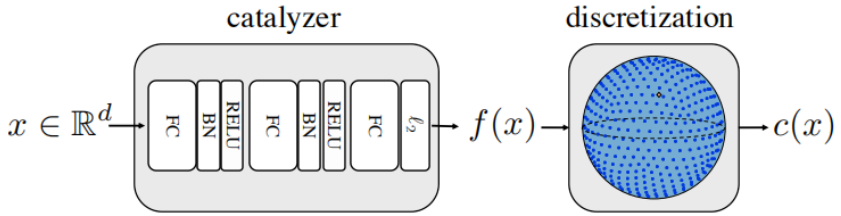


Figure 1: Our method learns a network that encodes the input space $\mathbb{R}^d$ into a code $c(x)$. It is learned end-to-end, yet the part of the network in charge of the discretization operation is fixed in advance, thereby avoiding optimization problems. The learnable function f, namely the "catalyzer", is optimized to increase the quality of the subsequent coding stage.

$$loss = loss_{rank} + \lambda loss_{KoLeo}$$

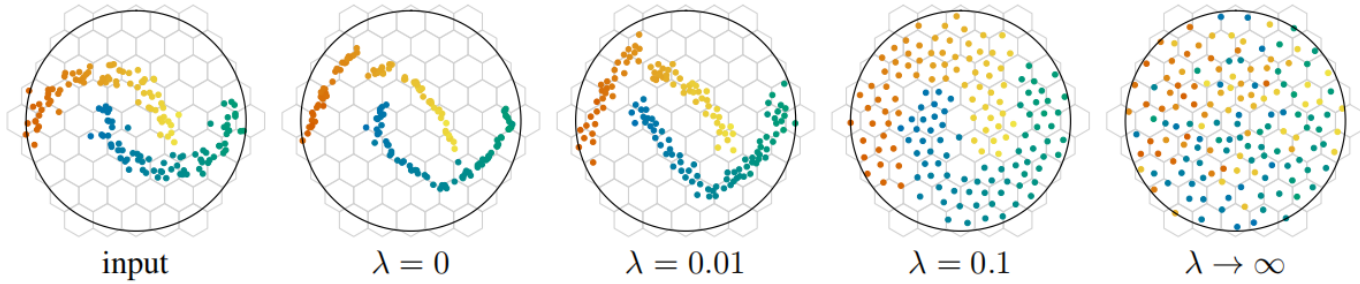


Figure 2: Illustration of our method, which takes as input a set of samples from an unknown distribution. We learn a neural network that aims at preserving the neighborhood structure in the input space while best covering the output space (uniformly). This trade-off is controlled by a parameter λ. The case $\lambda = 0$ keeps the locality of the neighbors but does not cover the output space. On the opposite, when the loss degenerates to the differential entropic regularizer ($\lambda \to \infty$), the neighbors are not maintained by the mapping. Intermediate values offer different trade-offs between neighbor fidelity and uniformity, which is proper input for an efficient lattice quantizer (depicted here by the hexagonal lattice A_2).

[5] Sablayrolles A, Douze M, Schmid C, et al. Spreading vectors for similarity search[C]. ICLR, 2019.

河海大学

谢谢聆听，敬请指正！