Transitive Hashing Network for Heterogeneous Multimedia Retrieval

Zhangjie Cao[1], Mingsheng Long[1], Jianmin Wang[1], and Qiang Yang[2]

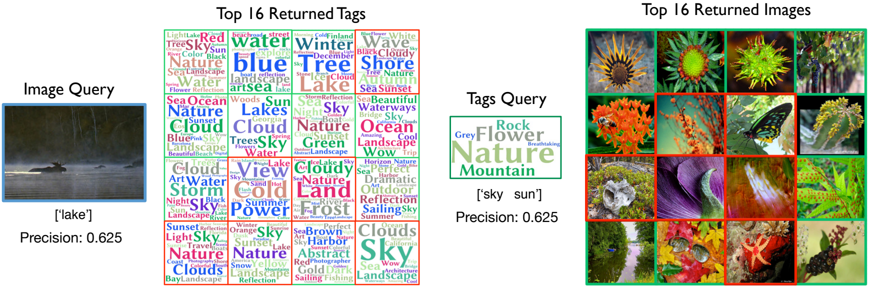
[1]School of Software
Tsinghua University

[2]Department of Computer Science
Hong Kong University of Science and Technology

The Thirty-first AAAI Conference on Artificial Intelligence, 2017

Cross-modal Retrieval

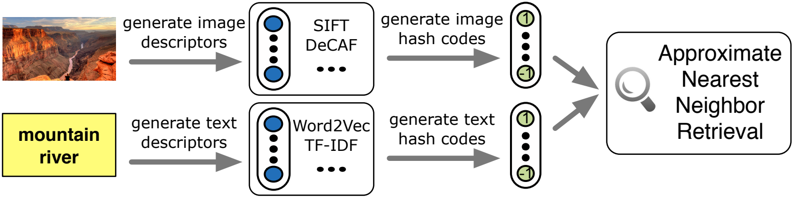
- Nearest Neighbor (NN) similarity retrieval across modalities
  - Database: $\mathcal{X}^{img} = \{\mathbf{x}_1^{img}, \ldots, \mathbf{x}_N^{img}\}$ and Query: $\mathbf{q}^{txt}$
  - Cross-modal NN: $\mathrm{NN}\left(\mathbf{q}^{txt}\right) = \min_{\mathbf{x}^{img} \in \mathcal{X}^{img}} d\left(\mathbf{x}^{img}, \mathbf{q}^{txt}\right)$



(a) $I \rightarrow T$ (Image Query on Text DB) (b) $T \rightarrow I$ (Text Query on Image DB)

Figure: Cross-modal retrieval: similarity retrieval across media modalities.

Hashing Methods



Superiorities

Memory

- 128-d float : 512 bytes → 16 bytes
- 1 billion items : 512 GB → 16 GB

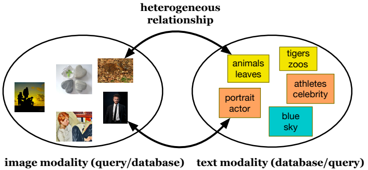
Time

- Computation: x10 - x100 faster
- Transmission (disk / web): x30 faster

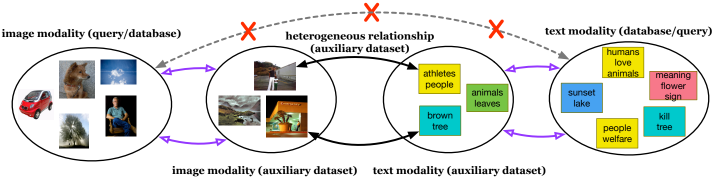
Applications

- Approximate nearest neighbor search
- Compact representation, Feature Compression for large datasets
- Distribute and transmit data online
- Construct index for large-scale database

Traditional VS. Transitive

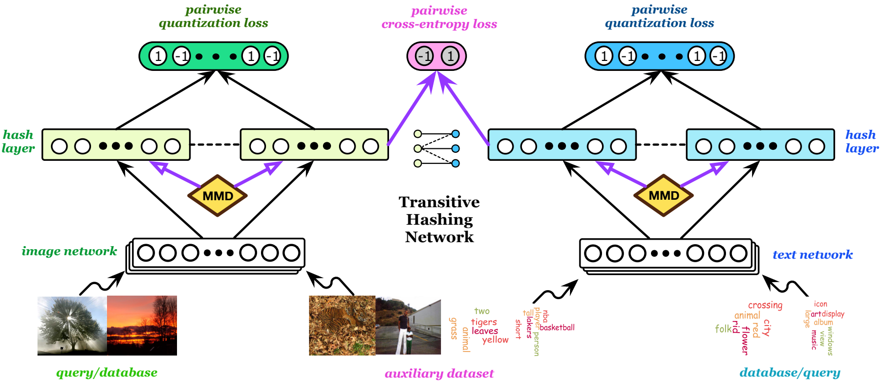


traditional cross-modal hashing

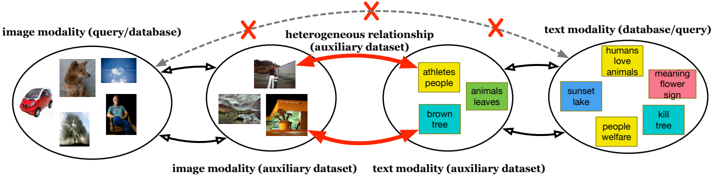


transitive cross-modal hashing

Network Architecture



Heterogeneous Relationship Learning



Given heterogeneous relationship $\mathcal{S} = \{s_{ij}\}$,

Logarithm Maximum a Posteriori estimation

$$\begin{aligned} \log p\left(\boldsymbol{H}^x, \boldsymbol{H}^y | \mathcal{S}\right) &\propto \log p\left(\mathcal{S} | \boldsymbol{H}^x, \boldsymbol{H}^y\right) p\left(\boldsymbol{H}^x\right) p\left(\boldsymbol{H}^y\right) \\ &= \sum_{s_{ij} \in \mathcal{S}} \log p\left(s_{ij} | \boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right) p\left(\boldsymbol{h}_i^x\right) p\left(\boldsymbol{h}_j^y\right), \end{aligned} \quad (1)$$

where $p(\mathcal{S}|\boldsymbol{H}^x, \boldsymbol{H}^y)$ is likelihood function, and $p(\boldsymbol{H}^x)$ and $p(\boldsymbol{H}^y)$ are prior distributions.

Heterogeneous Relationship Learning

Likelihood function

For each pair of points $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$, $p(s_{ij}|\boldsymbol{h}_i^x, \boldsymbol{h}_j^y)$ is the conditional probability of their relationship s_{ij} given their hash codes $\boldsymbol{h}_i^x$ and $\boldsymbol{h}_j^y$, which can be defined using the pairwise logistic function,

$$p\left(s_{ij}|\boldsymbol{h}_i^x, \boldsymbol{h}_j^y\right) = \begin{cases} \sigma\left(\left\langle \boldsymbol{h}_i^x, \boldsymbol{h}_j^y \right\rangle\right), & s_{ij} = 1 \\ 1 - \sigma\left(\left\langle \boldsymbol{h}_i^x, \boldsymbol{h}_j^y \right\rangle\right), & s_{ij} = 0 \end{cases}$$
$$= \sigma\left(\left\langle \boldsymbol{h}_i^x, \boldsymbol{h}_j^y \right\rangle\right)^{s_{ij}} \left(1 - \sigma\left(\left\langle \boldsymbol{h}_i^x, \boldsymbol{h}_j^y \right\rangle\right)\right)^{1-s_{ij}}, \quad (2)$$

where $\sigma(x) = 1/(1+e^{-x})$ is the sigmoid function and $\boldsymbol{h}_i^x = \operatorname{sgn}(\boldsymbol{z}_i^x)$ and $\boldsymbol{h}_i^y = \operatorname{sgn}(\boldsymbol{z}_i^y)$.

Heterogeneous Relationship Learning

Prior

For ease of optimization, continuous relaxation that $\boldsymbol{h}_i^x = \boldsymbol{z}_i^x$ and $\boldsymbol{h}_i^y = \boldsymbol{z}_i^y$ is applied to the binary constraints.
Then, to control the quantization error and close the gap between Hamming distance and its surrogate for learning accurate hash codes, a new cross-entropy prior is proposed over the continuous activations $\{\boldsymbol{z}_i^*\}$ as

$$p\left(\boldsymbol{z}_i^*\right) \propto \exp\left(-\lambda H\left(\frac{\mathbf{1}}{b}, \frac{|\boldsymbol{z}_i^*|}{b}\right)\right), \tag{3}$$

where $* \in \{x, y\}$, and λ is the parameter of the exponential distribution.

Heterogeneous Relationship Learning

Optimization problem for heterogeneous relationship

$$\min_{\Theta} J = L + \lambda Q, \tag{4}$$

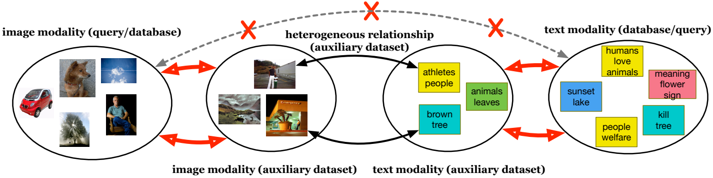
where λ is the trade-off parameter between pairwise cross-entropy loss L and pairwise quantization loss Q, and Θ is network parameters.

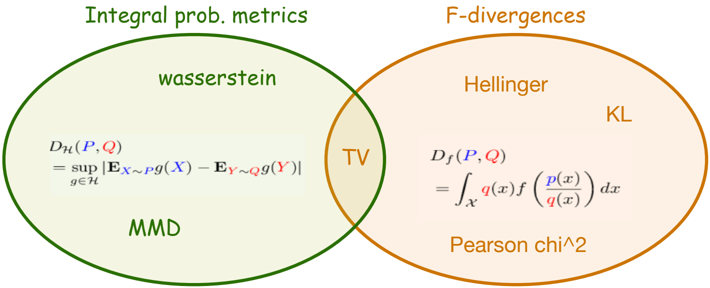
Specifically,

$$L = \sum_{s_{ij} \in \mathcal{S}} \log \left(1 + \exp\left(\left\langle \mathbf{z}_i^x, \mathbf{z}_j^y \right\rangle\right)\right) - s_{ij} \left\langle \mathbf{z}_i^x, \mathbf{z}_j^y \right\rangle. \tag{5}$$

$$Q = \sum_{s_{ij} \in \mathcal{S}} \sum_{k=1}^{b} (-\log(|z_{ik}^x|) - \log(|z_{jk}^y|)). \tag{6}$$

Homogeneous Distribution Alignment





Homogeneous Distribution Alignment

Maximum Mean Discrepancy (MMD) [jmlr 12']

MMD is a nonparametric distance measure to compare different distributions P_q and P_x in reproducing kernel Hilbert space $\mathcal{H}$ (RKHS) endowed with feature map ϕ and kernel k, formally defined as $D_q \triangleq \left\| \mathbb{E}_{\boldsymbol{h}^q \sim P_q} \left[\phi \left(\boldsymbol{h}^q \right) \right] - \mathbb{E}_{\boldsymbol{h}^x \sim P_x} \left[\phi \left(\boldsymbol{h}^x \right) \right] \right\|_{\mathcal{H}}^2$, where P_q and P_x are the distribution of the query set $\mathcal{X}^q$, and the auxiliary set $\bar{\mathcal{X}}$.

MMD between auxiliary dataset $\bar{\mathcal{X}}$ and query set $\mathcal{X}^q$

$$D_q = \sum_{i=1}^{\hat{n}} \sum_{j=1}^{\hat{n}} \frac{k\left(\boldsymbol{z}_i^q, \boldsymbol{z}_j^q\right)}{\hat{n}^2} + \sum_{i=1}^{\bar{n}} \sum_{j=1}^{\bar{n}} \frac{k\left(\boldsymbol{z}_i^x, \boldsymbol{z}_j^x\right)}{\bar{n}^2} - 2 \sum_{i=1}^{\hat{n}} \sum_{j=1}^{\bar{n}} \frac{k\left(\boldsymbol{z}_i^q, \boldsymbol{z}_j^x\right)}{\hat{n}\bar{n}}, \tag{7}$$

where $k(\boldsymbol{z}_i, \boldsymbol{z}_j) = \exp(-\gamma ||\boldsymbol{z}_i - \boldsymbol{z}_j||^2)$ is the Gaussian kernel.

Transitive Hashing Network

Unified optimization problem

$$\min_{\Theta} C = J + \mu \left(D_q + D_d \right), \tag{8}$$

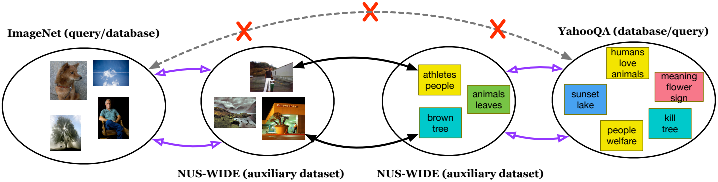
where μ is a trade-off parameter between the MAP loss J and the MMD penalty $(D_q + D_d)$.

Extensions

When the auxiliary set is small, we assume that the auxiliary set is related to database and query sets. However, if the auxiliary set is large, this requirement can be aborted since it's common that the auxiliary set has relationship with other sets. Thus, we can use a large set as auxiliary set in any task. We can even use relationship model pre-trained from large-scale datasets and fine-tune it with our homogeneous alignment method. This widely-used pre-training and fine-tuning strategy makes our method more easily deployable.

Experiments Setup

- **Datasets:**



- **Protocols:** MAPs, Precision-Recall Curve
- **Parameter selection:** cross-validation by jointly assessing
- **Methods to compare with:** two unsupervised methods Cross-View Hashing (CVH) and Inter-Media Hashing (IMH), two supervised methods Quantized Correlation Hashing (QCH) and Heterogeneous Translated Hashing (HTH), and one deep hashing method Deep Cross-Modal Hashing (DCMH).

Results and Discussion

Table: MAP Comparison of Cross-Modal Retrieval Tasks on NUS-WIDE and ImageNet-YahooQA

Task	Method	NUS-WIDE				ImageNet-YahooQA			
		8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
$I \rightarrow T$	IMH	0.5821	0.5794	0.5804	0.5776	0.0855	0.0686	0.0999	0.0889
	CVH	0.5681	0.5606	0.5451	0.5558	0.1229	0.1180	0.0941	0.0865
	QCH	0.6463	0.6921	0.7019	0.7127	0.2563	0.2494	0.2581	0.2590
	HTH	0.5232	0.5548	0.5684	0.5325	0.2931	0.2694	0.2847	0.2663
	DCMH	<u>0.7887</u>	<u>0.7397</u>	<u>0.7210</u>	<u>0.7460</u>	<u>0.5133</u>	<u>0.5109</u>	<u>0.5321</u>	<u>0.5087</u>
	THN	**0.8252**	**0.8423**	**0.8495**	**0.8572**	**0.5451**	**0.5507**	**0.5803**	**0.5901**
$T \rightarrow I$	IMH	0.5579	0.5593	0.5528	0.5457	0.1105	0.1044	0.1183	0.0909
	CVH	0.5261	0.5193	0.5097	0.5045	0.0711	0.0728	0.1116	0.1008
	QCH	0.6235	0.6609	0.6685	0.6773	0.2761	0.2847	0.2795	0.2665
	HTH	0.5603	0.5910	0.5798	0.5812	0.2172	0.1702	0.3122	0.2873
	DCMH	<u>0.7882</u>	<u>0.7912</u>	<u>0.7921</u>	<u>0.7718</u>	<u>0.5163</u>	<u>0.5510</u>	<u>0.5581</u>	<u>0.5444</u>
	THN	**0.7905**	**0.8137**	**0.8245**	**0.8268**	**0.6032**	**0.6097**	**0.6232**	**0.6102**

Results and Discussion

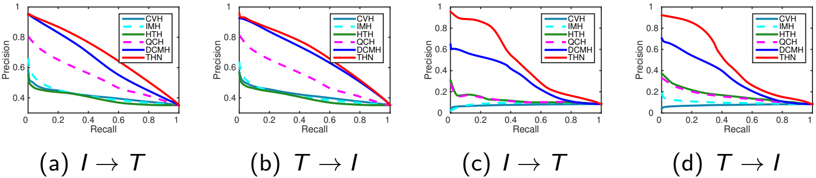


Figure: Precision-recall curves of Hamming ranking @ 24-bits codes on NUS-WIDE (a)-(b) and ImageNet-YahooQA (c)-(d).

Key Observations

- THN is a new state of the art method for the more conventional cross-modal retrieval problems where the relationship between query and database is available for training as in the NUS-WIDE dataset.
- The homogeneous distribution alignment module of THN effectively closes this shift by matching the corresponding data distributions with the maximum mean discrepancy.

Empirical Analysis

THN-ip: *is the variant which uses the pairwise inner-product loss instead of the pairwise cross-entropy loss*
THN-D: *is the variant without using the unsupervised training data*
THN-Q: *is the variant without using the quantization loss*

Table: MAP of THN variants on ImageNet-YahooQA

Method	$I \to T$				$T \to I$			
	8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
THN-ip	0.2976	0.3171	0.3302	0.3554	0.3443	0.3605	0.3852	0.4286
THN-D	<u>0.5192</u>	0.5123	0.5312	<u>0.5411</u>	0.5423	0.5512	0.5602	0.5489
THN-Q	0.4821	<u>0.5213</u>	<u>0.5352</u>	0.4947	<u>0.5731</u>	<u>0.5592</u>	<u>0.5849</u>	<u>0.5612</u>
THN	**0.5451**	**0.5507**	**0.5803**	**0.5901**	**0.6032**	**0.6097**	**0.6232**	**0.6102**

Empirical Analysis

Key Observations

- THN outperforms THN-ip by very large margins, confirming the importance of well-defined loss functions for heterogeneous relationship learning.
- THN outperforms THN-D, which convinces that THN can further exploit the unsupervised training data to bridge the Hamming spaces of auxiliary dataset (NUS-WIDE) and query/database sets (ImageNet-YahooQA) such that the auxiliary dataset can be leveraged as a bridge to transfer knowledge between query and database.
- THN outperforms THN-Q, indicating pairwise quantization loss can reduce the quantization errors when binarizing continuous representations to hash codes.

Summary

- We designe a new transitive deep hashing problem for heterogeneous multimedia retrieval without direct relationship between query set and database set.
- We propose a pairwise cross-entropy and a pairwise quantization loss for heterogeneous relationship learning.
- We achieve homogeneous distribution alignment by minizing MMD loss between query/database set and auxiliary set.
- In the future, we plan to extend the method to social media problems.