Deep Cross-Modal Hashing

Qing-Yuan Jiang & Wu-Jun Li

LAMDA Group, Department of Computer Science and Technology, Nanjing University, Nanjing, China.

jiangqy@lamda.nju.edu.cn, liwujun@nju.edu.cn

NANJING UNIVERSITY

Learning And Mining from DatA

CVPR July 21-26 HONOLULU 2017

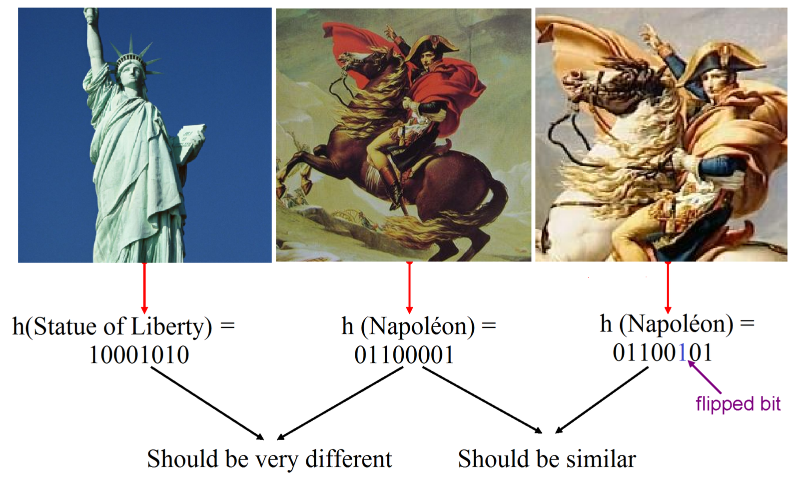
Introduction

Nearest Neighbor Search (NNS)

- Given a query point q, return the points closest to q in the database (e.g., image retrieval).
- Challenges for NNS in big data applications: curse of dimensionality; storage cost; query speed

Hashing

- Similarity preserved hashing is to map the data points from the original space into a Hamming space of binary codes with similarity preserved.
- Hashing can solve the above challenges.



Cross-Modal Hashing (CMH)

- Cross-modal retrieval: the modality of the query point is different from the modality of the points in database.
- CMH: hashing for cross-modal retrieval. Low storage cost and fast query speed.

Motivation

- Almost all existing CMH methods are based on hand-crafted features.
- Hand-crafted features might not be compatible for hash-code learning.

Contribution

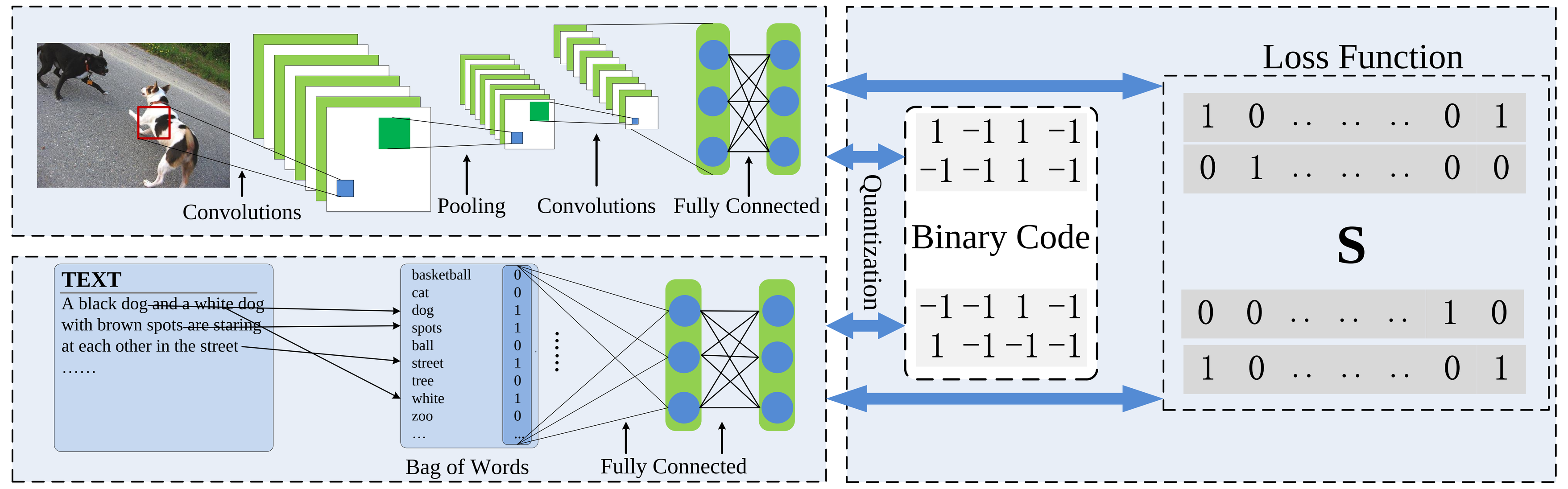
- A novel CMH method, called deep cross-modal hashing (DCMH), for cross-modal retrieval applications.
- DCMH is an end-to-end learning framework with deep neural networks, one for each modality, to perform feature learning from scratch.
- DCMH achieves the state-of-the-art performance on three datasets.

Notation

- $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$: n points of image modality.
- $\mathbf{Y} = \{\mathbf{y}_j\}_{j=1}^n$: n points of text modality.
- $\mathbf{S} = \{S_{ij}\}_{n\times n}$: cross-modal similarities.
- $f(\mathbf{x}_i;\theta_x)$: the output of deep neural network for image modality.
- $g(\mathbf{y}_j;\theta_y)$: the output of deep neural network for text modality.

Model

The end-to-end deep learning framework of DCMH model.



- Feature learning part:

This part contains two deep neural networks, one for image modality and the other for text modality. Their configurations are shown in the following tables.

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

Configuration of the deep neural network for text modality.

Layer	Configuration
full1	8192
full2	Hash code length c

Deep neural network input:

- Image deep neural network: raw image.
- Text deep neural network: Bag-of-words (BOW) feature.

- Hash-code learning part:

$$\min_{\mathbf{B},\theta_x,\theta_y} \mathcal{J} = -\sum\nolimits_{i,j=1}^{n}(S_{ij}\Theta_{ij} - \log(1+e^{\Theta_{ij}})) + \gamma(\|\mathbf{B}-\mathbf{F}\|_F^2 + \|\mathbf{B}-\mathbf{G}\|_F^2) + \eta(\|\mathbf{F1}\|_F^2 + \|\mathbf{G1}\|_F^2)$$
$$s.t.\ \mathbf{B}\in\{-1,+1\}^{c\times n}.$$

- $\mathbf{B}\in\{-1,+1\}^{c\times n}$: binary codes, where c is the code length.
- $\mathbf{F}\in\mathbb{R}^{c\times n}$ with $\mathbf{F}_{*i} = f(\mathbf{x}_i;\theta_x)$.
- $\mathbf{G}\in\mathbb{R}^{c\times n}$ with $\mathbf{G}_{*j} = g(\mathbf{y}_j;\theta_y)$.
- $\Theta_{ij} = \frac{1}{2}\mathbf{F}_{*i}^T\mathbf{G}_{*j}$.

Learning

Alternating Learning Algorithm

- Learn θ_x, with θ_y and $\mathbf{B}$ fixed.

BP for updating θ_x. For each sampled point $\mathbf{x}_i$, compute the gradient:

$$\frac{\partial\mathcal{J}}{\partial\mathbf{F}_{*i}} = \frac{1}{2}\sum\nolimits_{j=1}^{n}(\sigma(\Theta_{ij})\mathbf{G}_{*j} - S_{ij}\mathbf{G}_{*j}) + 2\gamma(\mathbf{F}_{*i}-\mathbf{B}_{*i}) + 2\eta\mathbf{F1}.$$

- Learn θ_y, with θ_x and $\mathbf{B}$ fixed.

BP for updating θ_y. For each sampled point $\mathbf{y}_j$, compute the gradient:

$$\frac{\partial\mathcal{J}}{\partial\mathbf{G}_{*j}} = \frac{1}{2}\sum\nolimits_{i=1}^{n}(\sigma(\Theta_{ij})\mathbf{F}_{*i} - S_{ij}\mathbf{F}_{*i}) + 2\gamma(\mathbf{G}_{*j}-\mathbf{B}_{*j}) + 2\eta\mathbf{G1}.$$

- Learn $\mathbf{B}$, with θ_x and θ_y fixed.

$$\mathbf{B} = \text{sign}(\gamma(\mathbf{F}+\mathbf{G})).$$

Experiment

Datasets

- MIRFLICKR-25K: 25,000 image-text pairs which are annotated with one of the 24 unique labels.
- IAPR TC-12: 20,000 image-text pairs which are annotated using 255 labels.
- NUS-WIDE: 260,648 image-text pairs. Each point is annotated with one or multiple labels from 81 concept labels. We select 195,834 image-text pairs that belong to the 21 most frequent concepts.
- For MIRFLICKR-25K and IAPR TC-12: 2000/10000 test/training points. For NUS-WIDE: 2100/10500 test/training points.

Hamming Ranking Task (Mean Average Precision)

Comparison to State-of-the-Art Baselines

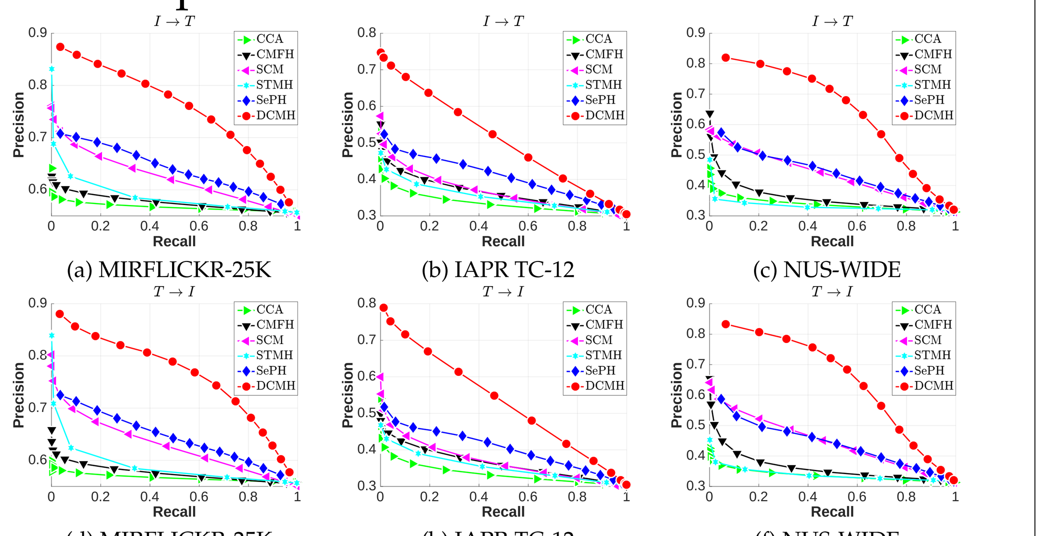
Task	Method	MIRFLICKR-25K			IAPR TC-12			NUS-WIDE		
		16	32	64	16	32	64	16	32	64
$I\to T$	DCMH	**.741**	**.747**	**.749**	**.453**	**.473**	**.484**	**.590**	**.603**	**.609**
	SePH	.657	.660	.662	.411	.416	.420	.479	.487	.489
	STMH	.592	.595	.598	.358	.373	.382	.397	.408	.415
	SCM	.629	.640	.648	.383	.390	.388	.465	.471	.482
	CMFH	.582	.581	.581	.368	.373	.379	.357	.362	.366
	CCA	.570	.566	.564	.335	.325	.319	.341	.334	.328
$T\to I$	DCMH	**.783**	**.790**	**.793**	**.519**	**.538**	**.547**	**.639**	**.651**	**.657**
	SePH	.648	.652	.655	.402	.407	.413	.449	.454	.459
	STMH	.580	.585	.586	.345	.357	.369	.361	.374	.384
	SCM	.620	.630	.637	.370	.373	.370	.437	.443	.450
	CMFH	.579	.577	.578	.362	.369	.377	.362	.367	.372
	CCA	.569	.566	.564	.334	.326	.320	.339	.332	.327

Comparison to Baselines with CNN-F Features

Task	Method	MIRFLICKR-25K			IAPR TC-12			NUS-WIDE		
		16	32	64	16	32	64	16	32	64
$I\to T$	DCMH	**.741**	**.747**	**.749**	**.453**	**.473**	**.484**	.590	.603	.609
	SePH	.712	.719	.723	.444	.456	.464	**.604**	**.617**	**.621**
	STMH	.613	.622	.627	.378	.400	.413	.471	.486	.494
	SCM	.685	.692	.700	.369	.367	.380	.541	.549	.555
	CMFH	.638	.642	.645	.419	.423	.425	.490	.505	.510
	CCA	.572	.569	.567	.342	.336	.330	.360	.349	.339
$T\to I$	DCMH	**.783**	**.790**	**.793**	**.519**	**.538**	**.547**	**.639**	**.651**	**.657**
	SePH	.722	.726	.732	.442	.456	.465	.598	.603	.611
	STMH	.607	.615	.622	.369	.390	.404	.447	.468	.478
	SCM	.694	.701	.706	.345	.341	.347	.534	.541	.548
	CMFH	.637	.640	.643	.417	.421	.428	.503	.519	.523
	CCA	.574	.571	.569	.349	.344	.338	.361	.349	.340

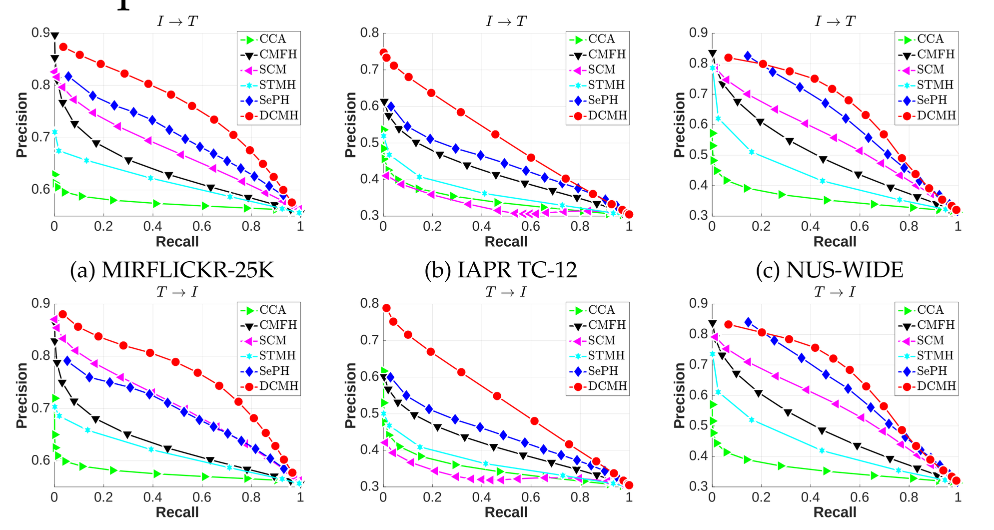
Hash Lookup Task (Precision Recall Curve)

Comparison to State-of-the-Art Baselines



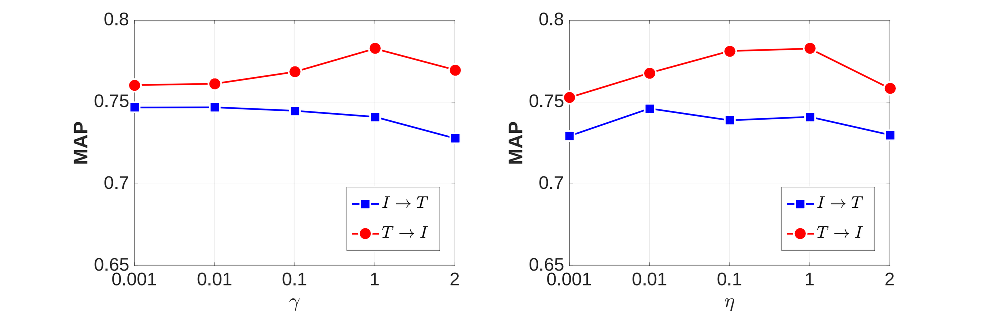
(a) MIRFLICKR-25K (b) IAPR TC-12 (c) NUS-WIDE (d) MIRFLICKR-25K (e) IAPR TC-12 (f) NUS-WIDE

Comparison to Baselines with CNN-F Features

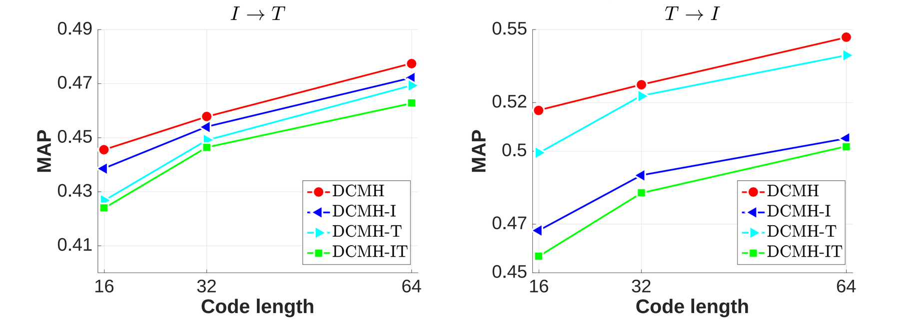


(a) MIRFLICKR-25K (b) IAPR TC-12 (c) NUS-WIDE (d) MIRFLICKR-25K (e) IAPR TC-12 (f) NUS-WIDE

Sensitivity to Parameters



Further Analysis



Conclusion

- DCMH is an end-to-end deep learning framework which can perform simultaneous feature learning and hash-code learning.
- DCMH can significantly outperform other baselines to achieve the state-of-the-art performance.