Correlation Autoencoder Hashing for Supervised Cross-Modal Search

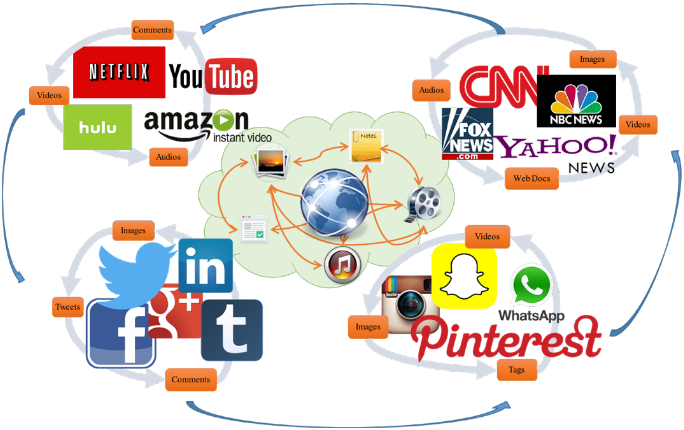
Yue Cao, Mingsheng Long, Jianmin Wang, and Han Zhu

School of Software
Tsinghua University

The Annual ACM International Conference on Multimedia Retrieval
ICMR 2016

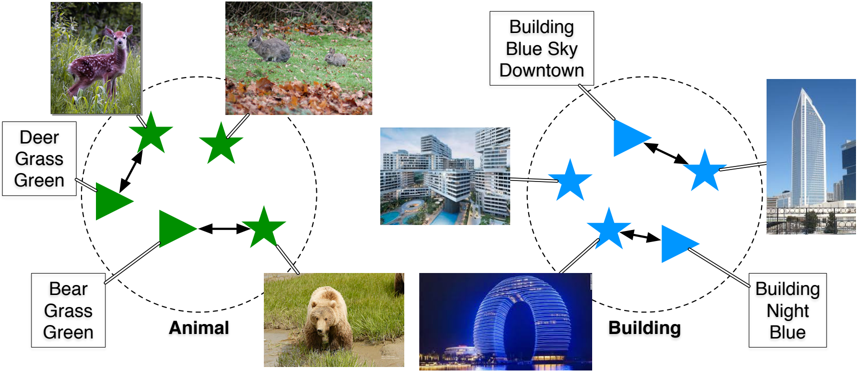
Background

- In the big data era, the amount of *multimedia* data has exploded
- An object or topic can be described by data of multiple modalities



Cross-Modal Similarity Search

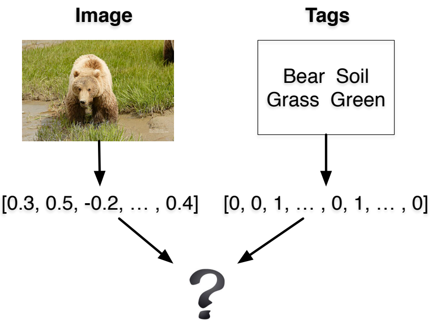
- Use a query from one modality to search for semantically relevant items from another modality
  - e.g. search for animal images using textual tags 'bear, deer …'



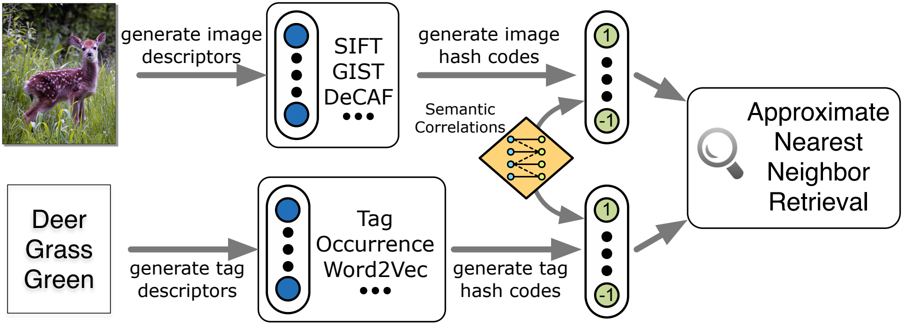
Challenges

- Trillions of images and texts are generated

- Features from different modalities are heterogeneous
  - Different dimensions
  - Distinct distributions
  - ...



Cross-Modal Hashing



Memory

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

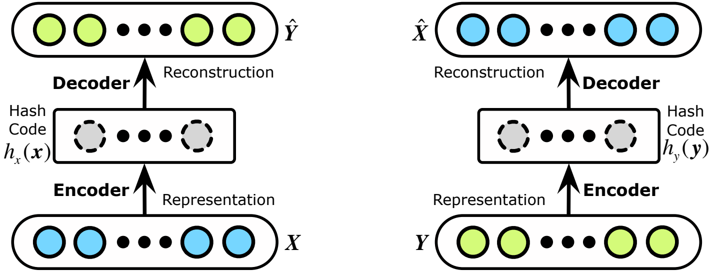
Time

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

Homogeneous Architecture

Key Points

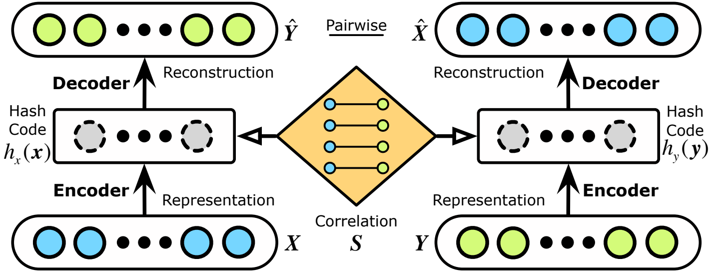
- Homogeneous Architecture: image and text can use the same deep architecture



Feature Correlations

Key Points

- Feature correlations can be maximized to reduce heterogeneity across modalities, using pairwise correlations (solid lines)



Feature Correlation Maximization

Key Points

- Use pairwise correlations for reconstructive embedding

Within-modal Reconstructive Embedding

$$\min_{\mathbf{V}_x, \mathbf{V}_y} \sum_{i=1}^{n} \left(\|\mathbf{x}_i - \mathbf{V}_x h_x(\mathbf{x}_i)\|_2^2 + \|\mathbf{y}_i - \mathbf{V}_y h_y(\mathbf{y}_i)\|_2^2 \right), \tag{1}$$

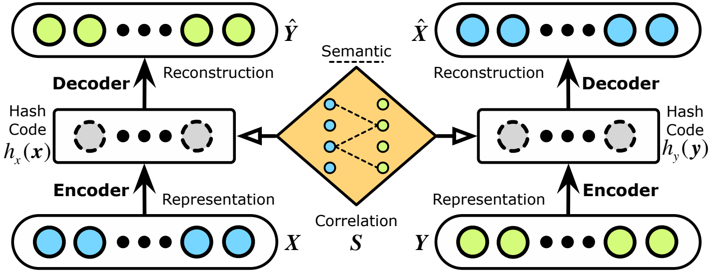
Cross-modal Reconstructive Embedding

$$\min_{\mathbf{V}_x, \mathbf{V}_y} L = \sum_{i=1}^{n} \left(\|\mathbf{x}_i - \mathbf{V}_x h_y(\mathbf{y}_i)\|_2^2 + \|\mathbf{y}_i - \mathbf{V}_y h_x(\mathbf{x}_i)\|_2^2 \right), \tag{2}$$

Semantic Correlations

Key Points

- Due to semantic gap, semantic correlations (dashed lines) need to be maximized



Semantic Correlation Maximization

Key Points

- Construct a Nearest Neighbor Affinity Matrix $\mathbf{A}$

Nearest Neighbor Affinity Matrix

$$A_{ij} = \begin{cases} d\left(\mathbf{x}_i, \mathbf{y}_j\right), & \text{if } \mathbf{l}_i = \mathbf{l}_j \wedge \begin{cases} \mathbf{x}_i \in \mathcal{N}_k\left(\mathbf{x}_j\right) \vee \mathbf{x}_j \in \mathcal{N}_k\left(\mathbf{x}_i\right) \\ \mathbf{y}_i \in \mathcal{N}_k\left(\mathbf{y}_j\right) \vee \mathbf{y}_j \in \mathcal{N}_k\left(\mathbf{y}_i\right) \end{cases} \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

$$d\left(\mathbf{x}_i, \mathbf{y}_j\right) = e^{-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2 / 2\sigma_x^2} + e^{-\left\|\mathbf{y}_i - \mathbf{y}_j\right\|_2^2 / 2\sigma_y^2} \tag{4}$$

where $\mathcal{N}_k(\mathbf{x})$ represents the k-nearest neighbors of $\mathbf{x}$.

Semantic Correlation Maximization

Key Points

- Construct a within-category and a between-category similarity matrix

Similarity Matrices

$$
\begin{aligned}
S_{ij}^{b} &= \begin{cases} A_{ij}\left(1/n - 1/n_c\right), & \text{if } \mathbf{l}_i = \mathbf{l}_j = c \\ A_{ij}/n, & \text{if } \mathbf{l}_i \neq \mathbf{l}_j, \end{cases} \\
S_{ij}^{w} &= \begin{cases} A_{ij}/n_c, & \text{if } \mathbf{l}_i = \mathbf{l}_j = c \\ 0, & \text{if } \mathbf{l}_i \neq \mathbf{l}_j, \end{cases}
\end{aligned} \tag{5}
$$

where n_c is the number of objects within the c-th category.

Semantic Correlation Maximization

Key Points

- Maximize the inter-category separation margin
- Circumvent the large intra-class variance

Cross-modal Semantic Correlation

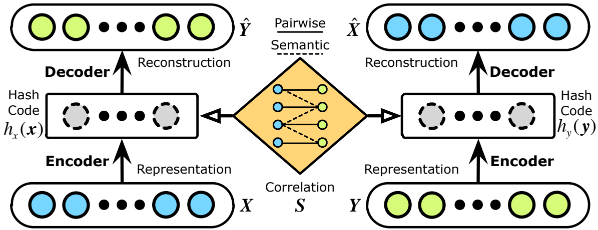
$$\min_{\mathbf{W}_x, \mathbf{W}_y} R = \sum_{i=1}^{n} \sum_{j=1}^{n} S_{ij} \left\| h_x(\mathbf{x}_i) - h_y(\mathbf{y}_j) \right\|_2^2, \quad (6)$$

$$S_{ij} = \begin{cases} A_{ij}\left(2/n_c - 1/n\right), & \text{if } \mathbf{l}_i = \mathbf{l}_j = c \\ -A_{ij}/n, & \text{if } \mathbf{l}_i \neq \mathbf{l}_j. \end{cases} \quad (7)$$

Correlation Autoencoder Hashing

Key Points

- enhances feature correlation by cross-modal reconstruct embedding
- maximizes the inter-category separation margin for learning more discriminative hash codes
- minimizes the intra-category variance by further exploring the cross-modal locality information



Correlation Autoencoder Hashing

Unified Optimization Problem

$$\min_{\mathbf{V}_x,\mathbf{V}_y,\mathbf{W}_x,\mathbf{W}_y} O = L + \lambda R$$
$$h_x(\mathbf{x}) = \operatorname{sgn}\left(\mathbf{W}_x^{\mathsf{T}}\mathbf{x}\right), h_y(\mathbf{y}) = \operatorname{sgn}\left(\mathbf{W}_y^{\mathsf{T}}\mathbf{y}\right), \tag{8}$$

where λ is a penalty parameter for trading off the relative importance of feature correlation and semantic correlation.

Learning Algorithm

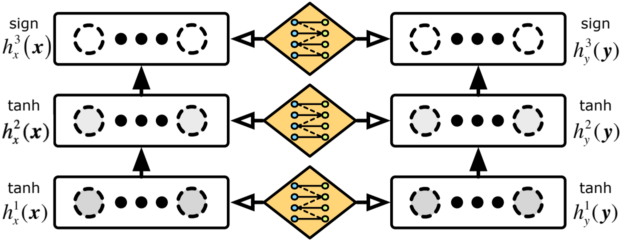
By back-propagation (BP) using mini-batch SGD

$$\frac{\partial O(\mathbf{x}_i, \mathbf{y}_i)}{\partial \mathbf{W}_{pq}^{x}} = \frac{\partial L(\mathbf{y}_i)}{\partial \mathbf{W}_{pq}^{x}} + \lambda \frac{\partial R(\mathbf{x}_i)}{\partial \mathbf{W}_{pq}^{x}}, \tag{9}$$

Deep Architecture

Key Points

- A three-layer stacked auto-encoder architecture
- The feature correlations and semantic correlations are distilled in each layer and can be strengthened layer by layer
- Use hyperbolic tangent function *tanh* as the activation function to reduce the large binarization loss



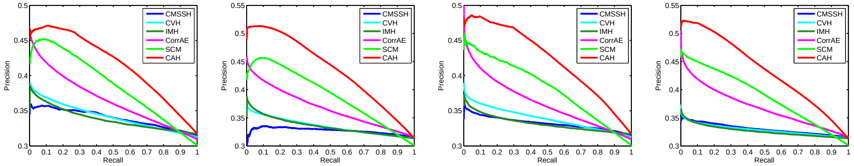
Experiment Setup

- **Datasets:** Nus-wide, Wiki and MIR-Flickr
- **Protocols:** Mean Average Precisions, Precision-Recall Curves
- **Parameter selection:** cross-validation
- **Comparison Methods**
  - Unsupervised Shallow Hashing: IMH
  - Supervised Shallow Hashing: SCM
  - Unsupervised Deep Hashing: CorrAE + Sign
  - Supervised Deep Hashing: Our approach CAH
- **Variants**
  - CAH only with feature correlation (CAH-F)
  - CAH without using data locality (CAH-L)

Results and Discussion [Nus-wide]

- CAH outperforms unsupervised deep hashing (CorrAE), supervised hashing (SCM) and unsupervised shallow hashing (IMH).
- CAH also outperforms CAH-F and CAH-L, verifying the vital importance of every component newly-crafted in this paper.

Dataset	Method	$I \to T$				$T \to I$			
		8 bits	16 bits	32 bits	64 bits	8 bits	16 bits	32 bits	64 bits
Nus-wide	IMH	0.4345	0.4399	0.4203	0.4115	0.4380	0.4582	0.4186	0.4051
	SCM	0.4693	0.4648	0.4619	0.4851	0.4449	0.4859	0.5105	0.5259
	CorrAE	0.4398	0.4522	0.4699	0.4944	0.4303	0.4501	0.4634	0.4880
	CAH-F	0.4439	0.4711	0.4922	0.5234	0.4433	0.4666	0.4885	0.5157
	CAH-L	0.4880	0.5050	0.5219	0.5581	0.4933	0.5053	0.5205	0.5250
	CAH	**0.4920**	**0.5084**	**0.5407**	**0.5628**	**0.5019**	**0.5135**	**0.5451**	**0.5800**

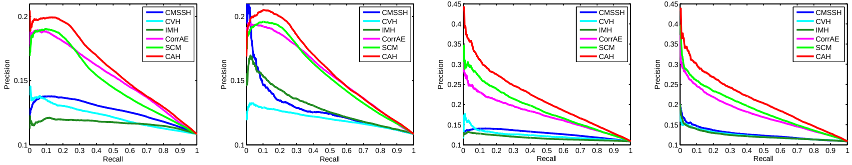


(a) $I \to T$ @ 16 bits (b) $I \to T$ @ 32 bits (c) $T \to I$ @ 16 bits (d) $T \to I$ @ 32 bits

Results and Discussion [Wiki]

- The low quality of the image modality leads to that task $I \rightarrow T$ is more difficult than task $T \rightarrow I$. Almost all the methods achieve better results on task $T \rightarrow I$.

Dataset	Method	$I \rightarrow T$				$T \rightarrow I$			
		8 bits	16 bits	32 bits	64 bits	8 bits	16 bits	32 bits	64 bits
Wiki	IMH	0.1734	0.1896	0.1714	0.1601	0.2394	0.2227	0.2333	0.1896
	SCM	0.2258	0.2372	0.2381	0.2378	0.3157	0.3698	0.4239	0.4369
	CorrAE	0.1990	0.2078	0.2105	0.2177	0.2712	0.2948	0.3111	0.3220
	CAH-F	0.2276	0.2323	0.2233	0.2339	0.2608	0.3311	0.3418	0.3693
	CAH-L	0.2208	0.2342	0.2420	0.2456	0.3302	0.3744	0.4156	0.4325
	CAH	**0.2308**	**0.2415**	**0.2465**	**0.2530**	**0.3424**	**0.3956**	**0.4284**	**0.4569**

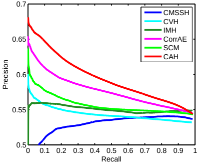


(e) $I \rightarrow T$ @ 16 bits (f) $I \rightarrow T$ @ 32 bits (g) $T \rightarrow I$ @ 16 bits (h) $T \rightarrow I$ @ 32 bits

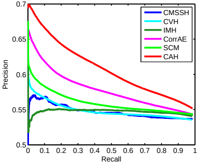
Results and Discussion [MIR-Flickr]

- CAH also achieve the state-of-the-art results on large-scale dataset.

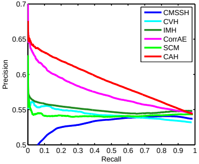
Dataset	Method	$I \to T$				$T \to I$			
		8 bits	16 bits	32 bits	64 bits	8 bits	16 bits	32 bits	64 bits
Flickr	IMH	0.5449	0.5646	0.5936	0.5539	0.5374	0.5536	0.5513	0.5583
	SCM	0.6361	0.6493	0.6495	0.6440	0.6037	0.5998	0.5805	0.6078
	CorrAE	0.6301	0.6329	0.6357	0.6401	0.6142	0.6198	0.6247	0.6431
	CAH-F	0.6493	0.6470	0.6544	0.6786	0.6324	0.6406	0.6508	0.6765
	CAH-L	0.6520	0.6584	0.6710	0.6920	0.6328	**0.6734**	**0.6978**	0.7201
	CAH	**0.6608**	**0.6875**	**0.7035**	**0.7072**	**0.6496**	0.6612	0.6908	**0.7263**



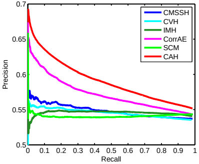
(i) $I \to T$ @ 16 bits



(j) $I \to T$ @ 32 bits



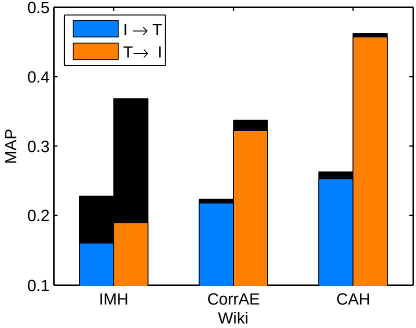
(k) $T \to I$ @ 16 bits



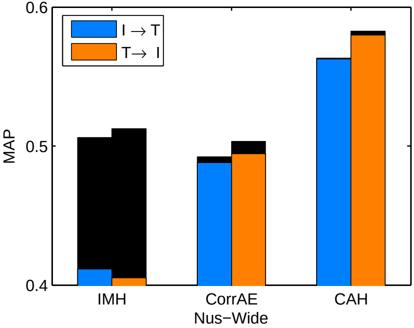
(l) $T \to I$ @ 32 bits

Quantization Error

- Quantization error: search quality loss due to binarization from continuous features to binary codes (black bars).
- CAH incurs significantly less loss on search quality than other two baselines, due to that $\operatorname{sgn}(x) \approx \tanh(x)$ is a more accurate surrogate than the widely-adopted spectral relaxation $\operatorname{sgn}(x) \approx x$.



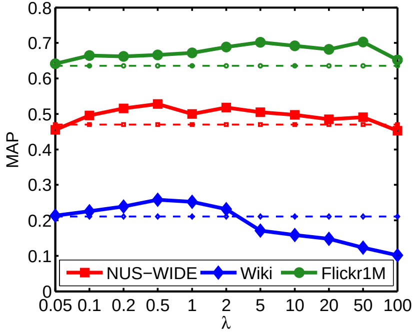
(m) Quantization Error on Wiki



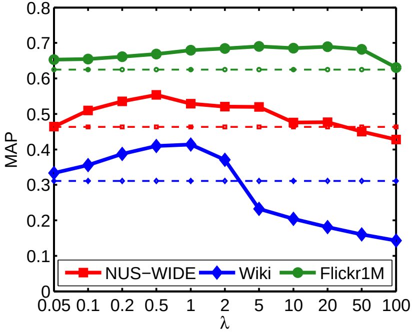
(n) Quantization Error on NUS-WIDE

Parameter Sensitivity

- CAH consistently outperforms the strongest baseline CorrAE on all datasets when λ is varied in a large range $[0.1, 2]$.



(o) $I \to T$ @ 32 bits



(p) $T \to I$ @ 32 bits

Summary

- Correlation Autoencoder Hashing (CAH) for cross-modal search
- Three key points
  - Explore the feature correlations by reconstructing feature vectors of one modality from corresponding hash codes of another modality
  - Explore the semantic correlations by maximizing the inter-category separation margin and minimizing the intra-category variance
  - Enhance both cross-modal correlations in a deep architecture, which will make the embedded hash codes generalize better across different modalities

- Future work
  - Hybrid Deep Architecture: Use Convolutional Neural Net to model images, and use Autoencoder to model texts

Thanks for your listening!

Q & A