

DOMAIN UNCERTAINTY BASED ON INFORMATION THEORY FOR CROSS-MODAL HASH RETRIEVAL

Wei Chen¹, Nan Pu¹, Yu Liu², Erwin M. Bakker¹, Michael S. Lew¹,

¹Leiden University, The Netherlands

²ESAT-PSI, KU Leuven, Belgium

{w.chen, n.pu, e.m.bakker, m.s.k.lew}@liacs.leidenuniv.nl, yu.liu@esat.kuleuven.be

ABSTRACT

Cross-modal hash retrieval has received considerable interest in the area of deep learning. Here hash codes of data of different modalities are learned where pair-wise loss functions control feature similarity in a shared embedding space. In this paper we improve on feature similarity by using Shannon's information entropy with respect to the modality information that is left in learning superior hash codes. We introduce a novel network for predicting the domain from the learned features while the protagonist network uses a loss function based on Shannon's information entropy to learn to maximize the domain uncertainty and therefore the information content. Additionally, according to the number of common labels between each similar image-text pair, we define a multi-level similarity matrix as supervisory information, which constrains all similar pairs with different weights. We show with extensive experiments that our novel approach to domain uncertainty leads to a cross-modal hash retrieval that outperforms the state-of-the-art.

Index Terms— Information entropy, cross-modal hash retrieval, domain uncertainty, multi-level similarity.

1. INTRODUCTION

The use of digital devices has become an integrated part of our daily life. This leads to an exponential increase of video, audio and text data. Finding effective ways of organizing and searching for these huge multimodal datasets has become an important research problem. Cross-modal retrieval has been a compelling research topic in recent years [1][2][3][4][5][6], which aims to accurately index semantically relevant samples from one modality, such as finding a text that describes a given image and vice versa. Meanwhile, to optimize retrieval and storage costs, binary representation learning (*a.k.a* hash code learning) has received increasing attention.

Reducing the modality gap and preserving feature similarity are two key issues being exploited in cross-modal hash retrieval. Since AlexNet [7] was proposed to implement image classification in ImageNet2012, Convolutional Neural Networks (CNNs) have demonstrated powerful feature learning

capacity. Discriminative features for each modality are separately learned well using deep learning methods. However, features from different modalities have usually heterogeneous distributions and representations. For instance, textual features are often more abstract than visual features. A common approach is to map features for different modalities into a joint Hamming space where hash codes can be assessed directly and modality gap is diminished. Existing methods for feature projection, where semantic correlation should be preserved, are categorized into unsupervised [4][5] and supervised [1][6]. Compared to unsupervised methods, supervised hash approaches can achieve superior performance with the help of semantic labels or relevant information.

In recent years, metric learning is used for preserving feature similarity when projecting modality features into a joint space, such as ranking loss [8], or contrastive loss [1][9]. In the joint space, features for similar pairs are projected together, while for dissimilar pairs features will be pushed away. Those loss functions focus on each pair separately and learn their features according to their affinity information, which preserve semantic similarity for each pair. However, using these loss functions cannot guarantee that the feature distributions for image and text are consistent. To tackle this limitation, adversarial learning is incorporated to study the levels of agreement between feature distributions from image and text when classified into their corresponding modality labels [6][8][9]. To obtain a suitable joint space, the gradient needs to be reversed by the optimizing adversarial networks. However, there are still some limitations. First, discrimination for image and text will tend to the semantically-similar image-text pairs far away because they belong to different modalities; Second, modality labels are needed in adversarial learning which limits the generalization to those cases where modalities are not just image and text; Third, the gradient reversal in adversarial learning is not straightforward.

Furthermore, for multi-label datasets, an affinity matrix is used as supervisory information to preserve feature similarity learning. Herein, the differences for all similar pairs are constrained equally [1][9]. Each objective value in the affinity matrix is set to one if an image and text have at least one

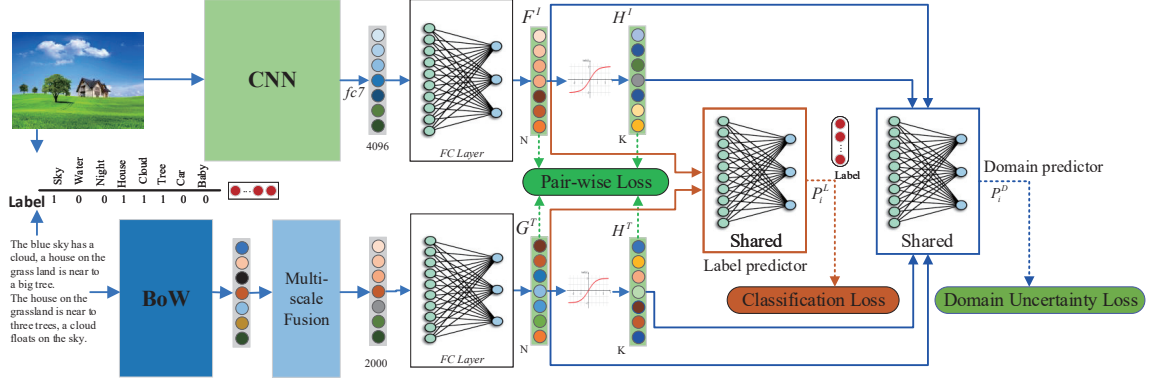


Fig. 1. Our proposed framework of cross-modal hash retrieval. Domain uncertainty loss is based on information theory (Section 3.1); Pair-wise loss is constrained by binary similarity matrix \mathbf{S} and multi-level similarity matrix \mathbf{W} (Section 3.1); Classification loss is introduced in Section 3.3.

common label. However, similar image-text pairs may have different levels of similarity depending on the number of common labels they have.

In this paper, we address above limitations by proposing a novel network for cross-modal hash retrieval as shown in Figure 1. Our contributions are two-fold. First, we propose to incorporate Shannon’s information entropy [10] to directly map features for image and text into a joint space where their heterogeneous modality properties are not exhibited. Specifically, given a hash code which corresponds to image or text, the network, after being trained well, will yield a high uncertainty with respect to modality the hash code belongs to. To the best of our knowledge, this work is the first to use information entropy [10] for cross-modal hash retrieval. Second, we propose a multi-level feature similarity which takes into account the number of common labels between similar image-text pairs to constrain these pairs with different weights.

2. RELATED WORK

Recently, a variety of cross-modal hash learning methods are proposed to bridge the modality gap. Regarding supervised methods to improve retrieval performance, Jiang et al. [1] proposed DCMH to integrate deep feature learning and hash code learning into a unified structure where a similarity matrix was used as supervisory information. Aiming at learning a joint latent space for image and text, Li et al. [9] introduced a three stream self-supervised hashing network where embedded features in a joint space were used to predict semantic labels. For those methods, each similar image-text pair could be well projected as semantically-related feature vectors. However, the holistic feature distributions of two modalities are still inconsistent (*i.e.* showing a heterogeneous modality gap). To mitigate this issue, adversarial learning methods are incorporated [6][9][8]. Chi et al. [6] introduced a dual structure for common representation learning in which new samples are generated via Generative Adversarial Networks (GANs)

[11] and original ones are reconstructed. Their method can solve the problem of adding new categories in cross-modal retrieval; Wang et al. [8] introduced a feature projector and domain classifier which run as minimax game with adversarial learning, but the Gradient Inversal Layer (GRL) [12] and domain labels are needed in their approach.

In this work, we consider a holistic feature distribution in the joint embedding space and incorporate the information entropy [10] to maximize the uncertainty of visual and textual domain, such that modality properties are not exhibited, while preserving the semantic similarity of hash codes by using pair-wise and classification-based loss functions.

3. PROPOSED APPROACH

For the image-text dataset with n samples, we use $\mathbf{X} = \{x_i, l_i\}_{i=1}^n$ to denote the images and their labels, we use $\mathbf{Y} = \{y_i, l_i\}_{i=1}^n$ to denote the text and their labels. Here $l_i = [l_{i1}, l_{i2}, \dots, l_{ic}]$ are multi-label annotations of images and text, and c is the total number of classes. We define a binary similarity matrix \mathbf{S} where $S_{ij} = 1$ when x_i and y_i have at least one common label, otherwise $S_{ij} = 0$. Additionally, we define a multi-level similarity weight $w_{ij} = t_{ij}/c$ where t_{ij} is the number of common labels between x_i and y_i . Given these training data and a supervised matrix, the task of the cross-modal hash retrieval is to learn two *sign* functions for the two modalities: $B(x_i) = \text{sign}(F(x, \theta_v)) \in \{-1, 1\}^K$, $B(y_i) = \text{sign}(G(y, \theta_t)) \in \{-1, 1\}^K$, where K is the length of hash codes, θ_v and θ_t are the network parameters for feature learning for two modalities. According to the binary similarity matrix \mathbf{S} , similar pairs $(F(x_i), G(y_i))$ should be represented by similar hash codes $(B(x_i), B(y_i))$ in the Hamming space. Usually, as $B(\cdot)$ is a discrete function and it is not differentiable, a soft continuous relaxation $H(\cdot) = \tanh(\cdot)$ is used to replace $B(\cdot)$. The hash code can be optimized using:

$$L_q = (\|\mathbf{H}^v - \mathbf{B}^v\|_F^2) + (\|\mathbf{H}^t - \mathbf{B}^t\|_F^2) \quad (1)$$

The aim of our method is to learn a better joint space for real-valued features $\mathbf{F}(\cdot)$, $\mathbf{G}(\cdot)$ and hash codes $\mathbf{H}(\cdot)$, $\mathbf{B}(\cdot)$ where multi-level similarity degrees are also preserved. The whole

framework is depicted in Figure 1.

3.1. Information Theory and Domain Uncertainty

As shown in Figure 2(a), real-valued features extracted from visual and textual domains (F^I and G^T in Figure 1, respectively) are semantically similar but inconsistently distributed. Samples from two domains have different domain-related properties. For example, textual data have more abstract semantics than visual data. Those properties will often result in feature distributions which still hold this information giving higher certainty on the domain to which the input data belongs (*i.e.* the visual domain or textual domain). More specifically, when it is possible to identify a feature in the joint space coming from the visual domain with higher probability (P_i) rather than coming from textual domain with lower probability ($P_t=1-P_i$), domain uncertainty is not achieved. Thus, for a given feature, it can not be determined which domain it originally belongs to, it means that this feature is identified from two domains with equal probability ($P_i = P_t = 0.5$), and the joint space has highest uncertainty corresponding to highest information entropy. As in [10], we incorporate information entropy to measure the uncertainty of two domains. Figure 2(b) illustrates that two domains with equal probability leads to highest information entropy and information content.

Domain uncertainty is in proportional to information entropy [10], as shown in Figure 2(c). Based on this observation, we devise a domain uncertainty loss function using information entropy. When the objective function is minimized, the information entropy will be maximized, which means that the joint space maximizes domain uncertainty. Specifically, we build domain predictor network \mathcal{D} which includes three fully-connected (fc) layers. The output probability is $P_j^d(\cdot) = \mathcal{D}(\cdot, \theta_d)$, “.” indicates features from image or text shared with the parameter θ_d . The output neurons of prediction layer are M :

$$\begin{aligned} \min_{\theta_v, \theta_t, \theta_d} (L_d^r + L_d^b) &= \sum_{i=1}^N \sum_{j=1}^M (P_{d,j}^r(\mathcal{F}(\cdot)) * \log(P_{d,j}^r(\mathcal{F}(\cdot))) \\ &\quad + P_{d,j}^b(\mathcal{H}(\cdot)) * \log(P_{d,j}^b(\mathcal{H}(\cdot)))) \quad (2) \\ \text{s.t. } \mathcal{F}(\cdot) &= \mathbf{F}(x_i, \theta_v) \text{ or } \mathbf{G}(y_i, \theta_t), \\ \mathcal{H}(\cdot) &= \mathbf{H}(x_i, \theta_v) \text{ or } \mathbf{H}(y_i, \theta_t), \end{aligned}$$

where L_d^r is the loss component for the real-valued features used to predict domain probability and L_d^b indicates the loss component for the binary features used for domain prediction. N is the number of training samples, and M is set to 2, which denotes the number of domains in this task.

3.2. Multi-level Feature Preserving

A binary similarity matrix \mathbf{S} can be used to preserve pair-wise similarity. Each $S_{ij} = 1$ when the corresponding image and text have at least one common label. However, similar image-text pairs may have different levels of similarity. Namely, different pairs can have different number of common labels, but the matrix \mathbf{S} constrains these pairs equally. Considering this

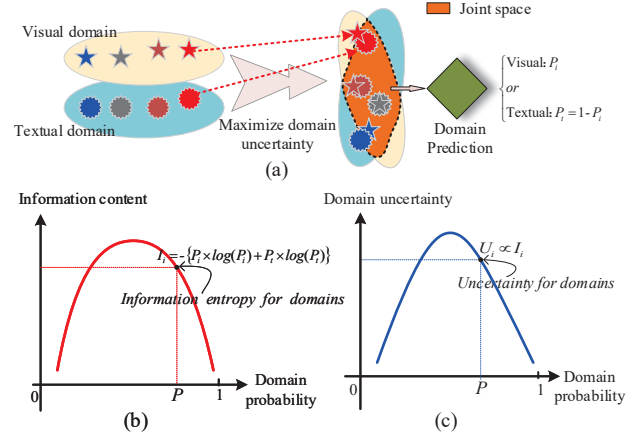


Fig. 2. (a): Images and text are embedded into a joint space. Embedded features of paired samples are close while those of unpaired samples are far away. (b): Relationship between domain uncertainty and predicted probability for each domain. When the probabilities for two domains are equal, the joint space maximizes modality uncertainty. (c): Relationship between domain probability and information entropy. The more uncertain, the more information entropy a joint space will convey. Information entropy is based on information theory.

limitation of \mathbf{S} , we define a multi-level similarity matrix \mathbf{W} , which holds different similarity weights for all similar pairs. We depict the multi-level similarity matrix and binary similarity matrix in Figure 3. Each value w_{ij} in \mathbf{W} is normalized by the total number of class in a dataset.

We denote the real-valued features and binary features of image x_j as a triplet vector $\{F^{x_i}, H^{x_i}, B^{x_i}\}$, and the feature of a text y_j as triplet $\{G^{y_j}, H^{y_j}, B^{y_j}\}$. Then, \mathbf{W} can be used to constrain a more specific feature learning for similar pairs by using the loss function:

$$\begin{aligned} \min_{\theta_v, \theta_t} (L_m^r + L_m^b) &= \sum_{i,j=1}^N \left((\delta(2\Delta_{ij}^r) - w_{ij})^2 + (\delta(2\Gamma_{ij}^b) - w_{ij})^2 \right) \quad (3) \\ \text{s.t. } w_{ij} &= t_{ij}/c, \end{aligned}$$

where L_m^r and L_m^b correspond to real-valued and binary features, $\delta(\cdot)$ is the sigmoid function, w_{ij} is the above defined multi-level similarity weight. $\Delta_{ij}^r = \frac{1}{2}(F_{*i}^T)(G_{*j})$ and $\Gamma_{ij}^b = \frac{1}{2}(H_{*i}^T)(H_{*j})$ denote the inner product of image and text features; H_{*i} and H_{*j} correspond to soft visual and textual hash codes, respectively.

As suggested in [1][9], we also use the binary similarity matrix \mathbf{S} to define the pair-wise objective function. Specifically, for S_{ij} , the conditional probability for each pair (F^{x_i}, G^{y_j}) and (H^{x_i}, H^{y_j}) can be computed by using:

$$p(S_{ij}|B) = \begin{cases} \delta(\psi_{ij}) & S_{ij} = 1, \\ 1 - \delta(\psi_{ij}) & S_{ij} = 0, \end{cases} \quad (4)$$

where $\delta(\psi_{ij})$ is the sigmoid function and ψ_{ij} is the inner product, larger inner product means that samples have a higher probability being similar. Pair-wise objective function is:

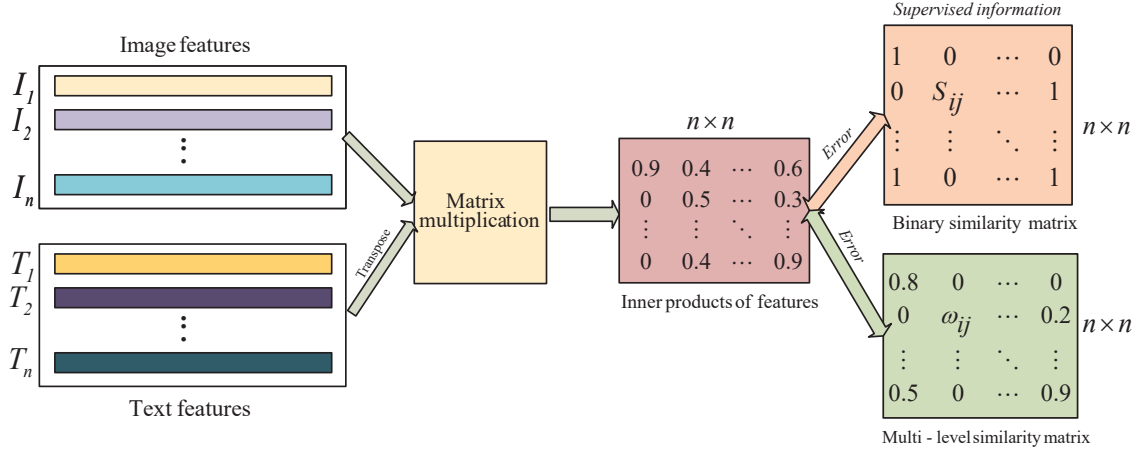


Fig. 3. Multi-level similarity matrix and binary similarity matrix are used as supervised information for feature learning.

$$\min_{\theta_v, \theta_t} (\underbrace{L_{pairs}^r + L_{pairs}^b}_{\theta_v, \theta_t}) = \sum_{i,j=1}^N ((S_{ij} \Delta_{ij}^r - \log(1 + e^{\Delta_{ij}^r})) + (S_{ij} \Gamma_{ij}^b - \log(1 + e^{\Gamma_{ij}^b}))) \quad (5)$$

where Δ_{ij}^r and Γ_{ij}^b are set as in Eq. 3.

3.3. Classification-based objective function

Furthermore, as shown Figure 1, we build a label predictor \mathcal{L} to output the probability $P_i^l(\cdot) = \mathcal{L}(\cdot, \theta_l)$. We only use the length-fixed real-valued feature $F(\cdot)$ and $G(\cdot)$ for label prediction because the length of hash codes is changed. The objective function for the label prediction is defined as:

$$\min_{\theta_v, \theta_t, \theta_l} (\underbrace{L_i^v + L_i^t}_{\theta_v, \theta_t, \theta_l}) = - \sum_{i=1}^N (l_i \cdot \log(p_i^{v,t}) + (1 - l_i) \cdot \log(1 - p_i^{v,t})) \quad (6)$$

where $p_i^v = \mathcal{L}(F(x_i), \theta_l)$, $p_i^t = \mathcal{L}(G(y_i), \theta_l)$ denote the sigmoid output probabilities of label predictor, l_i are the ground-truth labels. The dimension of p_i^v and p_i^t are equal to the number of labels in each dataset.

Finally, the global objective will be:

$$L = \alpha L_d + \beta L_{pairs} + \gamma L_l + \eta L_m + \epsilon L_q \quad (7)$$

3.4. Implementation Details

We utilize the CNN-F from [13] as the backbone network for visual feature learning. As shown in Figure 1, activations from $fc7$ layer are projected into a joint space using a 3 fc layer ($4096 \rightarrow K \rightarrow N$), where K is the dimension of the common feature. We use BoW to embed textual features and then adopt a multi-scale (MS) fusion fc layer ($T \rightarrow MS \rightarrow 2000 \rightarrow K \rightarrow N$) to learn the textual features. Following [9], MS fusion model has five-level pooling layers. The label predictor and domain predictor consist of 3 fc layers in which the number of neurons go from ($512 \rightarrow 256 \rightarrow c$) and ($512 \rightarrow 256 \rightarrow 2$), respectively, where c is 24 for the MIRFlickr-25K and 21 for the NUS-WIDE dataset. For all fc layers, we set

the dropout rate to 0.9. For optimizing the network, we adopt the alternating learning strategy from [1] where we fine-tune visual parameters and fix textual parameters. Regarding to the hyperparameters in Eq.7, we analyze the parameter sensitivity, as reported in Figure 4. Based on these observations, we set $\alpha = 100, \beta = \gamma = \epsilon = 1, \eta = 0.1$. The learning rate varies from 10^{-4} to 10^{-8} .

4. EXPERIMENTS AND EVALUATIONS

4.1. Datasets

The **MIRFLICKR-25K** [14] dataset contains 25,000 instances. We follow the experiment protocols given in [1]. In total, 20,015 image-text pairs are selected for our experiment. The text for each sample is embedded into a 1386 dimensional BoW representation. There are 24 labels for each pair. The number of training pairs is 10,000 and the number of query pairs is 2,000.

The **NUS-WIDE** [15] dataset contains 269,648 images. There are 81 ground-truth concepts that have been annotated manually. Following the protocols in [9], we select the 21 most frequent concepts as the training set (190,421 in total) in which the number of training samples is equal to 10,500 and query set has size of 2,100. Each annotation is embedded into a 1000 dimensional BoW representation.

4.2. Evaluation

We adopt Hamming ranking and hash lookup to evaluate the performance. For hash based retrieval, the Hamming ranking procedure ranks the candidates in the retrieval set according to their Hamming distance to the given query items in ascending order. Mean average precision (mAP) is the commonly-used criteria to measure the accuracy of the Hamming ranking distances. The accuracy of the hash look-up returns all the candidates within a certain Hamming radius. A precision-recall curve is widely used to implement hash look-up evaluation.

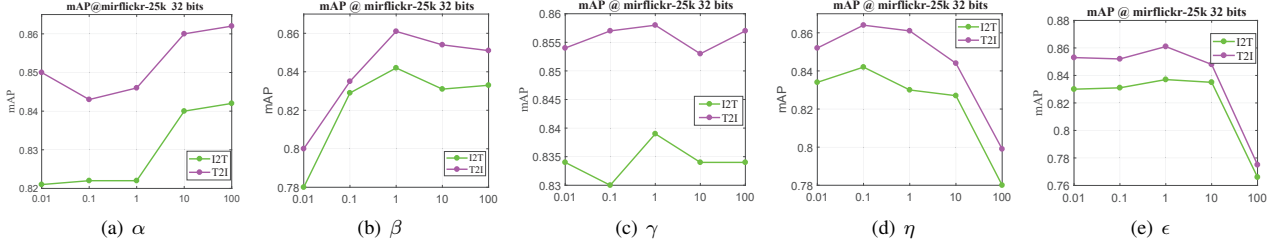


Fig. 4. Sensitivity analysis of the hyperparameters in the global objective function.

For performance comparison, we compare with recent relevant work in DCMH [1] and SSAH [9], both of which use deep learning methods.

4.3. Performance

Common Feature Dimension. The dimension of the common feature is an important parameter for cross-modal hash retrieval. Before conducting our experiments, we evaluate the effect of the common feature dimension (*i.e.* the N in Figure 1). The results are reported in Table 1 where “Image→Text” and “Text→Image” mean that the query items are image and text, respectively. We can see that for $N = 512$, the mAP score is highest. Therefore, in our experiments, we use a 512 dimensional (*i.e.* $N = 512$) common feature.

Table 1. mAP for different N on the MIRFlickr-25k dataset

Task \ N	64	128	256	512	1024
Image→Text	0.802	0.818	0.831	0.833	0.829
Text→Image	0.819	0.850	0.852	0.859	0.844

Hamming Ranking. To demonstrate the precision of our proposed method, we conduct and compare methods using CNN-F features on the MIRFlickr-25k and NUS-WIDE, as shown in Table 2. The baseline results are from SSAH [9] and we find that our method outperforms these baseline methods. Specifically, the proposed method achieves better significantly results than other counterparts. For instance, when the length of the hash codes is equal to 32 bits, the results for “Image→Text” and “Text→Image” are improved by 5.4% and 8.1%, respectively, when compared to state-of-the-art method SSAH. Meanwhile, for another dataset NUS-WIDE, where more instances and contents are included within an image, which makes it hard to train and perform cross-modal retrieval. However, the proposed method also outperforms the other methods. For instance, our method has 2.5% and 2.9% improvement respectively, compared to SSAH using a hash codes of 32 bits. Therefore, all the results in Table 2 demonstrate the effectiveness of using information entropy for mitigating the modality gap. Furthermore, we could find that for different tasks and using a different hash code length, we can find the retrieval performance improves when the hash code length is set to 32 bits.

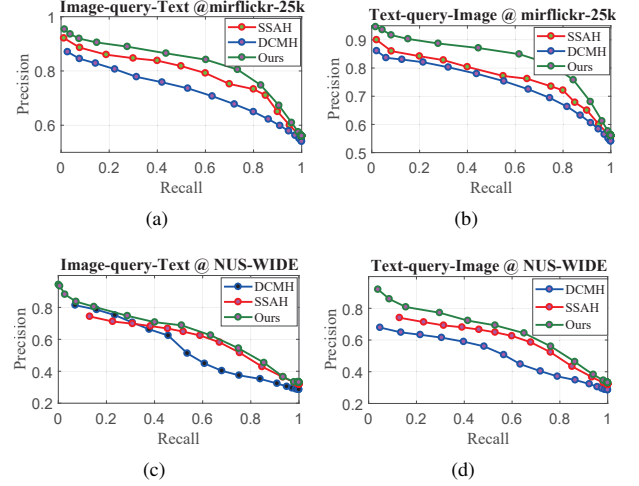


Fig. 5. Precision-recall curves for different approaches. The hash code length is 16 bits.

Hash Lookup. For this procedure, we compute the precision and recall for the retrieval results with respect to a different Hamming radius. In this experiment, we vary the Hamming radius from 0 to 50 with step-size 1. For each radius, the retrieval algorithms will return the correct items, larger covered area of the precision-recall curve indicates a better retrieval performance. The results are shown in Figure 5. For fair comparison, we used the source codes provided by the authors, and a hash code length of 16 bits for this experiment. For both the “Image→Text” and “Text→Image” tasks, our proposed method has curves that have a larger covered areas than these competitive deep learning methods. The result further demonstrates the superiority of the proposed method.

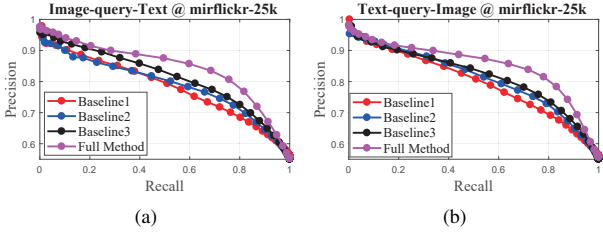
Ablation Study. We conduct an ablation study for our method on the MIRFlickr-25k dataset with 32-bits hash codes. Specifically, we build three baselines using different objective functions. Our *Baseline1* is only based on $L_{pairs} + L_q$; *Baseline2* is based on $L_{pairs} + L_q + L_l$, which illustrates the effectiveness of label predictor; *Baseline3* is based on $L_{pairs} + L_q + L_l + L_m$, demonstrating the effect of a multi-level similarity objective function. Finally, we incorporate all loss functions as *full-method*. The results are reported in Table 3. Furthermore, we compare the corresponding precision-recall curves, as shown in Figure 6. We can see that the mAP is highest when domain uncertainty is used.

Table 2. mAP results on MIRFlickr-25k and NUS-WIDE datasets

Tasks and Methods		MIRFlickr-25K			NUS-WIDE		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
Image→Text	DCMH [1]	0.735	0.737	0.750	0.478	0.486	0.488
	SSAH [9]	0.782	0.790	0.800	0.642	0.636	0.639
	Ours	0.825	0.833	0.838	0.648	0.652	0.647
Text→Image	DCMH [1]	0.763	0.764	0.775	0.638	0.651	0.657
	SSAH [9]	0.791	0.795	0.803	0.669	0.662	0.666
	Ours	0.845	0.859	0.861	0.671	0.681	0.669

Table 3. Ablation study for the proposed method

Tasks	Image→Text	Text→Image
Baselines		
Baseline1	0.773	0.792
Baseline2	0.795	0.814
Baseline3	0.810	0.827
Full-method	0.834	0.859

**Fig. 6.** Precision-recall curves for ablation study. The hash code length is 32 bits.

5. CONCLUSION

In this work, we have exploited modality information for cross-modal hash retrieval. We devised a novel network to predict visual domain and textual domain based on the features learned from these two modalities. The protagonist network depends on a objective function by using Shannon’s information entropy to maximize domain uncertainty. Maximizing the domain uncertainty is beneficial for bridging the gap between two modalities because it minimizes the influence of the individual modality. Furthermore, we considered multi-level similarity for feature learning where all similar image-text pairs are constrained with different weights according to the number of common labels between these similar pairs. Extensive experiments implemented on two multi-label datasets demonstrate the effectiveness of the proposed method which outperforms the state-of-the-art.

6. ACKNOWLEDGMENT

This work is supported by LIACS MediaLab at Leiden University and China Scholarship Council (CSC). We would like to thank NVIDIA for the donation of GPU cards.

7. REFERENCES

[1] Q. Jiang, W. Li, Deep cross-modal hashing, In CoRR .

[2] Y. Shen, L. Liu, L. Shao, J. Song, Deep binaries: Encoding semantic-rich cues for efficient textual-visual cross retrieval, in: In ICCV, IEEE, 4117–4126, 2017.

[3] X. Ji, W. Wang, M. Zhang, Y. Yang, Cross-domain image retrieval with attention modeling, in: In ACM MM, 1654–1662, 2017.

[4] J. Song, Y. Yang, Y. Yang, Z. Huang, H. T. Shen, Inter-media hashing for large-scale retrieval from heterogeneous data sources, in: In SIGMOD, ACM, 785–796, 2013.

[5] G. Wu, Z. Lin, J. Han, L. Liu, G. Ding, B. Zhang, J. Shen, Unsupervised Deep Hashing via Binary Latent Factor Models for Large-scale Cross-modal Retrieval., in: In IJCAI, 2854–2860, 2018.

[6] J. Chi, Y. Peng, Dual Adversarial Networks for Zero-shot Cross-media Retrieval., in: In IJCAI, 663–669, 2018.

[7] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: In NeurIPS, 1097–1105, 2012.

[8] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H. T. Shen, Adversarial cross-modal retrieval, in: In ACM MM, 154–162, 2017.

[9] C. Li, C. Deng, N. Li, W. Liu, X. Gao, D. Tao, Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval, in: In CVPR, 4242–4251, 2018.

[10] C. E. Shannon, A mathematical theory of communication, Bell system technical journal 27 (3) (1948) 379–423.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: In NeurIPS, 2672–2680, 2014.

[12] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, In arXiv:1409.7495 .

[13] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, In arXiv:1405.3531 .

[14] M. J. Huiskes, M. S. Lew, The MIR flickr retrieval evaluation, in: In ACM CIVR, 39–43, 2008.

[15] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, NUS-WIDE: a real-world web image database from National University of Singapore, in: In ACM CIVR, 48, 2009.