Exploring deep learning for intelligent image retrieval

Chen, W.

Chapter 3

Domain Uncertainty based on Information Theory for Cross-modal Hash Retrieval

In the previous chapter, we gave a comprehensive review about intelligent image retrieval. Semantic information that helps us understand the world usually comes from different modalities. We can express the same concept by using different ways so that we can search the images of interest by submitting any media content at hand (*e.g.* a phrase, or an image) as the query item. Therefore, cross-modal hash retrieval, as a natural searching way, 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 chapter, we improve on feature similarity by using Shannon's information entropy with respect to the modality information that is present 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.

Keywords

Information entropy, cross-modal hash retrieval, domain uncertainty, multi-level similarity

This chapter is based on the following publication [34]:

- Chen, W., Pu, N., Liu, Y., Bakker, E. and Lew, M.S., "Domain Uncertainty Based On Information Theory for Cross-Modal Hash Retrieval." IEEE International Conference on Multimedia and Expo (ICME), 2019, pp 43-48.

3.1 Introduction

Cross-modal retrieval has been a compelling research topic in recent years [173, 174, 175]. It 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 heterogeneity gap [176] and the semantic gap [10] (*i.e.* retaining feature similarity) are two key issues being explored in cross-modal hash retrieval. Since the data in different modalities are described by different statistical properties, the heterogeneity gap characterizes the difference between feature vectors from different modalities that have similar semantics but are distributed in different spaces. Similarities between these feature vectors are not well associated so that these vectors are not directly comparable, leading to inconsistent distributions. The semantic gap characterizes the difference, in any application, between the high-level concepts of humans and the low-level features typically derived from images (*i.e.* pixels or symbols) [10].

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. Textual features are often more abstract than visual features. A common practice is to map features for different modalities into a common Hamming space where hash codes can be assessed directly and the heterogeneity gap is diminished. Existing methods for feature projection are categorized into unsupervised [174] and supervised [173, 175]. 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 to retain feature similarity when projecting modality features into a common space, such as ranking loss [177], and contrastive loss [173, 178]. In the common space, features of similar pairs are projected together, while for dissimilar pairs features will be pushed away. These loss functions focus on each pair separately and learn their features according to their affinity information. 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 [175, 177, 178]. To obtain a suitable common space, the gradients need to be reversed by the optimizing adversarial networks. However, there still exist 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 these cases where modalities are not just image and text; Third, the gradient reversal in adversarial learning is not straightforward.

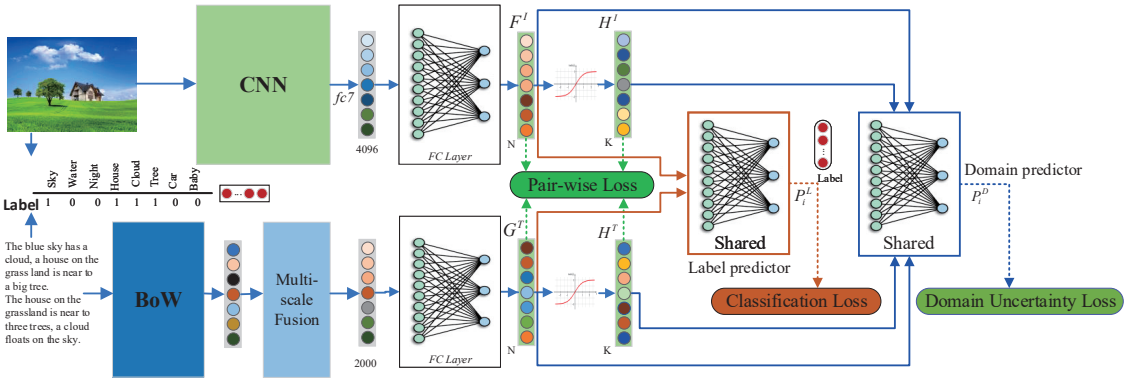

Figure 3.1: The framework for cross-modal hash retrieval. Domain uncertainty loss is based on information theory (Section 3.3.1); Pair-wise loss is constrained by binary similarity matrix **S** and multi-level similarity matrix **W** (Section 3.3.1); Classification loss is introduced in Section 3.3.3.

For multi-label datasets, an affinity matrix is used as binary supervisory information to constrain feature similarity. Herein, all similar pairs are constrained equally [173, 178]. Each objective value in the affinity matrix is set to 1 if an image and text have at least one common label. However, similar image-text pairs may have different levels of similarity depending on the number of common labels they have.

In this chapter, we address above limitations by proposing a novel network, as shown in Figure 3.1. The novelty of this chapter is summarized as two-fold. First, we incorporate Shannon's information entropy [179] to directly map features for image and text into a common 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 [179] for cross-modal hash retrieval. Second, we propose a multi-level feature similarity which considers the number of common labels between similar image-text pairs to constrain these pairs with different weights.

3.2 Cross-modal Hash Learning

Recently, a variety of cross-modal hash learning methods are proposed to minimize the heterogeneity gap. Regarding supervised methods to improve retrieval performance, Jiang *et al.* [173] 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 common latent space for image and text, Li *et al.* [178] introduced a three stream self-supervised hashing network where embedded features in a common space were used to predict semantic labels. For these 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 heterogeneity gap). To mitigate this issue, adversarial learning methods are incorporated [175, 177, 178]. Chi *et al.* [175] introduced a dual structure for common representation learning in which new samples are generated via Generative Adversarial Networks (GANs) [180] and original ones are reconstructed. Their method can solve the problem of adding new categories in cross-modal retrieval; Wang *et al.* [177] introduced a feature projector and domain classifier which run as minimax game with adversarial learning, but the Gradient Reversal Layer (GRL) [181] and domain labels are needed in their approach.

We consider a holistic feature distribution in the common space and incorporate the information entropy [179] to maximize the uncertainty of visual and textual domains, such that modality properties are not exhibited, while preserving the semantic similarity of hash codes by using pair-wise and classification-based loss functions.

3.3 Domain Uncertainty Measurement via Information Theory

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 $l_i = [l_{i1}, l_{i2}, ..., l_{ic}]$ 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: $B(x_i) = sign(F(x_i, \boldsymbol{\theta}_v)) \in \{-1, +1\}^K$, $B(y_i) = sign(G(y_i, \boldsymbol{\theta}_t)) \in \{-1, +1\}^K$, where K is the length of hash codes, $\boldsymbol{\theta}_v$ and $\boldsymbol{\theta}_t$ 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 differentiate, a soft continuous relaxation $H(\cdot) = tanh(\cdot)$ is used to replace $B(\cdot)$.The hash code can be optimized using:

$$L_q = \left(\|\mathbf{H}^v - \mathbf{B}^v\|_F^2\right) + \left(\|\mathbf{H}^t - \mathbf{B}^t\|_F^2\right) \tag{3.1}$$

The aim of our method is to learn a better common 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 3.1.

3.3.1 Information theory and domain uncertainty

As shown in Figure 3.2(a), real-valued features extracted from visual and textual domains (F^I and G^T in Figure 3.1, respectively) are semantically similar but in-

consistently distributed. Samples from two domains have different domain-related properties. For example, textual data have more abstract semantics than visual data. These 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 common 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 common space has highest uncertainty corresponding to highest information entropy. As in [179], we incorporate information entropy to measure the uncertainty of two domains. Figure 3.2(b) illustrates that two domains with equal probability leads to highest information entropy and information content.

Domain uncertainty is in proportional to information entropy [179], as shown in Figure 3.2(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 common space maximizes domain uncertainty. Specifically, we build domain predictor network $\mathcal{D}$ which includes three fully-connected (FC) layers. The output probability is $P_j^d(\cdot) = \mathcal{D}(\cdot, \boldsymbol{\theta}_d)$, "$\cdot$" indicates features from image or text shared with the parameter $\boldsymbol{\theta}_d$. The output neurons of prediction layer are M.:

$$\begin{aligned} \min \underbrace{(L_d^r + L_d^b)}_{\boldsymbol{\theta}_v, \boldsymbol{\theta}_t, \boldsymbol{\theta}_d} = & \sum_{i=1}^{N} \sum_{j=1}^{M} \Big(P_{d,j}^r\big(\mathcal{F}(\cdot)\big) * log\big(P_{d,j}^r(\mathcal{F}(\cdot))\big) \\ & + P_{d,j}^b\big(\mathcal{H}(\cdot)\big) * log\big(P_{d,j}^b(\mathcal{H}(\cdot))\big) \Big) \\ s.t. \quad & \mathcal{F}(\cdot) = \mathbf{F}(x, \boldsymbol{\theta}_v) \; or \; \mathbf{G}(y, \boldsymbol{\theta}_t), \\ & \mathcal{H}(\cdot) = \mathbf{H}(x, \boldsymbol{\theta}_v) \; or \; \mathbf{H}(y, \boldsymbol{\theta}_t), \end{aligned} \tag{3.2}$$

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.

3.3.2 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 limitation of $\mathbf{S}$, we define a multi-level similarity matrix $\mathbf{W}$, which holds different similarity weights for all similar pairs. We depict

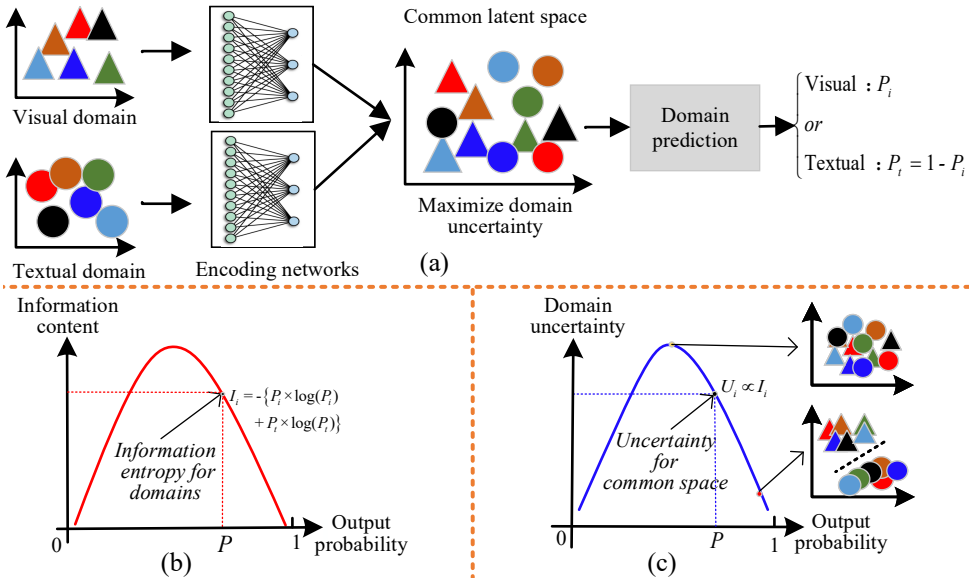

Figure 3.2: (a): Images and text are embedded via non-shared encoding sub-networks. The domain uncertainty can be predicted by using the output probabilities from a predictor. (b): Relationship between information entropy and predicted probability. (c): Relationship between domain uncertainty and output probabilities. When probabilities predicted for two modalities are identical, the shared space is intertwined into a domain confusion state (*i.e.* most uncertain). If one modality is identified with a higher probability (closer to 1) while another with a lower probability (closer to 0), the domain confusion state is not achieved.

the multi-level similarity matrix and binary similarity matrix in Figure 3.3. Each value w_{ij} in $\mathbf{W}$ is normalized by the total number of class in a dataset.

The real-valued features and binary features of image x_j are denoted 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 regularize a more specific similar pairs by using:

$$\begin{aligned} \min \underbrace{(L_m^r + L_m^b)}_{\boldsymbol{\theta}_v, \boldsymbol{\theta}_t} = \sum_{i,j=1}^{N} \Big(\big(\delta(2\Delta_{ij}^r) - w_{ij}\big)^2 + \big(\delta(2\Gamma_{ij}^b) - w_{ij}\big)^2 \Big) \\ s.t. \quad w_{ij} = t_{ij}/c, \end{aligned} \tag{3.3}$$

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 soften visual and textual hash codes, respectively.

As suggested in [173, 178], 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:

$$p(S_{ij}|B) = \begin{cases} \delta(\psi_{ij}) & S_{ij} = 1, \\ 1 - \delta(\psi_{ij}) & S_{ij} = 0, \end{cases} \tag{3.4}$$

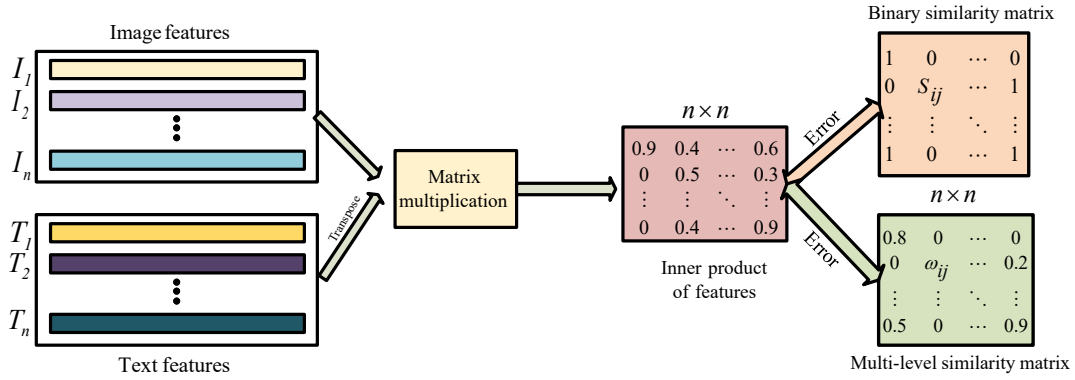

Figure 3.3: Multi-level similarity matrix and binary similarity matrix are used as supervised information for feature learning.

where $\delta(\psi_{ij})$ is the sigmoid function and ψ_{ij} is the inner product of input features. Pair-wise objective function is:

$$\min \underbrace{(L^r_{pairs} + L^b_{pairs})}_{\boldsymbol{\theta}_v, \boldsymbol{\theta}_t} = \sum_{i,j=1}^{N} \Big(\big(S_{ij}\Delta^r_{ij} - log(1 + e^{\Delta^r_{ij}})\big) + \big(S_{ij}\Gamma^b_{ij} - log(1 + e^{\Gamma^b_{ij}})\big) \Big) \tag{3.5}$$

where Δ^r_{ij} and Γ^b_{ij} are set as in Eq. 3.3.

3.3.3 Classification-based objective function

Furthermore, as shown Figure 3.1, we build a label predictor $\mathcal{L}$ to output the probability $P^l_i(\cdot) = \mathcal{L}(\cdot, \boldsymbol{\theta}_l)$. 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:

$$\min \underbrace{(L^v_l + L^t_l)}_{\boldsymbol{\theta}_v, \boldsymbol{\theta}_t, \boldsymbol{\theta}_l} = -\sum_{i=1}^{N} \Big(l_i \cdot log(p^{v,t}_i) + (1 - l_i) \cdot log(1 - p^{v,t}_i) \Big) \tag{3.6}$$

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

Finally, the global objective will be:

$$L = \alpha\ L_d + \beta\ L_{pairs} + \gamma\ L_l + \eta\ L_m + \epsilon\ L_q \tag{3.7}$$

3.4 Experiments and Evaluations

3.4.1 Implementation details

We utilize the CNN-F from [75] as the backbone network for visual feature learning. As shown in Figure 3.1, activations from *FC7* layer are projected into a common 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 ($\mathrm{T} \rightarrow \mathrm{MS} \rightarrow 2000 \rightarrow K \rightarrow N$) to learn the textual features. Following [178], 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 [173] 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}.

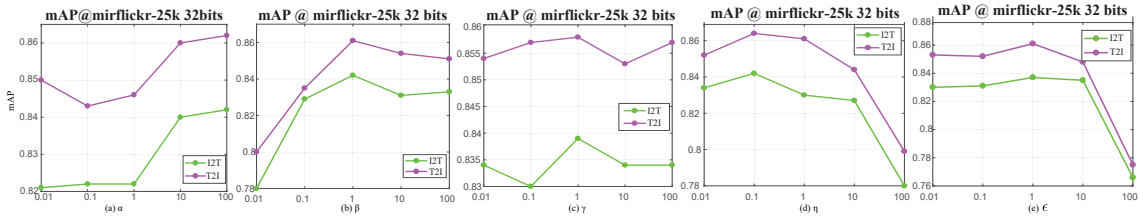

Figure 3.4: Sensitivity analysis of the hyperparameters in loss function in Eq. 3.7.

3.4.2 Datasets

The **MIRFLICKR-25K** [182] dataset contains 25,000 instances. We follow the experiment protocols given in [173]. 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** [183] dataset contains 269,648 images. There are 81 ground-truth concepts that have been annotated manually. Following the protocols in [178], 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.

3.4.3 Performance and 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

Table 3.1: mAP for different feature dimension N on the MIRFlickr-25k dataset.

Feature dimension $N =$	64	128	256	512	1024
Image-to-Text	0.802	0.818	0.831	**0.833**	0.829
Text-to-Image	0.819	0.850	0.852	**0.859**	0.844

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. For performance comparison, we compare with recent relevant work in DCMH [173] and SSAH [178], both of which use deep learning methods.

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 3.1). The results are reported in Table 3.1 where "Image-to-Text" and "Text-to-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.

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 3.2. The baseline results are from SSAH [178] 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-to-Text" and "Text-to-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%

Table 3.2: 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-to-Text	DCMH [173]	0.735	0.737	0.750	0.478	0.486	0.488
	SSAH [178]	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-to-Image	DCMH [173]	0.763	0.764	0.775	0.638	0.651	0.657
	SSAH [178]	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**

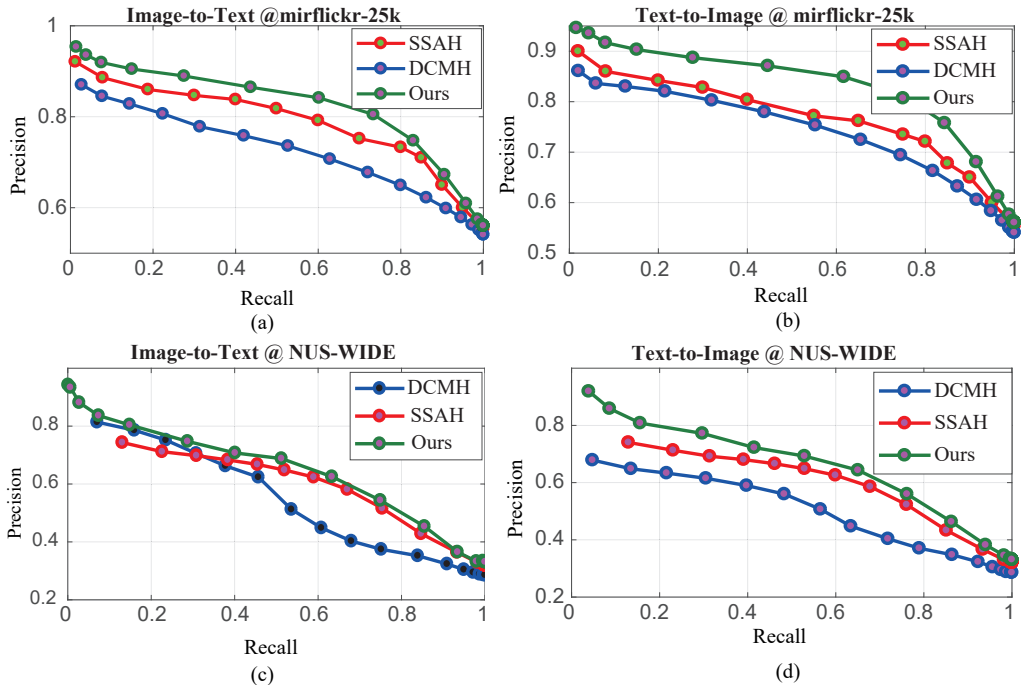

Figure 3.5: Precision-recall curves for three methods. The code length is 16 bits.

Table 3.3: Ablation study for the proposed method.

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

and 2.9% improvement respectively, compared to SSAH using a hash codes of 32 bits. Therefore, all the results in Table 3.2 demonstrate the effectiveness of using information entropy for mitigating the heterogeneity 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.

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 3.5. 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-to-Text" and "Text-to-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-

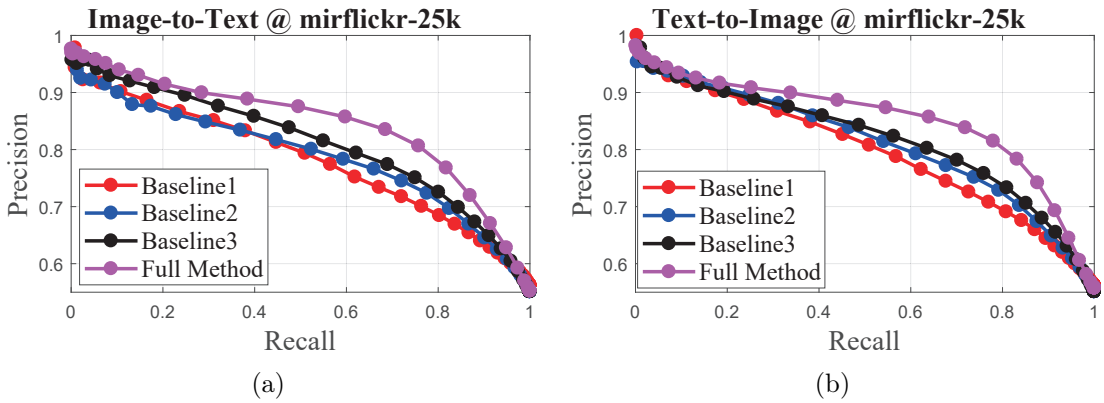

Figure 3.6: Precision-recall curves for ablation study. The code length is 32 bits.

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 3.3. Furthermore, we compare the corresponding precision-recall curves, as shown in Figure 3.6. We can see that the mAP is highest when domain uncertainty is used.

3.5 Chapter Conclusions

In this chapter, 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.