Exploring deep learning for intelligent image retrieval

Chen, W.

Chapter 4

Integrating Information Theory and Adversarial Learning for Cross-modal Retrieval

In this chapter, we further explore cross-modal retrieval to address the challenges posited by the heterogeneity gap and the semantic gap. To be specific, we propose integrating Shannon information theory and adversarial learning. In terms of the heterogeneity gap, we integrate modality classification and information entropy maximization adversarially. For this purpose, a modality classifier (as a discriminator) is built to distinguish the text and image modalities according to their different statistical properties. This discriminator uses its output probabilities to compute Shannon information entropy, which measures the uncertainty of the modality classification it performs. Moreover, feature encoders (as a generator) project uni-modal features into a commonly shared space and attempt to fool the discriminator by maximizing its output information entropy. Thus, maximizing information entropy gradually reduces the distribution discrepancy of cross-modal features, thereby achieving a domain confusion state where the discriminator cannot classify two modalities confidently. To reduce the semantic gap, Kullback-Leibler (KL) divergence and bidirectional triplet loss are used to associate the intra- and inter-modality similarity between features in the shared space. Furthermore, a regularization term based on KL-divergence with temperature scaling is used to calibrate the biased label classifier caused by the data imbalance issue.

Keywords

Cross-modal retrieval, Shannon information theory, Adversarial learning, Modality uncertainty, Data imbalance.

This chapter is based on the following publication [35]:

- Chen, W., Liu, Y., Bakker, E., and Lew, M.S., "Integrating Information Theory and Adversarial Learning for Cross-modal Retrieval." Pattern Recognition, 2021, 117, pp. 107983.

4.1 Introduction

Deep learning methods can effectively embed features from different modalities into a commonly shared space, and then measure the similarity between these embedded features. As mentioned in Chapter 3, the “heterogeneity gap” [176] and the “semantic gap” [10] are still challenges to be addressed for cross-modal retrieval. To achieve better retrieval performance, it is essential to address these gaps for associating the similarity between cross-modal features in the shared space.

To capture the semantic similarity between cross-modal features, many approaches have been proposed in recent years. Some approaches focus on designing effective structures from a deep networks perspective. For instance, graph convolutional networks are employed to model the dependencies within visual or textual data. Other approaches focus on designing similarity constraint functions from a deep features perspective. For example, bilinear pooling-based methods are applied to align image and text features to then accurately capture inter-modality semantic similarity. In other examples, coordinated representation learning methods, such as ranking loss [177, 184] are widely used to preserve similarity between cross-modal features. These constraint functions mainly aim at reducing the semantic gap by focusing on the similarity between two-tuple or three-tuple samples. However, they might not directly mitigate the heterogeneity gap caused by the inconsistent feature distributions in the different spaces.

Considering the limitations of similarity constraint functions, we propose a new method to perform cross-modal retrieval from two aspects. First, we reduce the heterogeneity gap by integrating Shannon information theory [179] with adversarial learning, in order to construct a better embedding space for cross-modal representation learning. Second, we combine two loss functions, including KL-divergence loss and bi-directional triplet loss, to preserve semantic similarity during the feature embedding procedure, thereby reducing the semantic gap.

To do this, we combine the information entropy predictor and the modality classifier in an adversarial manner. Information entropy maximization and modality classification are two processes trained with competitive goals. Since uni-modal features extracted from image or text data are characterized by different statistical properties, it can be used to distinguish the original modalities these features belong to. As a result, when these features in the shared space are correctly classified into their original modalities with high confidence, then their feature distributions convey less information content, and the modality classifier performs modality classification with lower uncertainty. In contrast, when cross-modal features become modality-invariant and show their commonalities, these features cannot be classified into the modality they originally belong to. In this case, the feature distributions in the shared space conveys more information content and higher modality uncertainty.

According to Shannon's information theory [179], we can measure the modality uncertainty in the shared space by computing information entropy. This basic proportional relation provides the principle to mitigate the heterogeneity gap. For this purpose, we integrate modality uncertainty measurement into cross-modal representation learning. As shown in Figure 4.1, a modality classifier (in the following we call it a *discriminator*) is devised to classify image and text modality, rather than perform a "true/false" binary classification. This discriminator also provides its output probabilities to calculate the information entropy of the cross-modal feature distributions. At the start of training, the discriminator can classify images and text modalities with high confidence due to their different statistical properties. In contrast, the feature encoders (in the following we call it a *generator*) project features into a shared space and attempt to fool the discriminator and make it perform an incorrect modality classification until features in the shared space are fused heavily into a confusion state, maximizing the modality uncertainty.

On the basis of this heavily-fused state, we further use similarity constraints on the feature projector to reduce the semantic gap. Specifically, KL-divergence loss is used to preserve semantic similarity between image and text features by using instance labels as supervisory information. More importantly, we consider the issue of data imbalance and introduce a regularization based on KL-divergence with temperature scaling to calibrate the biased label classifier. Afterwards, we adopt the commonly used bi-directional triplet loss and instance label classification loss (*i.e.* categorical cross-entropy loss) to achieve good retrieval performance.

4.2 Related Work

4.2.1 Cross-modal representation learning and matching

Preserving the similarity between cross-modal features should consider two aspects: inter-modality and intra-modality. Supervision information (*e.g.* class label or instance label), if available, is beneficial for learning features from these two aspects. Preserving feature similarity can be realized by using methods such as joint representation learning and coordinated representation learning. Joint representation learning methods project the uni-modal features into the shared space using straightforward strategies such as feature concatenation, summation, and inner product. Subsequently, more complicated bilinear pooling methods, such as multimodal compact bilinear (MCB) pooling, are proposed to explore the semantic similarity of cross-modal features. To regularize the joint representations, deep networks are commonly trained by using objective functions, such as regression-based loss [185].

Coordinated representation learning methods process image and text features separately but impose them under certain similarity constraints. In general, these constraints can be categorized into classification-based and verification-based methods in supervised scenarios. In terms of classification-based methods, both image and text features are used to make a label classification by using categorical cross-entropy loss function. Because a paired image-text input has the same class label, their features can be associated in the shared space. However, classification-based methods cannot preserve the similarity between inter-modality features well because the similarity between image and text features is not directly regularized.

Verification-based methods, based on metric learning, are proposed to further optimize inter-modality feature learning. Given a similar (or dissimilar) image-text pair, their corresponding features should be verified as similar (or dissimilar). Therefore, the goal of deep networks is to push features of similar pairs closer, while keeping features of dissimilar pairs further apart. Verification-based methods include pair-wise constraints and triplet constraints, which focus on inferring the matching scores of image-text feature pairs [185].

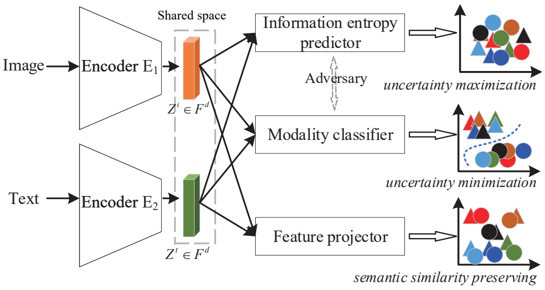


Figure 4.1: Illustration of combining information theory and adversarial learning. The features $Z^i \in F^d$ and $Z^t \in F^d$ with dimension d for image-text pairs are extracted using deep neural networks. Shape indicates modality and color denotes pair-wise similarity information.

Triplet constraints optimize the distance between positive pairs to be smaller than the distance between negative pairs by a margin. They can capture both intra-modality and inter-modality semantic similarity. For example, bi-directional triplet loss has been employed to optimize image-to-text and text-to-image ranking [177]. Although triplet constraints are widely used for cross-modal retrieval, the difficulties are in the mining strategy for negative pairs and the selection of a margin value, which are usually task-specific and empirically selective.

4.2.2 Adversarial learning for cross-modal retrieval

The aforementioned joint and coordinated representation learning approaches focus on two-tuple or three-tuple samples, which may be insufficient for achieving overall good retrieval performance. Adversarial learning, as an alternative method, has shown its powerful capability for modeling feature distributions and learning discriminative representations between modalities when deep networks are trained with competitive objective functions [177].

Recent progress in using adversarial learning for cross-modal retrieval can be categorized as feature-level and loss function-level discriminative models.

From a feature-level perspective, it is possible to preserve semantic consistency by performing a min-max game between inter-modality feature pairs [177]. A straightforward way is to build a discriminator, making a "true/false" classification between image features (regarded as true), corresponding matched text features (regarded as fake), and unmatched image features from other categories (also regarded as fake) [177]. Alternatively, a cross-modal auto-encoder can be combined to generate features for another modality. For example, a generator attempts to generate image features from textual data and then regards them as true, while for a discriminator, image features extracted from original images and these from the generated "images" are labeled as true and fake, respectively. The adversarial training explores the semantic similarity of cross-modal representations. Intra-modality discrimination also can be considered in cross-modal adversarial learning, forcing the generator to learn more discriminative features. In this case, the discriminator tends to discriminate the generated features from its original input.

From a loss function-level perspective, instead of making a binary classification (*i.e.* true or fake), adversarial learning is used to train two groups of loss functions or two processes with competitive goals. This idea is applied in recent work for cross-modal retrieval [177]. Specifically, a feature projector is trained to generate modality-invariant representations in the shared space, while a modality classifier is constructed to classify the generated representations into two modalities. Similarly, we combine two networks and train them with two competitive goals.

4.2.3 Information-theoretical feature learning

As noted before, feature vectors from different modalities are distributed in different spaces, resulting in the heterogeneity gap, which affects the accuracy of cross-modal retrieval. Therefore, it becomes essential to reduce feature distribution discrepancies and thereby reduce the heterogeneity gap. The solution for this is to measure and then minimize distribution discrepancy. For example, distribution disparity of cross-modal features can be characterized by Maximum Mean Discrepancy (MMD), which is a differentiable distance metric between distributions. However, MMD suffers from sensitive kernel bandwidth and weak gradients during training.

Information-theoretical based methods measure the differences of feature distributions and learn better cross-modal features. As an example, the cross-entropy loss function is widely used to estimate the errors between inference probabilities and ground-truth labels where the gradients are computed according to the errors. Once the gradients are computed, deep networks can further update their parameters via the back-propagation algorithm. KL-divergence (also called relative entropy) is another popular criterion to characterize the difference between two probability distributions. Minimizing the difference is beneficial for retaining the semantic similarity between features. For example, Zhang *et al.* [186] employ the KL-divergence to measure the similarity between projected features and supervisory information.

Recently, Shannon information entropy [179] has been used for performing cross-modal hash retrieval [34]. This study indicates that Shannon entropy can be used for multimodal representation learning by estimating uncertainty [179]. Take generative adversarial networks as an example: if the generator makes image features and text features close and minimizes their discrepancy, then the discriminator will become less-certain or under-confident, *i.e.*, having a high information entropy to predict which modality each feature comes from. We applied this principle in our previous work [34] to design an objective function to maximize the domain uncertainty over cross-modal hash codes in a commonly shared space. Deep networks trained by using information entropy construct a domain confusion state where the heterogeneity gap can be effectively reduced. On the basis of this state, other loss functions, such as ranking loss, can be further applied to regularize feature similarity.

4.3 Proposed Approach

4.3.1 Problem formulation

We consider a supervised scenario for cross-modal retrieval. Denote X^i as the input images and the corresponding descriptive sentences as X^t. Each image and its descriptive sentences have the same instance label Y. Therefore, we can organize an input pair (x^i, x^t, y) to train a deep network. To be specific, feature encoders $E_1(\cdot;\boldsymbol{\theta}_{E_1})$ and $E_2(\cdot;\boldsymbol{\theta}_{E_2})$ extract image and text features, respectively, and then further embed these uni-modal features into a shared space by using non-shared sub-networks. The embedded features with dimension d are denoted as $Z^i = E_1(X^i;\boldsymbol{\theta}_{E_1})$ and $Z^t = E_2(X^t;\boldsymbol{\theta}_{E_2})$, Z^i,$Z^t \in R^d$. Note that the parameters in the non-shared sub-networks for uni-modal image and text feature embedding have been included into $\boldsymbol{\theta}_{E_1}$ and $\boldsymbol{\theta}_{E_2}$, respectively. The goal is to train a deep network to make the embedded features Z^i and Z^t modality-invariant and semantically discriminative, improving the retrieval accuracy.

As shown in Figure 4.1, the networks E_1, E_2, and the information entropy predictor act as a generator, while the modality classifier acts as a discriminator. The training of the generator and the discriminator is formulated as an min-max game to mitigate the heterogeneity gap. The feature projector preserves feature similarity under several constraints, which are introduced in Section 4.4.2, 4.4.3, and 4.4.4.

4.3.2 Integrating information theory & adversarial learning

4.3.2.1 Information entropy and modality uncertainty

Uni-modal features from different modalities have similar semantics but are distributed in different spaces. Their similarities are not well associated so that these features are not directly comparable. It is required to further embed them into a shared space (*i.e.* Z^i and Z^t in Figure 4.1). Uni-modal features are characterized by different statistical properties. Therefore, as shown in Figure 3.2(a) in Chapter 3, it is possible to identify a feature in the shared space coming from a visual modality with higher probability P_i (more certain classification) than coming from a textual modality with lower probability P_t=1–P_i (less certain classification). In other words, these cross-modal features are not intertwined heavily. As a result, the domain confusion state is not achieved. Conversely, if a given feature can not be distinguished which modality this feature originally comes from, it indicates that this feature has identical probability ($P_i = P_t$) coming from each modality. In this case, the shared space has highest uncertainty and the cross-modal features are intertwined into a domain confusion state, which corresponds to highest information content. We use information entropy [179] to measure the uncertainty of the shared space. Figure 3.2(b) in Chapter 3 illustrates that two modalities with an equal probability leads to the highest Shannon information entropy and thus information content.

Modality uncertainty refers to the unreliability of classification that the discriminator classifies image features and text features into two modalities. It is proportional to Shannon information entropy [179], as shown in Figure 3.2(c) in Chapter 3. Based on this observation [34], we design the discriminator to measure its output modality uncertainty by using information entropy as a criterion. Maximizing information entropy means that the discriminator becomes least-confident in classifying the original modality of image and text features, resulting in the greatest reduction of the heterogeneity gap.

4.3.2.2 Adversarial learning and information entropy

To make cross-modal features modality-invariant, we devise a generator and a discriminator, as shown in Figure 4.1. The discriminator performs modality classification to identify visual modality and textual modality based on cross-modal features. Following [177], we define the modality label as Y_c^* for these two modalities (for visual modality $* = i$ and textual modality $* = t$). Using output probabilities of the discriminator, we can compute cross-entropy loss to realize modality classification [177]. Once the network convergences under the constraint of this loss function, visual modality and textual modality are clearly identified and classified, thereby minimizing the modality uncertainty.

Conversely, the generator is designed to maximize the modality uncertainty over the cross-modal feature distributions. To achieve this, the generator learns modality-invariant features to fool the discriminator, maximizing the uncertainty of modality classification the discriminator performs. If the modality uncertainty is maximized, the discriminator is most likely to make an incorrect modality classification and be least-confident about its classification results. In this case, cross-modal features are intertwined into a domain confusion state and become indistinguishable.

To this end, we explore the ways to integrate information entropy and adversarial learning into an end-to-end network, which is introduced in Section 4.4.1. For better understanding, we also explore another combining paradigm in the Experimental Section.

4.3.3 KL-divergence for cross-modal feature projection

To reduce the semantic gap, we use KL-divergence to characterize the differences between projected cross-modal features (Z^i and Z^t in Figure 4.1) and a supervisory matrix computed from their instance labels, *i.e.* $KL((f(Z^i, Z^t)||f(Y_l^\top, Y_l))$, (see Eq. 4.9). In this way, the semantic similarity among cross-modal features can be preserved. We illustrate this process in Figure 4.2. It is important to note that when using KL-divergence to preserve semantic similarity of cross-modal features, all positive and negative pairs in a mini-batch are considered. As for the supervisory matrix $f(Y_l^\top, Y_l)$, it is computed by using matrix multiplication and is normalized to the range from 0 and 1.

We argue that different operations to realize $f(Z^i, Z^t)$ affect similarity preserving. Directly, the operation $f(\cdot)$ can be an inner product on cross-modal features Z^i and Z^t. However, using the inner product has some implicit drawbacks. First, when multiplying one image feature vector with all text feature vectors, the results of the inner product are not optimally comparable due to the non-normalized text features, and vice versa. Second, the angles between each image feature vector and each text feature vector, as well as their whole feature distributions, are changing when training the deep network, which makes it problematic for an inner product to measure feature similarity.

To tackle the above limitations, we adopt a cross-modal feature projection to characterize the similarity between features. The idea is related to the work in [186]. Cross-modal feature projection is based on the same distribution and operates on the normalized features. For instance, an image feature vector, $z_j^i \in Z^i$, can be projected to the distribution of a text feature vector $z_k^t \in Z^t$, then each projected feature vector from image to text (termed "$i \to t$") can be formulated as:

$$\begin{aligned} \hat{z}_j^{i\to t} &= |z_j^i| * \frac{<z_j^i, z_k^t>}{|z_j^i||z_k^t|} * \frac{z_k^t}{|z_k^t|} \\ &= <z_j^i, \bar{z}_k^t> * \bar{z}_k^t \end{aligned} \tag{4.1}$$

where "i" and "t" represent the visual and the textual modality, respectively, "j" and "k" represent the index of each image feature and text feature in the shared space, respectively, $\bar{z}_k^t$ denotes the normalized feature. Therefore, the length of $\hat{z}_j^{i\to t}$ is equal to $|\hat{z}_j^{i\to t}| = |<z_j^i, \bar{z}_k^t>|$, and denotes the similarity between image feature z_j^i and text feature z_k^t. When associating each image feature z_j^i with all text features Z^t, we obtain all different lengths, Therefore, when projecting all image features into all text features Z^t, we get a similarity matrix $A_{i\to t}$, which is formulated as

$$A_{i\to t}(Z^i, Z^t) = \sum_{j=1}^{N}\sum_{k=1}^{N} |<z_j^i, \bar{z}_k^t>| = Z^i(\bar{Z}^t)^\top \tag{4.2}$$

Similarly, if projecting all text features into all image features Z^i, we obtain another similarity matrix $A_{t\rightarrow i}$:

$$A_{t\rightarrow i}(Z^t, Z^i) = \sum_{k=1}^{N}\sum_{j=1}^{N} |{<}z_k^t, \bar{z}_j^i{>}| = Z^t(\bar{Z}^i)^\top \tag{4.3}$$

In the above two equations, Z^i and Z^t represent the cross-modal features from two modalities. N is the number of samples in a mini-batch. These two similarity matrices are normalized by a softmax function. Afterwards, we use KL-divergence to characterize the difference between the normalized matrices and the supervisory matrix, *i.e.* $KL((f(Z^i, Z^t)||f(Y_l^\top, Y_l))$. The specific objective function is introduced in Section 4.4.2.

4.4 Implementation and optimization

We introduce the implementation and optimization of our proposed approach in this section. We employ four convolutional neural networks such as ResNet-152 [13] and MobileNet [187] to obtain image features and a Bi-directional LSTM (Bi-LSTM) [188] to extract text features. All the extracted image and text features are uni-modal. Later, we borrow the protocols of non-shared encoding sub-networks (fully-connected layers) in [186] to get the cross-modal features Z^i and Z^t.

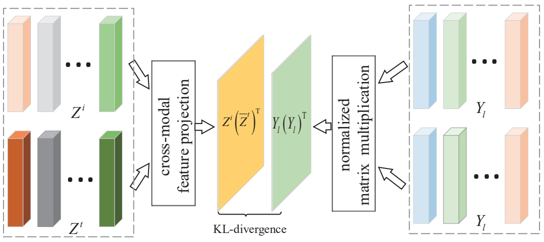


Figure 4.2: KL-divergence for cross-modal feature projection, which considers all features Z^i and Z^t in the shared space. Each paired image feature and text feature share the same instance label, indicated by the same color. The cross-modal feature projection module is critical to explore the similarity between image features and normalized text features. The projection process is formulated in Eqs. 4.2 and 4.3.

Once the cross-modal features are obtained, we use the proposed algorithm to train the networks based on the above theoretical analysis. The algorithm includes combining information entropy and adversarial learning to mitigate the heterogeneity gap, and loss function terms (*i.e.* KL-divergence loss, categorical cross-entropy loss, and bi-directional triplet loss) to preserve semantic similarity between cross-modal features.

4.4.1 Combining information theory & adversarial learning

We combine information entropy predictor and modality classifier in Figure 4.1 into a unified sub-network, as shown in Figure 4.3. In this paradigm, the discriminator D with parameters $\boldsymbol{\theta}_D$ performs a modality classification and computes the Shannon information entropy. The backbone nets E_1 and E_2 for feature extraction act as the generator G. The whole structure forms a generative adversarial network. The information entropy computed from the discriminator back-propagates to the feature encoders. Specifically, when the discriminator is fixed, and its parameters are $\boldsymbol{\theta}^{\star}_{D}$, then the information entropy $H(P^{\star}_{_D}) = \mathbb{E}_{i,t}(-P^{\star}_{_D} * log(P^{\star}_{_D}))$ is computed from its output probabilities $P^{\star}_{_D}(D|Z^{i,t};\boldsymbol{\theta}^{\star}_{D})$ across the features for all classes. Based on the information entropy, we can design a negative entropy loss $L_s = -H(P^{\star}_{_D})$ (see Eq. 4.4) to train the network. The gradients computed from L_s update the parameters of feature extractors. The negative information entropy L_s is label-free during training, and it regularizes the whole feature distribution to be modality-invariant.

The discriminator consists of some fully-connected layers. The last layer with two neurons yields probabilities that correspond to two modalities. This discriminator classifies whether the input features Z^i and Z^t are from the visual or the textual modality given the pre-defined modality label Y^{*}_{c}. In contrast, the generator ($i.e.$ E_1 and E_2) aims at learning modality-invariant features to fool the discriminator to make an incorrect modality classification so that the generator gradually maximizes the output information entropy from the discriminator. Therefore, the learning process of the discriminator affects that of the generator in an indirect way. The objective function is calculated using the output probabilities $P_{_D}(D|Z^{i,t};\boldsymbol{\theta}_D)$ of the discriminator.

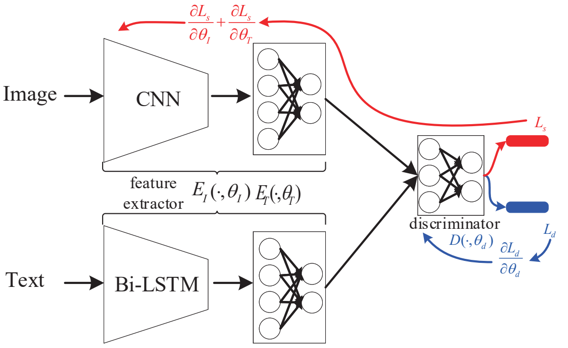

$\frac{\partial L_s}{\partial \theta_I} + \frac{\partial L_s}{\partial \theta_T}$
Image
CNN
Text
Bi-LSTM
feature extractor
$E_I(\cdot,\theta_I)$ $E_T(\cdot,\theta_T)$
L_s
discriminator
$D(\cdot,\theta_d)$
$\frac{\partial L_d}{\partial \theta_d}$
L_d

Figure 4.3: The implementation of integrating information entropy predictor and modality classifier in Figure 4.1 into a unified discriminator. Together with the feature extractors, the whole framework is in the form of generative adversarial network. For clarity, we ignore the feature projector, which includes label classification loss, bi-directional triplet loss, and KL-divergence loss.

For the generator E_1 and E_2:

$$\begin{aligned} L_s = & \frac{1}{N}\sum_{j=1}^{N}\sum_{m=1}^{M}\Big(P^{i}_{D,m}(D^i|Z^i_j;\boldsymbol{\theta}_D)*log(P^{i}_{D,m}(D^i|Z^i_j;\boldsymbol{\theta}_D)) \\ & + P^{t}_{D,m}(D^t|Z^t_j;\boldsymbol{\theta}_D)*log(P^{t}_{U,m}(D^t|Z^t_j;\boldsymbol{\theta}_D))\Big) \\ & s.t. \sum_{m=1}^{M} P^{*}_{D,m}(D^*|Z^*_j;\boldsymbol{\theta}_D) = 1,\ P^{*}_{D,m}(D^*|Z^*_j;\boldsymbol{\theta}_D) \geq 0 \end{aligned} \tag{4.4}$$

It is expected for the generator G to maximize the information entropy $H(P^{\star}_{_D})$, and subsequently the modality uncertainty (see Figure 3.2 in Chapter 3). Since L_s is a negative entropy ($L_s =-H(P^{\star}_{_D})$) to maximize $H(P^{\star}_{_D})$, it is minimized to optimize the parameters $\boldsymbol{\theta}_{E_1}$ and $\boldsymbol{\theta}_{E_2}$ of the generator during training. For the discriminator D, depending on the modality label Y_c^i and Y_c^t and its output probabilities $P_{_D}(D|Z^{i,t};\boldsymbol{\theta}_D)$, the modality classification cross-entropy loss function is formulated as:

$$L_c = -\frac{1}{N}\sum_{j=1}^{N}\Big(Y_c^i{*}log\big(P_D^i(D^i|Z_j^i;\boldsymbol{\theta}_D)\big) + Y_c^t{*}log\big(P_D^t(D^t|Z_j^t;\boldsymbol{\theta}_D)\big)\Big) \tag{4.5}$$

L_c refers to the negative cross-entropy loss of the discriminator and is minimized to clearly classify image and text features into two modalities during training. Note that the gradients calculated from term L_s are only used to optimize the parameters $\boldsymbol{\theta}_{E_1}$ and $\boldsymbol{\theta}_{E_2}$ of the generator, whereas the gradients from term L_c are only for optimizing the parameters $\boldsymbol{\theta}_D$ of the discriminator, as shown in Figure 4.3. Minimizing loss L_c and L_s when trained iteratively will reduce the heterogeneity gap. The optimization method is straightforward, even though the gradients calculated from L_c will not directly affect the parameters of the feature encoders E_1 and E_2. The output probabilities of the discriminator change when updating its parameters, which will affect the Shannon information entropy and affect the output features from E_1 and E_2 in the end.

4.4.2 KL-divergence for similarity preserving

We also compute KL-divergence directly across Z^i and Z^t to further preserve semantic similarity. KL-divergence focuses on the projections of image and text features and is computed by $L_{kl} = KL((f(Z^i, Z^t)||f(Y_l^\top, Y_l))$. Here, superscript “$^\top$” means matrix transpose. L_{kl} focuses on constraining the whole feature distributions and is complementary to the following bi-directional triplet loss function. We have introduced the process of cross-modal feature projection in Section 4.3.3. Given the similarity matrices (*i.e.* $A_{i\to t}(Z^i, Z^t)$ and $A_{t\to i}(Z^t, Z^i)$), we use the softmax function to normalize these matrices in Eq. 4.6 and Eq. 4.7. The supervisory matrix is normalized after matrix multiplication as in Eq. 4.8. Similar to [186], since we project features from visual (or textual) modality into textual (or visual) modality, the KL-divergence regularizes the semantics in bi-directional feature projection, which is formulated in Eq. 4.9 as:

$$P_{i\to t} = \frac{exp\big(A_{i\to t}(Z^i, Z^t)\big)}{\sum exp\big(A_{i\to t}(Z^i, Z^t)\big)} \tag{4.6}$$

$$P_{t\to i} = \frac{exp\big(A_{t\to i}(Z^t, Z^i)\big)}{\sum exp\big(A_{t\to i}(Z^t, Z^i)\big)} \tag{4.7}$$

$$Q_y = \frac{exp(Y_l^\top Y_l)}{\sum exp(Y_l^\top Y_l)} \tag{4.8}$$

$$\begin{aligned} L_{kl} &= L_{kl_{i \to t}} + L_{kl_{t \to i}} \\ &= \frac{1}{N}\Big(\sum\sum P_{i \to t} * log(\frac{P_{i \to t}}{Q_y + \varepsilon}) + \sum\sum P_{t \to i} * log(\frac{P_{t \to i}}{Q_y + \varepsilon})\Big) \end{aligned} \tag{4.9}$$

where ε is a small constant to avoid division by zero. Loss L_{kl} refers to the KL-divergence between the projections of image-text features and their supervisory matrix. This loss is minimized and the gradients computed from L_{kl} are used to update the parameters $\boldsymbol{\theta}_{E_1}$ and $\boldsymbol{\theta}_{E_2}$ of the generator, thereby the semantics between image features and text features can be associated.

4.4.3 Instance label classification

4.4.3.1 Categorical cross-entropy loss

Label classification is a popular idea for cross-modal features learning [186]. We use the instance labels provided on the datasets for label classification. For categorical cross-entropy loss, we apply the norm-softmax strategy and feature projection in [186] to learn more discriminative cross-modal features. On the one hand, the normalized parameters $\boldsymbol{\theta}_P$ in the label classifier encourage cross-modal features to distribute more compactly so that the softmax classifier performs label classification correctly. On the other hand, projection between image and text features strengthens their similarity association and is beneficial for label classification [186]. Feature projection can be computed using Eq. 4.1. Subsequently, given the instance label y_l, categorical cross-entropy loss L_{ce} is defined by Eq. 4.10 and is minimized during training[1]:

$$\begin{aligned} L_{ce} &= \mathbb{E}_{i,t}(-y_l * log(p_{_P}(c|Z^{i,t}; \boldsymbol{\theta}_P))) \\ = -\frac{1}{N}\Big(\sum_{j=1}^{N} y_{l,j} * log\Big(&\frac{exp(\mathbf{W}_{y_{l,j}}^{\top} \hat{z}_j^{i \to t})}{\sum_j exp(\mathbf{W}_j^{\top} \hat{z}_j^{i \to t})}\Big) + \sum_{j=1}^{N} y_{l,j} * log\Big(\frac{exp(\mathbf{W}_{y_{l,j}}^{\top} \hat{z}_j^{t \to i})}{\sum_j exp(\mathbf{W}_j^{\top} \hat{z}_j^{t \to i})}\Big)\Big) \\ s.t. \quad &||\mathbf{W}_j|| = 1; \hat{z}_j^{i \to t} = <z_j^i, \bar{z}_j^t> * \bar{z}_j^t; \hat{z}_j^{t \to i} = <z_j^t, \bar{z}_j^i> * \bar{z}_j^i \end{aligned} \tag{4.10}$$

where N is the number of image-text pairs in a mini-batch. $W_{y_{l,j}}$ and W_j represent the $y_{l,j}$-th and the j-th column of weights $\mathbf{W}$ in classifier parameters $\boldsymbol{\theta}_P$ according to [186]. $\hat{z}_j^{i \to t}$ and $\hat{z}_j^{t \to i}$ are the projections image to text and the projections text to image, respectively, by using Eq. 4.1.

4.4.3.2 KL-divergence for data imbalance

Label classification using categorical cross-entropy loss can preserve semantic similarity between cross-modal features. However, we argue that there also exists a data imbalance issue when training the label classifier because each image is described

[1] We omit the bias term for simplicity

by more than one sentence (*e.g.* each image has five description sentences in the Flickr30K dataset). In the end, it causes the learned label classifier to prefer text features.

The issue of data imbalance in cross-modal retrieval can be resolved by constructing an augmented semantic space to re-align features. In this work, we use the temperature scaling [189] to tackle the data imbalance issue. The biased label classifier can be calibrated by re-scaling its output probabilities *i.e.*, $p^{i \to t}{=}softmax(\frac{\mathbf{W}^\top \hat{z}^{i\to t}}{\tau})$ and $p^{t \to i} = softmax(\frac{\mathbf{W}^\top \hat{z}^{t\to i}}{\tau})$, respectively. Re-scaling the probabilities with temperature τ raises the output entropy so better image-text matching can be observed [189]. Subsequently, we use KL-divergence to measure the differences between the re-scaled probabilities. Since the magnitudes of the gradients produced by the re-scaling probabilities scale as $1/\tau^2$, it is important to multiply them by τ^2. Finally, the KL-divergence loss on the scaling probabilities for data imbalance can be formulated as L_{di}:

$$\begin{aligned} L_{di} = & \frac{\tau^2}{N} \sum \sum \Big(p^{i\to t} * log(\frac{p^{i\to t}}{p^{t\to i}+\varepsilon}) + p^{t\to i} * log(\frac{p^{t\to i}}{p^{i\to t}+\varepsilon}) \Big) \\ s.t.\ & p^{i\to t}{=}softmax\Big(\frac{\mathbf{W}^\top \hat{z}^{i\to t}}{\tau}\Big), p^{t\to i} = softmax\Big(\frac{\mathbf{W}^\top \hat{z}^{t\to i}}{\tau}\Big) \end{aligned} \tag{4.11}$$

where ε is a small constant to avoid division by zero. With $\tau = 1$, we recover the original KL-divergence. As reported in Table 4.5, we find that the parameter τ can affect the effectiveness of loss L_{di}. Minimizing loss L_{di} effectively reduces the influence of data imbalance issue and improves retrieval accuracy. The final objective function for label classification is $(L_{ce} + L_{di})$. The gradients calculated from loss $(L_{ce} + L_{di})$ are used to optimize the parameters $\boldsymbol{\theta}_{E_1}$, $\boldsymbol{\theta}_{E_2}$, and $\boldsymbol{\theta}_P$ in the generator and the label classifier, respectively.

4.4.4 Bi-directional triplet constraint

The triplet constraint is commonly used for feature learning. To achieve the baseline performance, we use this constraint from an inter-modality and an intra-modality perspective to strengthen the discrimination of cross-modal features.

Given cross-modal features Z^i and Z^t in the shared space, the cosine function is used to measure global similarity between feature vectors, *i.e.* $S_{jk} = (Z^i_j)^\top Z^t_k$. We adopt the hard sampling strategy to select three-tuples features from an inter-modality and an intra-modality viewpoint. Hence, the inter-modality and intra-modality triplet loss functions are formulated as:

$$L_{inter} = \frac{1}{N}\Big(\sum_{j,k^+,k^-}^{N} \max[0, m - S_{j,k^+} + S_{j,k^-}] + \sum_{k,j^+,j^-}^{N} \max[0, m - S_{k,j^+} + S_{k,j^-}] \Big) \tag{4.12}$$

Algorithm 1: Whole network training and optimization pseudocode

```
Input: mini-batch images $X^i$, text $X^t$, instance label $Y$, modality label ($Y_c^i$,
  $Y_c^t$), total training batch $S$, pre-trained parameters $\boldsymbol{\theta}_{E_1}$, update steps $k$
Output: the embedded cross-modal features $Z^i$ and $Z^t$ in Figure 4.1
Initialize hash functions: learning rate $lr_1$, $lr_2$, $\boldsymbol{\theta}_{E_2}$, $\boldsymbol{\theta}_P$, $\boldsymbol{\theta}_D$
For $n = 1$ to $S$
    For $k$ steps
        cross-modal features embedding:
        $Z^i = E_1(X^i; \boldsymbol{\theta}_{E_1})$    //Embed image features into the shared space
        $Z^t = E_2(X^t; \boldsymbol{\theta}_{E_2})$    //Embed text features into the shared space
        loss computing and feature optimization:
        $L_{ce}, L_{di}, L_{tr}, L_{kl}$ calculation    //Eqs. 4.10, 4.11, 4.14, 4.9
        $P_D^i = D(Z^i; \boldsymbol{\theta}_D)$    //Discriminator D
        $P_D^t = D(Z^t; \boldsymbol{\theta}_D)$
        $L_s, L_c$ calculation //Eqs. 4.4, 4.5
        fix $\boldsymbol{\theta}_D$, update parameters $\boldsymbol{\theta}_{E_1}$, $\boldsymbol{\theta}_{E_2}$, $\boldsymbol{\theta}_P$:
        $\boldsymbol{\theta}_P \leftarrow \boldsymbol{\theta}_P - lr_2 \cdot \nabla_{\boldsymbol{\theta}_P}(L_{ce} + L_{di})$
        $\boldsymbol{\theta}_{E_1} \leftarrow \boldsymbol{\theta}_{E_1} - lr_1 \cdot \nabla_{\boldsymbol{\theta}_{E_1}}(L_{ce} + L_{di} + L_{tr} + L_{kl} + L_s)$
        $\boldsymbol{\theta}_{E_2} \leftarrow \boldsymbol{\theta}_{E_2} - lr_2 \cdot \nabla_{\boldsymbol{\theta}_{E_2}}(L_{ce} + L_{di} + L_{tr} + L_{kl} + L_s)$
    End for
    fixate $\boldsymbol{\theta}_P$, $\boldsymbol{\theta}_{E_1}$, $\boldsymbol{\theta}_{E_2}$, update parameters $\boldsymbol{\theta}_D$:
    $\boldsymbol{\theta}_D \leftarrow \boldsymbol{\theta}_D - lr_2 \cdot \nabla_{\boldsymbol{\theta}_D}(L_c)$
End for
```

$$L_{intra} = \frac{1}{N}\Big(\sum_{j,j^+,j^-}^{N} \max[0, m - S_{j,j^+} + S_{j,j^-}] + \sum_{k,k^+,k^-}^{N} \max[0, m - S_{k,k^+} + S_{k,k^-}]\Big) \tag{4.13}$$

$$L_{tr} = L_{inter} + L_{intra} \tag{4.14}$$

where m is the margin in the bi-directional triplet loss function. For instance, in case of inter-modality, $S_{j,k^+} = (Z_j^i)^\top Z_{k^+}^t$, where the anchor features are selected from the visual modality, while the positive features are selected from the textual modality. In case of intra-modality, $S_{j,j^+} = (Z_j^i)^\top Z_{j^+}^i$, both the anchor features and the positive features are selected from the visual modality. Minimizing bi-directional triplet loss L_{tr} keeps the correlated image-text pairs closer to each other, while the uncorrelated image-text pairs are pushed away. This loss directly operates on the cross-modal features Z^i and Z^t so that the gradients from it optimize the parameters $\boldsymbol{\theta}_{E_1}$ and $\boldsymbol{\theta}_{E_2}$ of the generator.

The problem of integrating information theory and adversarial learning for cross-modal retrieval is formally defined, in Eq. 4.15, as a min-max game using the previously defined loss terms. We further introduce the complete procedure of training and optimization in Algorithm 1. Finally, when trained to convergence, the network yields cross-modal features Z^i and Z^t in the shared space, as shown in Figure 4.1. These return cross-modal features are used for performing retrieval.

$$\begin{cases} \min\limits_{\boldsymbol{\theta}_{E_1},\boldsymbol{\theta}_{E_2},\boldsymbol{\theta}_P} \max\limits_{\boldsymbol{\theta}_D}(L_{ce} + L_{di} + L_{kl} + L_{tr} + L_s) \\ \min\limits_{\boldsymbol{\theta}_D} L_c \end{cases} \tag{4.15}$$

4.5 Experiments

4.5.1 Datasets and settings

We demonstrate the efficacy of the proposed method on the Flickr8K [190], Flickr30K [191], Microsoft COCO [192], and CUHK-PEDES [193] datasets. Each image in these datasets is described by several descriptive sentences. For Flickr8K, we adopt the standard dataset splitting method to obtain a training set (6K), a validation set (1K), and a test set (1K). For Flickr30K, we follow the previous work [186] and use 29,783 images for training, 1,000 images for validation and 1,000 images for testing. For MS-COCO, we follow the training protocol in [186] and split this dataset into 82,783 training, 30,504 validation and 5,000 test images, and then report the performance on both 5K and 1K test set. For CUHK-PEDES, it contains 40,206 pedestrian images of 13,003 identities. Following [186], we split this dataset into 11,003 training identities with 34,054 images, 1,000 validation identities with 3,078 images and 1,000 test identities with 3,074 images. Note that all captions for the same image are used as separate image-text pairs to train network.

Models are trained on GEFORCE TITAN X and Tesla K40 GPUs. To extract text features, the embedded words are fed into a Bi-LSTM to capture vectors with dimension 1024 (1024-D). We follow [186] and set the Bi-LSTM with dropout rate 0.3. For fair comparison, we adopt ResNet [13], MobileNet [187], and VGGNet [61] as the backbone to extract image features and further fine-tune them with learning rate $lr_1 = 2 \times 10^{-5}$, decaying every 2 epochs exponentially. The output 2048-D image features and 1024-D text features are further projected into a shared space. Then cross-modal features in the space are 512-D vectors (*i.e.* Z^i and Z^t in Figure 4.1). The batch size is set to 64 or 32 depending on available GPUs memory. For the bi-directional triplet loss function, initially, we treat the inter-modality and intra-modality sampling identically although each of them might have different contributions [194], we empirically set the margin to $m = 0.5$. The re-scaling parameter τ for data imbalance issue is set as $\tau = 4$ (see Table 4.5). In practice, the discriminator can classify image and text modality easily at the start of training, so the generator typically requires multiple (*e.g.*, 5) update steps per discriminator update step during training (see Algorithm 1).

Once trained to converge, the network yields image features Z^i and text features Z^t. We use the cosine function to measure their similarity. We use Recall@K (K=1,

Table 4.1: Comparison of retrieval results on the Flickr30K [191] and MS-COCO [192] dataset (R@K (K=1,5,10)(%))

		Flickr30K						MS-COCO					
Method	Backbone Net	Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
m-RNN [196]	VGG	35.4	63.8	73.7	22.8	50.7	63.1	41.0	73.0	83.5	29.0	42.2	77.0
RNN+FV [197]	VGG	35.6	62.5	74.2	27.4	55.9	70.0	41.5	72.0	82.9	29.2	64.7	80.4
DSPE+FV [194]	VGG	40.3	68.9	79.9	29.7	60.1	72.1	50.1	79.7	89.2	39.6	75.2	86.9
CMPM+CMPC† [186]	MobileNet	40.3	66.9	76.7	30.4	58.2	68.5	52.9	83.8	92.1	41.3	74.6	85.9
Word2VisualVec [198]	ResNet-152	42.0	70.4	80.1	-	-	-	-	-	-	-	-	-
sm-LSTM [199]	VGG	42.5	71.9	81.5	30.2	60.4	72.3	53.2	83.1	91.5	40.7	75.8	87.4
RRF-Net [200]	ResNet-152	47.6	77.4	87.1	35.4	68.3	79.9	56.4	85.3	91.5	43.9	78.1	88.6
Joint learning [143]	ResNet-152	48.6	73.6	83.6	32.3	62.5	74.0	55.3	82.7	90.2	41.7	75.0	87.4
CMPM+CMPC‡ [186]	ResNet-152	49.6	76.8	86.1	37.3	65.7	75.5	-	-	-	-	-	-
VSE++ [184]	ResNet-152	52.9	80.5	87.2	39.6	70.1	79.5	51.3	82.2	91.0	40.1	75.3	86.1
TIMAM [201]	ResNet-152	53.1	78.8	87.6	42.6	71.6	81.9	-	-	-	-	-	-
DAN [202]	ResNet-152	55.0	81.8	89.0	39.4	69.2	79.1	-	-	-	-	-	-
Dual-path stage I [203]	ResNet-152	44.2	70.2	79.7	30.7	59.2	70.8	52.2	80.4	88.7	37.2	69.5	80.6
Dual-path stage II [203]	ResNet-152	55.6	81.9	89.5	39.1	69.2	**80.9**	**65.6**	**89.8**	**95.5**	47.1	79.9	90.0
Our ITMeetsAL	VGG	38.5	66.5	76.3	30.7	59.4	70.3	44.2	76.1	86.3	37.1	72.7	85.1
Our ITMeetsAL	MobileNet	46.6	73.5	82.5	34.4	63.3	74.2	54.7	84.3	91.1	41.0	76.7	88.1
Our ITMeetsAL	ResNet-152	**56.5**	**82.2**	**89.6**	**43.5**	**71.8**	80.2	58.5	85.3	92.1	**48.3**	**82.0**	**90.6**

MS-COCO is tested on 1K setting. The best results are in boldface and the second best ones are underlined.

5, 10) for evaluation and comparison. Moreover, we adopt the precision-recall and mAP for the ablation study, and visualize their feature distributions by t-SNE [195]. Furthermore, we display the cross-modal retrieval results using our method.

4.5.2 Performance evaluation

4.5.2.1 Results on the Flickr30K and MS-COCO datasets

The retrieval results on Flickr30K and MS-COCO are reported in Table 4.1. Hereafter, "Image-to-Text" means using an image as a query item to retrieve semantically-relevant text from the textual gallery. "Text-to-Image" means using a text as query to retrieve images from the visual gallery. In most cases, our proposed approach shows the best performance when using three different deep networks. For the "Image-to-Text" task on the MS-COCO dataset, the best results are obtained by Zheng *et al.* [203], which adopted a deeper network for text feature learning and used a two-stage training strategy. However, for the "Text-to-Image" task and the "Image-to-Text" task on the Flickr30K dataset, our method performs better. Take ResNet-152 as an example, the results are R@1=43.5% on the Flickr30K and R@1=48.3% on the MS-COCO for "Text-to-Image" task; the results are R@1=56.5% on the Flickr30K dataset and R@1=58.5% on the MS-COCO dataset for "Image-to-Text" task.

The learning capacity of deep networks would affect retrieval performance significantly. For visual feature learning, deeper CNNs usually achieve better results than their shallower counterparts. This can be observed from Table 4.1, the retrieval results based on ResNet-152 are usually higher than those of MobileNet and VGG. Moreover, our method also has good performance using MobileNet. For instance, regarding the "Image-to-Text" task on the Flickr30K dataset, the recall

Table 4.2: Retrieval results on the CUHK-PEDES [193] dataset.

Method	Backbone Net	Text-to-Image		
		R@1	R@5	R@10
Latent co-attention [204]	VGG	25.94	-	60.48
Local-global association [205]	ResNet-50	43.58	66.93	76.26
CMPM [186]	MobileNet	44.02	-	77.00
Dual-path two-stage [203]	ResNet-152	44.40	66.26	75.07
MIA [206]	ResNet-50	48.00	70.70	79.30
CMPM+CMPC [186]	MobileNet	49.37	-	79.27
Our ITMeetsAL	VGG	44.43	68.26	77.50
Our ITMeetsAL	MobileNet	51.85	73.36	81.27
Our ITMeetsAL	ResNet-50	50.63	73.33	81.34
Our ITMeetsAL	ResNet-152	**55.72**	**76.15**	**84.26**

result of CMPM+CMPC [186] is R@1=40.3%, but the result from our method is R@1=46.6%, which is a significant improvement. Likewise, for textual modality, a powerful extractor provides better semantic-aware features, providing better results. This can be observed on the comparisons between our proposed "ITMeetsAL", m-RNN [196] and RNN+FV [197]. Concretely, both of them leverage VGG to extract image features, but m-RNN [196] and RNN+FV [197] extract textual features using RNN, which is less powerful than the Bi-LSTM as we used in our experiments.

We obverse that the strategy for network training is critical for retrieval tasks. Take [203] as an example, the backbone network (ResNet-152) is fixed at stage I (R@1=44.2% on "Image-to-Text" task on Flickr30K) and then fine-tuned with a small learning rate on stage II (R@1=55.6% on the "Image-to-Text" task on Flickr30K). In contrast, our network structure is trained end-to-end in only one stage (we fine-tune the backbone network with a small learning rate from the beginning). Our reported results are close to those in two-stage dual learning [203]. When tested on the Flickr30K dataset for the "Image-to-Text" task, the recall results are R@1=56.5%, R@5=82.2%, R@10=89.6%, which are the best overall previous methods.

Considering the two branches of "Image-to-Text" task and the "Text-to-Image" task, we think that the data imbalance issue still influences the performance of each branch. More specifically, for all listed methods, the "Image-to-Text" task has better performance, which indicates that the network still has more biases on text feature learning as a result of the issue of data imbalance. Thus, there exists more room for improvement using other strategies, such as data augmentation.

4.5.2.2 Results on CUHK-PEDES dataset

The "Text-to-Image" retrieval results on the CUHK-PEDES dataset are reported in Table 4.2. We evaluate the proposed method using four deep networks. All results indicate that our method outperforms other counterparts. The optimal results are achieved with R@1=55.72% using ResNet-152 as backbone network. The results using MobileNet are sub-optimal but also have some improvements. For

Table 4.3: Retrieval results on the Flickr8K [190] dataset (R@K (K=1,5,10)(%))

Method	Backbone Net	Image-to-Text		
		R@1	R@5	R@10
RNN+FV [197]	VGG	23.2	53.3	67.8
GMM+HGLMM [207]	VGG	31.0	59.3	73.7
Word2VisualVec [198]	ResNet-152	33.4	63.1	75.3
Joint learning [143]	ResNet-152	**40.6**	**67.8**	78.6
Our ITMeetsAL	VGG	28.0	52.7	63.1
Our ITMeetsAL	MobileNet	30.9	58.6	70.8
Our ITMeetsAL	ResNet-152	40.1	**67.8**	**79.2**

The best results are in boldface and the second best results are underlined.

Table 4.4: Component analysis on the Flickr30K [191] (R@1, R@10, and mAP (%))

Method using MobileNet	Flickr30K					
	Image-to-Text			Text-to-Image		
	R@1	R@10	mAP	R@1	R@10	mAP
Baseline1: Only L_{ce}+L_{tr}	40.6	80.8	23.1	31.9	72.2	31.9
Baseline2: L_{ce}+L_{tr}+L_{di}	42.3	80.6	24.4	32.5	73.0	32.5
Baseline3: L_{ce}+L_{tr}+L_{di}+L_{kl}	44.7	81.0	25.2	32.6	73.2	32.6
Full method: L_{ce}+L_{tr}+L_{di}+L_{kl}+L_s+L_c	46.6	82.5	26.3	34.4	74.1	34.4

example, CMPM+CMPC achieves a recall R@1=49.37% and R@10=79.27%, while our method obtains R@1=51.85% and R@10=81.27%. Moreover, the results of our method show that deeper networks achieve better retrieval performance, whereas the light-weight MobileNet has a similar performance as ResNet-50.

4.5.2.3 Results on Flickr8K dataset

The retrieval results on the Flick8K dataset are reported in Table 4.3. The best results R@1=40.6%, R@5=67.8%, R@10=78.6% are achieved by joint correlation learning [143] where a batch-based triplet loss, which considers all image-sentences pairs, is used for learning correlations. The second-best results are achieved using ResNet-152 (same as [143]) R@1=40.1%, R@5=67.8%, R@10=79.2%, which has better R@10 performance compared to [143]. Our method shows competitive results compared to other counterparts and also indicates that there exists room for further performance improvement.

4.5.3 Ablation study

For analyzing the effect of each component, the ablation study are conducted on the Flickr30K dataset using MobileNet as a backbone net, we use the commonly used categorical cross-entropy L_{ce} and bi-triplet loss function L_{tr} to construct the baseline in Table 4.4, we call this **Baseline1** configuration “Only $L_{ce} + L_{tr}$”.

4.5.3.1 Analysis of KL-divergence for data imbalance

Each image in a dataset (*e.g.* Flickr30k) has more than one description sentence. We think this leads to a data imbalance issue for cross-modal feature learning. The network has more text data for training, which causes the learned label classifier to prefer text features. Therefore, we adopt a regularization term L_{di} based on KL-divergence to calibrate this bias. To this end, the label classifier can be re-calibrated on the image features and text features. In Table 4.4, this **Baseline2** configuration is named " $L_{ce} + L_{tr} + L_{di}$". The Recall and mean Average Precision (mAP) show the effectiveness of this loss. Compared to Baseline1, the scaling KL-divergence loss L_{di} contributes more on Recall@1 for both the "Image-to-Text" (42.3%) and "Text-to-Image" task (32.5%).

4.5.3.2 Analysis of KL-divergence for cross-modal feature projection

KL-divergence is obtained by adding L_{kl} which constrains the image features and text features in the shared space under the supervision of supervisory matrix. It focuses on the whole feature distribution and is complementary to the bi-directional triplet loss function. We denote **Baseline3** as "$L_{ce} + L_{tr} + L_{di} + L_{kl}$" in Table 4.4. As we can see, Recall@1 of the "Image-to-Text" task has been improved significantly by 2.4%. However, the KL-divergence loss shows a slight improvement on the "Text-to-Image" task. The results indicate that the KL-divergence loss function contributes more to image feature learning, which might be caused by the issue of data imbalance of the dataset.

4.5.3.3 Analysis of adversary combining

The prior loss terms have been used to constrain the similarity of the image-text features in the shared space. Intuitively, two-tuple or three-tuple feature exemplars are helpful for reducing the "semantic gap" and further making the whole feature distribution close at the same time. However, the constraint loss functions (*e.g.* cosine similarity) cannot constrain the distribution discrepancy of the whole distribution because these loss functions are symmetrical. Focusing on the whole feature distribution, we combine the Shanon information entropy L_s and the modality classification loss L_c in an adversary training manner to reduce the heterogeneity gap. This **full method** is named "$L_{ce} + L_{tr} + L_{di} + L_{kl} + L_s + L_c$" and corresponding results are shown in Table 4.4. Compared to former baselines, the results obtained by using our method are improved significantly.

Furthermore, we compare the precision-recall curves for the above four configurations and baselines, the results are shown in Figure 4.4. The larger the area under the curve, the better the algorithm. Regarding the different tasks, the improvements are slightly different. Overall, we can see that each added component helps to improve the overall performance of the retrieval algorithm.

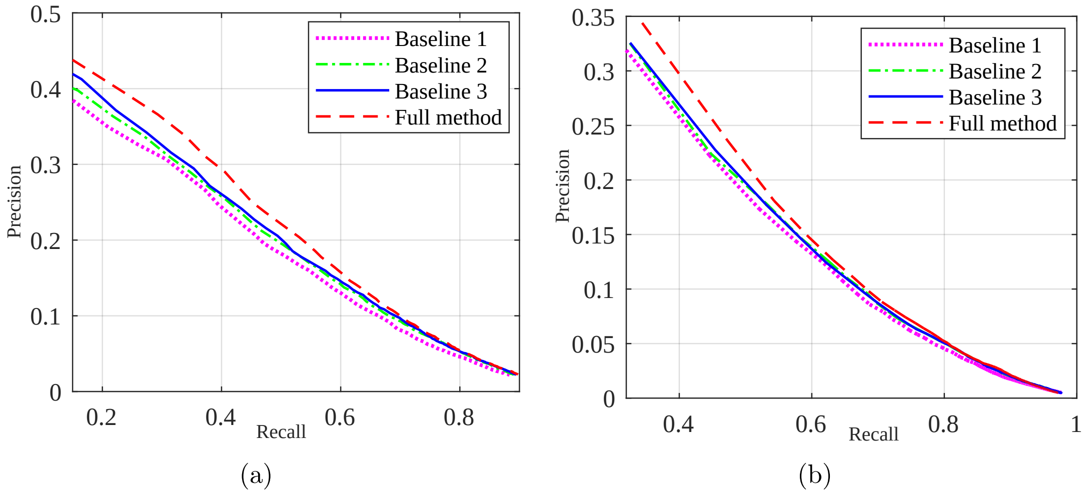


Figure 4.4: The precision_recall curves from "Baseline1" to "Full method" on Flickr30K, each line corresponds one experimental configuration in Table 4.4. The larger area under the line indicates better performance.

Table 4.5: Temperature scaling analysis for loss L_{di} (R@1, R@10, and mAP (%))

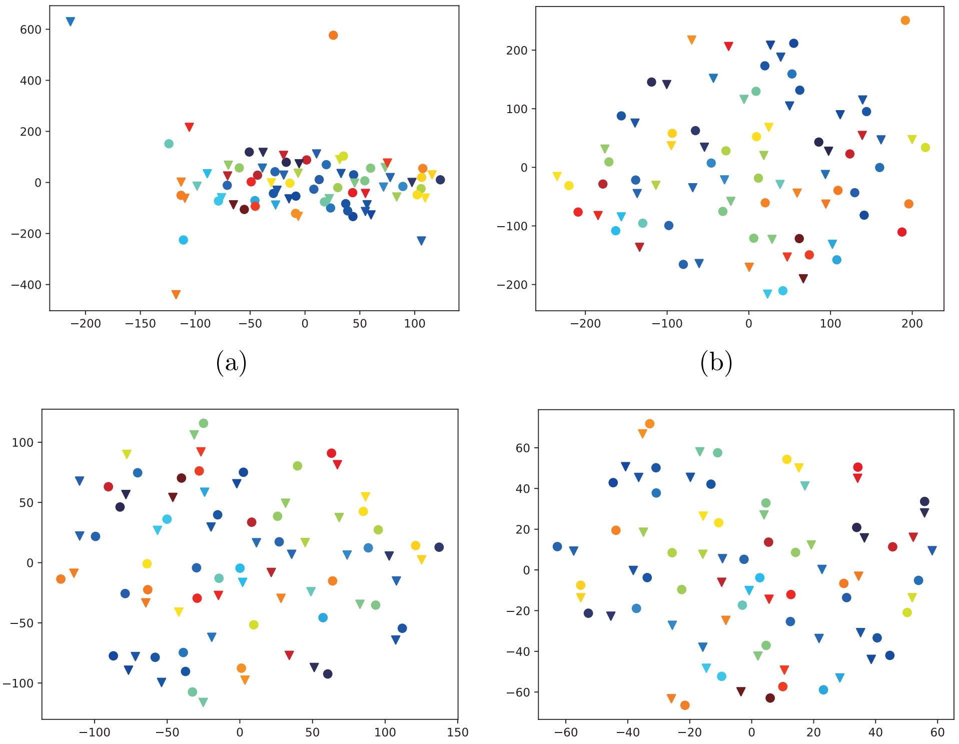
Temperature	Flickr30K					
	Image-to-Text			Text-to-Image		
	R@1	R@10	mAP	R@1	R@10	mAP
τ=1	44.0	80.6	24.8	32.9	73.5	32.9
τ=2	45.3	80.9	25.6	33.6	73.6	33.6
τ=3	46.2	**83.2**	25.7	33.3	73.4	33.3
τ=4	**46.6**	82.5	**26.3**	**34.4**	**74.2**	**34.4**
τ=5	46.0	81.6	26.1	34.3	73.9	34.3
τ=6	45.9	80.2	26.1	33.1	73.4	33.1

4.5.3.4 Analysis of temperature τ

We analyze the temperature parameter τ in loss L_{di} in Eq. 4.11. Other loss terms are kept the same with the full method, *i.e.* "$L_{ce} + L_{tr} + L_{di} + L_{kl} + L_s + L_c$". We vary this parameter τ from 1 to 6, and their corresponding results are reported in Table 4.5. We can observe that the optimal results are achieved if the classifier's output probabilities are re-scaled by $\tau = 4$. As claimed in [189], the temperature scaling raises the output entropy of the classifier with $\tau > 1$. In our experiments, we found it is beneficial for improving the image-text matching.

4.5.3.5 Distribution visualization

We choose 40 image-text pairs from the Flickr30K dataset to visualize their feature distributions using t-SNE [195]. We only choose the first description caption among the five sentences. In Figure 4.5, the circle and the triangle shape denote text features and image features, respectively. Label information is represented by a different color.



(a) (b)

(c) (d)

Figure 4.5: Feature distribution visualizations for the ablation study. The shape represents modality and the color indicates the label information. Sub-figures (a)∼(d) correspond to the four experimental configurations in Table 4.4. When each loss function is gradually applied, the paired image features and text features have smaller distances. Best viewed in color.

This distribution indicates the effectiveness of each component (*e.g.* KL-divergence for cross-modal feature projection, and the Shannon information entropy trained in an adversarial manner). In Figure 4.5(a), there exist several feature outliers within the distribution and the proximity relationship between pair-wise features is not obvious. When using the proposed components, the features distribute much better. For example, in Figure 4.5(d), all loss functions are utilized to constrain feature learning, the pair-wise feature shows a close proximity relationship. Moreover, image features and text features are distributed within smaller ranges (-60 ∼ 60). Few outliers exist among the whole distribution.

Qualitative retrieval results on the Flickr30K and the CUHK-PEDES dataset are shown in Figure 4.9. For the "Image-to-Text" task, the proposed method can return almost all paired text of the query image. The "Image-to-Text" task also has good performance, the proposed method retrieves the paired image correctly. Also, other retrieved images show contents relevant to the query sentence.

4.5.3.6 Analysis of complexity and stability

We analyze the complexity of the proposed method by evaluating FLOPs (network forward pass), parameter size, and inference time for each image-text pair on Flickr8k. The results are reported in Table 4.6. The complexity of the proposed framework, implemented by three networks, performs differently. It is well known that VGG has a larger model size and more parameters, which increase computation cost. As a result, the FLOPs of the VGG-based framework achieve 3.1×10^{10}, while the lightweight MobileNet reaches FLOPs to 1.1×10^{9}. Although ResNet-152 has more layers than VGG and MobileNet, it achieves in-between FLOPs to 2.2×10^{10}. The model complexity also leads to different inference times for each image-text input. Take MobileNet on Flickr30k as an example, its inference time is 14.8±3.2 *ms*, relatively faster than these of VGG and ResNet-152.

Table 4.6: Comparisons of model size and computation complexity. FLOPs: the number of FLoating-point OPerations;

Dataset	Backbone Net	FLOPs (forward pass)	#Parameters (million)	Inference time (ms) (per image-text pair)
Flickr8k	Based on VGG	3.1×10^{10}	147.2	114.32±2.5
	Based on MobileNet	1.1×10^{9}	14.6	15.6±3.1
	Based on ResNet-152	2.2×10^{10}	70.1	110.6±2.3

An algorithm is stable if it produces consistent predictions with respect to small perturbations of training samples [208, 209, 210]. Therefore, stability of a learning algorithm holds if statistical conclusions are robust or stable to appropriate perturbations to data [209]. According to this definition, we conduct a stability analysis based on Flickr8k using MobileNet. We add Gaussian noise $N \sim \mathcal{N}(\mu, \sigma^2)$ to change the image-text pairs, with a varying σ. For this purpose, first, we build up an upper-bound performance where no Gaussian noise is added. Second, we vary the σ and collect the corresponding output and then evaluate its Recall rate.

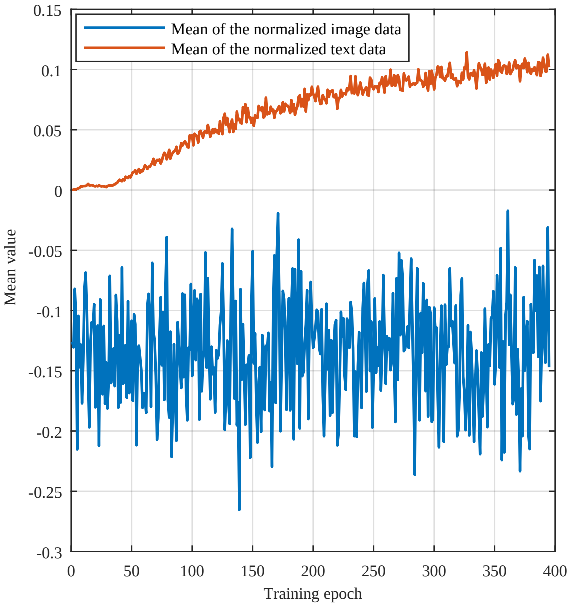
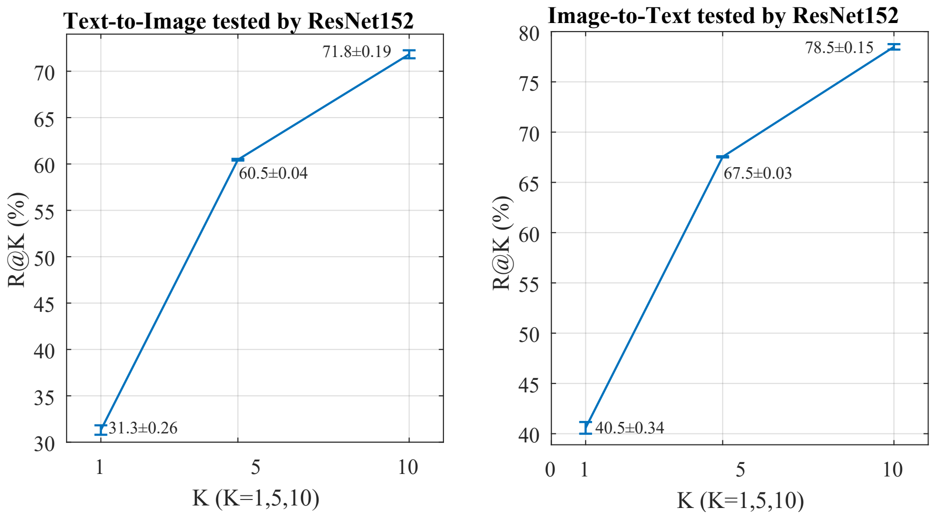


Figure 4.6: The means of the input image and text data (after normalization).

Since the training data have been normalized before feeding into the network and they have small means, as depicted in Figure 4.6. Since the magnitude of image and text inputs are small, we determine the mean of the Gaussian noise by the corresponding means of image and text inputs in each training epoch.

Table 4.7: Stability evaluation on Flickr8k using MobileNet as a backbone net.

Gaussian distribution		Text-to-Image			Image-to-Text		
		R@1	R@5	R@10	R@1	R@5	R@10
With no noise added		23.8	49.7	61.3	30.8	58.9	70.3
$\mu_x = Mean(X^i)$ $\mu_y = Mean(X^t)$	$\sigma = 0.025$	23.8	50.3	61.3	30.7	58.9	71.2
	$\sigma = 0.05$	22.7	47.5	59.0	30.3	56.6	68.6
	$\sigma = 0.075$	22.7	48.0	59.2	29.8	54.8	66.7
	$\sigma = 0.1$	22.2	46.6	58.3	27.7	54.0	65.3



(a)

(b)

Figure 4.7: Error analysis for the proposed model on Flickr8k based on ResNet-152.

The averaged results are reported in Table 4.7. We vary the variance from 0.025 to 0.1, the performance of the proposed framework is relatively stable. For example, for the "Text-to-Image" task, when varying $\sigma = 0.025$ to $\sigma = 0.1$, the result of R@1 changes from 23.8% to 22.2%, decreasing by about 6.7%.

Besides, we also perform error analysis for the performed framework on Flickr8k using ResNet-152. For "Text-to-Image" and "Image-to-Text" tasks, we consider the error bar calculation based on three times running. The results of R@K (K=1,5,10) are illustrated in Figure 4.7. In this error analysis, we observe that the recall results for "Text-to-Image" and "Image-to-Text" tasks have small variations.

4.5.4 Further exploring

We propose to integrate Shannon information entropy with the discriminator for cross-modal retrieval. That is, the discriminator performs modality classification and measures the information entropy at the same time (see Figure 4.3). Herein, we further explore a paradigm to integrate information entropy with adversarial learning. This combining paradigm is more straightforward to the structure in Figure 4.1. Concretely, we build two branches of sub-networks: an uncertainty predictor for modality uncertainty prediction and a modality classifier for modality classification. Then adversarial learning is implemented as an interplay between these two sub-networks with competitive objectives. The uncertainty predictor aims at maximizing the modality uncertainty of the shared space (measured by information entropy), while the modality classifier is to identify image inputs and text inputs by modality classification. We illustrate this combining paradigm in Figure 4.8. Compared to the former paradigm depicted in Figure 4.3, the optimization depicted in Figure 4.8 is different and more complex. The gradients computed by the classifier are used to update parameters $\boldsymbol{\theta}_I$ and $\boldsymbol{\theta}_T$ in the feature extractor. To learn modality-invariant features, the feature extractor *minimizes* the loss of the uncertainty predictor and it *maximizes* the loss $L_d = L_c$ (Eq. 4.5) of the modality classifier, which aims to make image features and text features as similar as possible [211]. The parameters of the modality classifier *minimize* its loss L_d. This training process needs to depend on the gradient reversal layer [211], which would multiply gradient values by -1 when executing back-propagating.

The training procedure is almost the same as used in Algorithm 1 except for the gradients from the modality classification loss that updates the backbone network, leading to a slower training process. The retrieval performance of these two combined methods presented in Figure 4.3 and Figure 4.8 (named as unified and separate, respectively) are given in Table 4.8. The backbone net for image feature extraction is ResNet-152. These two combined strategies show different performances on the four datasets when combining information entropy and modality classification into a unified discriminator. The performance improves slightly on the Flickr30K, MS-COCO, and Flickr8K datasets when adopting the combining strategy of Figure 4.3. However, the method depicted in Figure 4.8 has better performance on the CUHK-PEDES dataset, which is not the common objects dataset. This method has R@1 improved by 3.3% (from 65.58% to 67.79%), Also, the mAP has improved by 1.8% compared to the unified method depicted in Figure 4.3. In summary, the proposed framework of combining information entropy and adversarial learning in Figure 4.3 has better performance and has faster convergence during training.

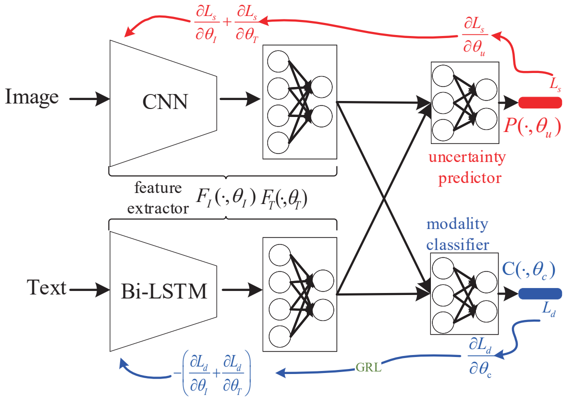

$\frac{\partial L_s}{\partial \theta_I} + \frac{\partial L_s}{\partial \theta_T}$
$\frac{\partial L_s}{\partial \theta_u}$
L_s
Image
CNN
$P(\cdot, \theta_u)$
uncertainty predictor
feature extractor
$F_I(\cdot, \theta_I)$ $F_T(\cdot, \theta_T)$
modality classifier
Text
Bi-LSTM
$C(\cdot, \theta_c)$
L_d
$\frac{\partial L_d}{\partial \theta_c}$
GRL
$-\left(\frac{\partial L_d}{\partial \theta_I} + \frac{\partial L_d}{\partial \theta_T}\right)$

Figure 4.8: The illustration of independent combining information entropy and modality classification into an adversary, which is an intuitive structure of the diagram in Figure 4.1. Other loss functions are kept the same, but we do not show in this graph for simplicity. The gradients computed from the modality classifier in this combining paradigm are used to optimize the parameters $\boldsymbol{\theta}_I$ and $\boldsymbol{\theta}_T$ of the feature extractor.

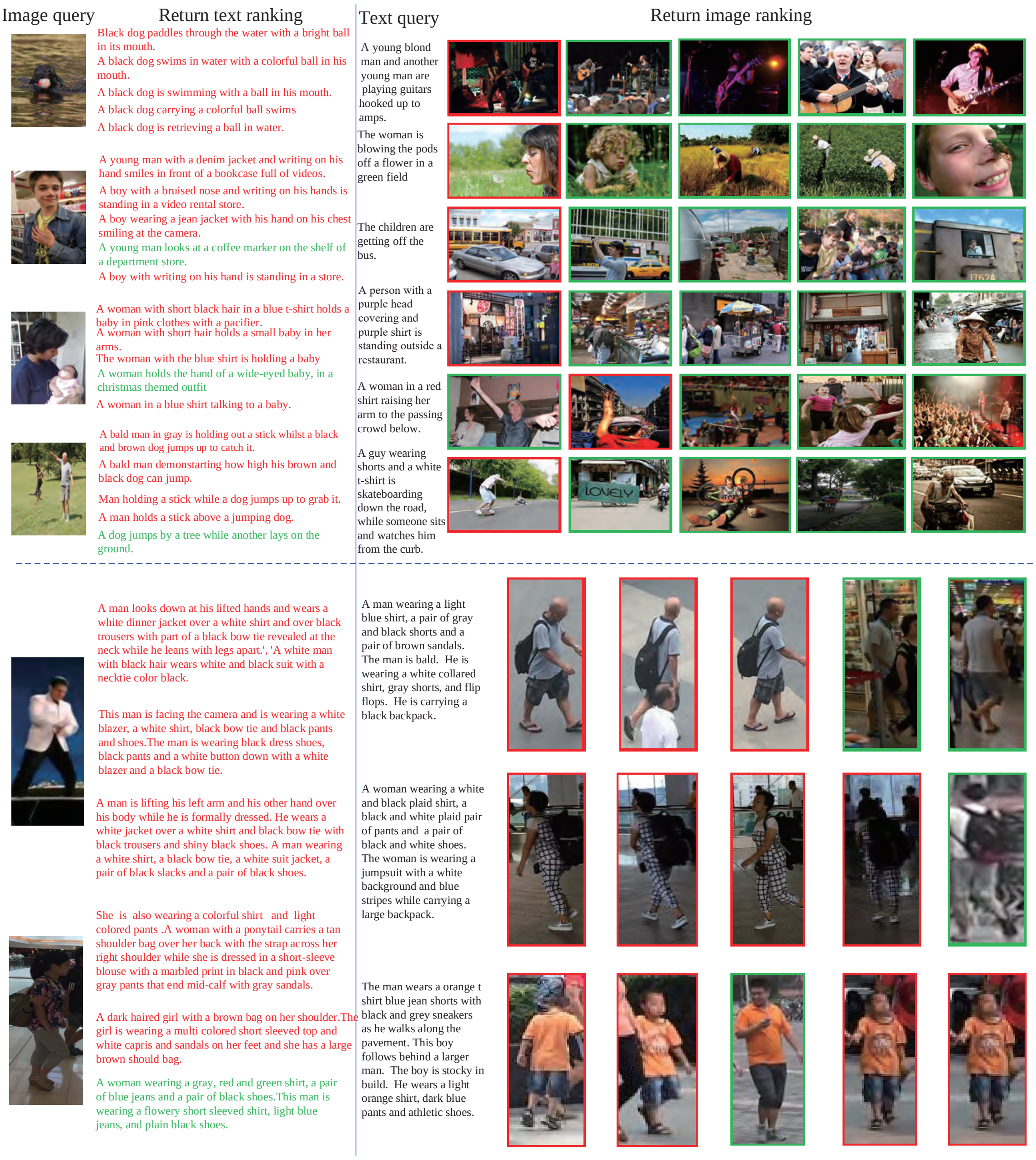
Table 4.8: Comparison of two combining paradigms on four retrieval datasets (R@1, R@10, and mAP(%))

Image-to-Text													
Combining strategy	Backbone Net	Flickr30K			MS-COCO			CUHK-PEDES			Flickr8K		
		R@1	R@10	mAP	R@1	R@10	mAP	R@1	R@10	mAP	R@1	R@10	mAP
Method in Figure 4.8	ResNet-152	55.30	88.30	32.23	57.00	92.10	35.12	67.79	93.75	34.79	39.00	77.70	22.33
Method in Figure 4.3	ResNet-152	56.50	89.60	32.58	58.50	92.10	36.28	65.58	93.60	34.17	39.90	77.90	22.46

4.6 Chapter Conclusions

In this work, we explored methods to improve the performance of cross-modal retrieval by integrating information theory and adversarial learning by analyzing the relation between information entropy and modality uncertainty. Based on this relation, we explored two different paradigms to combine information entropy maximization and modality classification in an adversarial manner. Training these two components iteratively reduces feature distribution discrepancies and further the heterogeneity gap. This is beneficial for preserving semantic similarity between cross-modal features by using bi-directional triplet loss and cross-entropy loss. In addition, we also considered the issue of data imbalance, which leads to a biased classifier and affects label classification. KL-divergence is used as an additional loss term to regularize the re-scaled probabilities computed from image features and text features. It is also used to constrain the cross-modal feature projections and is helpful for learning modality-invariant features. The efficacy of the proposed method was demonstrated by thorough experimental results on four well-known datasets using four deep models.

Successfully combining information entropy and adversarial learning depends on the competitive goals between the information entropy predictor and the modality classifier, and this leads to challenging directions worth further investigation. For example, we used instance labels as supervisory information in this work. Then the information entropy loss was computed only based on image modality and text modality. However, retrieval performance depends on the matching of each image-text feature pair. For some large-scale datasets, each category may include a large number of image-text pairs. Thus, it is valuable to make the information entropy loss specific for each category so that the discrepancy between two modalities can be reduced more granularly. Moreover, the problem of data imbalance leads to training a biased label classifier, which is an issue that can also be resolved by training strategies like data augmentation or by using other loss functions, *e.g.* knowledge distillation loss.



Black dog paddles through the water with a bright ball in its mouth.
A black dog swims in water with a colorful ball in his mouth.
A black dog is swimming with a ball in his mouth.
A black dog carrying a colorful ball swims
A black dog is retrieving a ball in water.

A young man with a denim jacket and writing on his hand smiles in front of a bookcase full of videos.
A boy with a bruised nose and writing on his hands is standing in a video rental store.
A boy wearing a jean jacket with his hand on his chest smiling at the camera.
A young man looks at a coffee marker on the shelf of a department store.
A boy with writing on his hand is standing in a store.

A woman with short black hair in a blue t-shirt holds a baby in pink clothes with a pacifier.
A woman with short hair holds a small baby in her arms.
The woman with the blue shirt is holding a baby
A woman holds the hand of a wide-eyed baby, in a christmas themed outfit
A woman in a blue shirt talking to a baby.

A bald man in gray is holding out a stick whilst a black and brown dog jumps up to catch it.
A bald man demonstarting how high his brown and black dog can jump.
Man holding a stick while a dog jumps up to grab it.
A man holds a stick above a jumping dog.
A dog jumps by a tree while another lays on the ground.

A man looks down at his lifted hands and wears a white dinner jacket over a white shirt and over black trousers with part of a black bow tie revealed at the neck while he leans with legs apart.', 'A white man with black hair wears white and black suit with a necktie color black.

This man is facing the camera and is wearing a white blazer, a white shirt, black bow tie and black pants and shoes.The man is wearing black dress shoes, black pants and a white button down with a white blazer and a black bow tie.

A man is lifting his left arm and his other hand over his body while he is formally dressed. He wears a white jacket over a white shirt and black bow tie with black trousers and shiny black shoes. A man wearing a white shirt, a black bow tie, a white suit jacket, a pair of black slacks and a pair of black shoes.

She is also wearing a colorful shirt and light colored pants .A woman with a ponytail carries a tan shoulder bag over her back with the strap across her right shoulder while she is dressed in a short-sleeve blouse with a marbled print in black and pink over gray pants that end mid-calf with gray sandals.

A dark haired girl with a brown bag on her shoulder.The girl is wearing a multi colored short sleeved top and white capris and sandals on her feet and she has a large brown should bag.

A woman wearing a gray, red and green shirt, a pair of blue jeans and a pair of black shoes.This man is wearing a flowery short sleeved shirt, light blue jeans, and plain black shoes.

A young blond man and another young man are playing guitars hooked up to amps.

The woman is blowing the pods off a flower in a green field

The children are getting off the bus.

A person with a purple head covering and purple shirt is standing outside a restaurant.

A woman in a red shirt raising her arm to the passing crowd below.

A guy wearing shorts and a white t-shirt is skateboarding down the road, while someone sits and watches him from the curb.

A man wearing a light blue shirt, a pair of gray and black shorts and a pair of brown sandals. The man is bald. He is wearing a white collared shirt, gray shorts, and flip flops. He is carrying a black backpack.

A woman wearing a white and black plaid shirt, a black and white plaid pair of pants and a pair of black and white shoes. The woman is wearing a jumpsuit with a white background and blue stripes while carrying a large backpack.

The man wears a orange t shirt blue jean shorts with black and grey sneakers as he walks along the pavement. This boy follows behind a larger man. The boy is stocky in build. He wears a light orange shirt, dark blue pants and athletic shoes.

Figure 4.9: Qualitative test results on the Flikcr30K and CUHK-PEDES datasets. We report Recall@5 of the "Image-to-Text" task and the "Text-to-Image" task from left to right. The correct retrieval images or text are in red and a red box, while the failure retrieval are in green. For Flickr30K, each image is described by 5 sentences. Hence, each text query also has a correct retrieved image, but other retrieved images have similar content as described by the sentence. For the CUHK-PEDES dataset, each category has more than one image, thus almost all correct images are retrieved according to the text query. The list is best viewed in color.