future internet

MDPI

Article

DA-GAN: Dual Attention Generative Adversarial Network for Cross-Modal Retrieval

Liewu Cai, Lei Zhu, Hongyan Zhang and Xinghui Zhu *

College of Information and Intelligence, Hunan Agricultural University, Changsha 410128, China; liewucai@stu.hunau.edu.cn (L.C.); leizhu@hunau.edu.cn (L.Z.); hongyan_zhang@hunau.edu.cn (H.Z.)
* Correspondence: zhuxh@hunau.edu.cn

Abstract: Cross-modal retrieval aims to search samples of one modality via queries of other modalities, which is a hot issue in the community of multimedia. However, two main challenges, i.e., heterogeneity gap and semantic interaction across different modalities, have not been solved efficaciously. Reducing the heterogeneous gap can improve the cross-modal similarity measurement. Meanwhile, modeling cross-modal semantic interaction can capture the semantic correlations more accurately. To this end, this paper presents a novel end-to-end framework, called Dual Attention Generative Adversarial Network (DA-GAN). This technique is an adversarial semantic representation model with a dual attention mechanism, i.e., intra-modal attention and inter-modal attention. Intra-modal attention is used to focus on the important semantic feature within a modality, while inter-modal attention is to explore the semantic interaction between different modalities and then represent the high-level semantic correlation more precisely. A dual adversarial learning strategy is designed to generate modality-invariant representations, which can reduce the cross-modal heterogeneity efficiently. The experiments on three commonly used benchmarks show the better performance of DA-GAN than these competitors.

Keywords: cross-model retrieval; deep representation learning; generative adversarial network; intra-modal attention; inter-modal attention

Citation: Cai, L.; Zhu, L.; Zhang, H.; Zhu, X. DA-GAN: Dual Attention Generative Adversarial Network for Cross-Modal Retrieval. *Future Internet* **2022**, *14*, 43. https://doi.org/10.3390/fi14020043

Academic Editor: Eirini Eleni Tsiropoulou

Received: 4 January 2022
Accepted: 13 January 2022
Published: 27 January 2022

1. Introduction

Cross-modal retrieval [1,2] is a hot issue in the field of multimedia [3]. As shown in Figure 1, it is aiming to find objects of one modality by queries of another modality. Recently, multimedia data [4] is growing exponentially, which is widely used in several scenarios, such as information retrieval, recommendation system [5], social network [6], etc. It makes this problem attract increasing interest by a growing number of researchers.

The main challenge of cross-modal retrieval is how to eliminate the heterogeneity between multimedia objects and how to bridge the semantic gap [7,8] by understanding cross-modal consistent semantic concepts. In the existing literature, the classic way to overcome this challenge is to construct a common latent subspace [9], in which the multimedia instances are represented in the same form and the semantic features can be aligned [10]. As a traditional approach, Canonical Correlation Analysis (CCA) [11] is adopted by many researches [12–15] to learn correlation between cross-modal instances with the same category label. Although these CCA-based methods are supported by classical statistical theory, they cannot represent the complex non-linear semantic correlation. To break this limitation, some non-linear extensions such as KCCA [11], RCCA [16], LPCCA [17], etc. have been proposed to enhance the cross-modal representation.

Thanks to the powerful representation ability of deep learning models [18–21], cross-modal semantic representation learning has been boosted significantly. For instance, several CCA-based approaches, e.g., deep CCA [22], DisDCCA [23], DCCAE [24], are extended by integrating CCA with DNNs. In recent years, attention mechanisms are exploited to support cross-modal feature learning, which is used to discover more significant semantic

details from heterogeneous cross-modal representations. With the help of the attention techniques, high-level semantics can be selectively focused on during the learning, which augments the semantic modeling and reduces the influence of noise on representation learning [25–29].

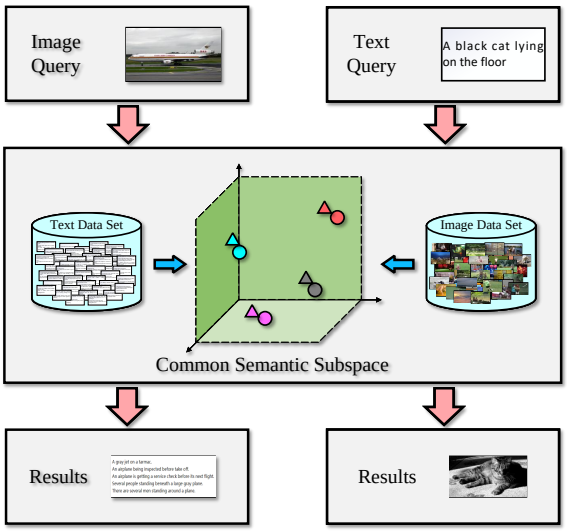


Figure 1. Illustration of cross-modal retrieval.

Our method. To implement the above idea, this paper proposes a new approach, named **D**ual **A**ttention **G**enerative **A**dversarial **N**etwork (**DA-GAN**). This method combines adversarial learning, intra-modal, and inter-modal attention mechanism to improve cross-modal representation capability. Specifically, the inputs are divided into three groups: an image-text pair $\langle \boldsymbol{I}_i, \boldsymbol{T}_i, \boldsymbol{L}_i \rangle$ with category label $\boldsymbol{L}_i$, a group of images and a group of texts with the same label $\boldsymbol{L}_i$. For the generator, we utilize visual CNN and textual CNN to generate visual and textual feature vectors respectively. Then these feature vectors are fed into a two-channel intra-attention model (each channel per modality) to learn intra-modal high-level semantic feature representation with the help of a group of images and texts. At the top of this model, a two-channel encoder is implemented by DNN to learn modality-consistent representations, at the top of which an inter-attention model captures the important semantic features across different modalities. Besides, a two-channel decoder is to re-construct the feature representation for intra-modal adversarial learning. In addition, two types of discriminators are used to form a dual adversarial learning strategy to narrow the heterogeneity gap.

Contributions. This paper has three-fold contributions, which are listed as follows.

- We propose a novel Dual Attention Generative Adversarial Network (DA-GAN) for cross-modal retrieval, which is an integration of the adversarial learning method with a dual attention mechanism.
- To narrow semantic gap and learn high-level semantic features, a dual attention mechanism is designed to capture important semantic features from cross-modal instances in both intra-modal view and inter-modal view, which enhances abstract concepts learning across different modalities.
- To reduce heterogeneity gap, a cross-modal adversarial learning model is employed to learn consistent feature distribution via intra-modal and inter-modal adversarial loss.

Roadmap. The rest of this paper is organized as follows: related works on cross-modal retrieval, attention models, and generative adversarial network are introduced in Section 2. In Section 3, the problem definition and related concepts are proposed. In Section 4, we discuss the details of the proposed DA-GAN. Section 5 presents the experiments and the results. At last, Section 6 concludes this paper.

2. Related Work

2.1. Cross-Modal Retrieval

The main challenge of cross-modal retrieval [30–33] is to diminish the heterogeneity gap and semantics gap by learning a consistent semantic subspace, in which the cross-modal similarity can be directly measured. The existing methods include CCA-based methods, deep learning-based methods, and hashing-based methods. We review them in brief as follows.

CCA-Based Methods. Rasiwasia et al. [34] is the first to use CCA [11] for cross-modal correlation learning. After this work, several CCA-based methods are proposed to enhance cross-modal representation learning. For example, Sharma et al. [14] studied a supervised extension of CCA, which is a general multi-view and kernelizable feature learning method. Pereira et al. [12] proposed three CCA-based approaches, namely correlation matching (CM), semantic matching (SM), and semantic correlation matching (SCM). Gong et al. [13] presented a three-view CCA model in which the abstract semantic information is learned by a third view module to support semantic correlation learning. In [15], cluster-CCA method is developed to generate discriminant cross-modal representations.

Deep Learning-Based Methods. Recently, deep learning [18,19,35] techniques have made great progress, which empowers the multimedia analysis [36–39] and cross-modal representation [40,41]. To learn non-linear correlations from different data modalities, Andrew et al. [42] proposed to integrate deep neural networks into the CCA method. It is a two-channel model, each of which is for one modality. Benton et al. [22] introduced Deep Generalized Canonical Correlation Analysis (DGCCA) to learn non-linear transformations of arbitrarily many views. Gu et al. [43] designed generative processes so as to learn global and local features from cross-modal samples. Zhen et al. [44] introduced a method named Deep Supervised Cross-modal Retrieval (DSCMR) with a weight-sharing strategy to explore the cross-modal consistent relationship.

2.2. Attention Models

Attention mechanism [45] is widely applied in image caption [46], action recognition [47], fine-grained image classification [48], visual questing answering [49], cross-modal retrieval [25] and etc. For example, Wu et al. [50] introduced a deep attention-based spatially recursive model to consider spatial dependencies during feature learning. Sudhakaran et al. [51] proposed Long Short-Term Attention method to capture features from spatial relevant parts across the video frames.

For cross-modal task, Peng et al. [25] proposed a modality-specific cross-modal similarity approach by using a recurrent attention network. Wang et al. [52] designed a hierarchically aligned cross-modal attention (HACA) model to fuse both global and local temporal dynamics of different modalities. Xu et al. [26] developed a Cross-modal Attention with Semantic Consistency (CASC) method to realize local alignment and multi-label prediction for image-text matching. Liu et al. [53] proposed a cross-modal attention-guided erasing approach to comprehend and align cross-modal information for referring expression grounding. Huang et al. [54] used object-oriented encoders along with inter-modal and intra-modal attention networks to improve inter-modal dependencies. Fang et al. [27] introduced subjective attention-based multi-task auxiliary cross-modal fusion method to enhance the robustness and contextual awareness of image fusion.

2.3. Generative Adversarial Network

Generative adversarial network (GAN) is devised by Goodfellow et al. [55], which is a powerful generative model applied in various multimedia tasks [56]. Wang et al. [57] is the first to employ GAN to learn modality-invariant features to diminish cross-modal heterogeneity. Liu et al. [58] presented an adversarial learning-based image-text embedding method to make the distributions of different modalities consistent. Huang et al. [59] studied an adversarial-based transfer model to realize knowledge transfer, and generate modality-indiscriminative representations.

With the support of GAN, many works proposed effective cross-modal hashing methods to realize efficient retrieval in binary Hamming space [60,61]. For example, [62] presented a GAN-based semi-supervised cross-modal hashing approach is presented, which is to learn semantic correlations from unlabeled samples via a minimax game.

3. Preliminaries

In this section, the formal problem definition and related notions are presented. Then, we review the theory of generative adversarial networks, which is the base of the proposed technique. Table 1 summarizes the mathematical notations used in this paper.

Table 1. The mathematical notations.

Notation	Definition
$\mathcal{D}$	a multimedia dataset
$\boldsymbol{I}_i$	the i-th image sample
$\boldsymbol{T}_i$	the i-th text sample
$\boldsymbol{L}_i$	a label vector
Q	a cross-modal query
$\mathcal{R}$	the set of results
$\Phi(\cdot)$	a non-linear mapping
$\mathcal{C}$	the set of semantic concepts
$\boldsymbol{\theta}$	the parameter vector of model
$\boldsymbol{\xi}_I^{(i)}$	the i-th visual convolutional representation
$\boldsymbol{\xi}_T^{(i)}$	the i-th textual convolutional representation
$\boldsymbol{\xi}_I^{\prime(i)}$	the attention-aware representation of image $\boldsymbol{I}_i$
$\boldsymbol{\xi}_T^{\prime(i)}$	the attention-aware representation of text $\boldsymbol{T}_i$
$\boldsymbol{F}_I^{(i)}$	the cross-modal common semantic representation of image $\boldsymbol{I}_i$
$\boldsymbol{F}_T^{(i)}$	the cross-modal common semantic representation of text $\boldsymbol{T}_i$
$\boldsymbol{F}_I^{\prime(i)}$	the attention-aware cross-modal common semantic representation of image $\boldsymbol{I}_i$
$\boldsymbol{F}_T^{\prime(i)}$	the attention-aware cross-modal common semantic representation of text $\boldsymbol{T}_i$
$\boldsymbol{h}$	a hidden vector
$\boldsymbol{K}$	a convolutional kernel
$\boldsymbol{M}$	a semantic correlation matrix
$\boldsymbol{A}$	an attention map
$\boldsymbol{U}$	a cross-modal semantic correlation matrix
$\boldsymbol{\zeta}(i)_I$	a reconstructed representation of i-th image
$\boldsymbol{\zeta}(i)_T$	a reconstructed representation of i-th text

3.1. Problem Definition

This work considers two common modalities: image and text. Let $\mathcal{D} = \{\langle \boldsymbol{I}_i, \boldsymbol{T}_i, \boldsymbol{L}_i \rangle\}_{i=1}^{n}$ be a multimedia dataset that contains n image-text pairs, where $\boldsymbol{I}_i \in \mathbb{R}^{\lambda_I}$ and $\boldsymbol{T}_i \in \mathbb{R}^{\lambda_T}$ represent i-th image sample and text sample in their original space respectively, λ_I and λ_T are the dimensions of image and text original space. Each pair is assigned a semantic label vector that is denoted as $\boldsymbol{L}_i = (L_i^{(1)}, L_i^{(2)}, \ldots, L_i^{(\lambda_L)}) \in \mathbb{R}^{\lambda_L}$, where λ_L is the number of semantic categories in $\mathcal{D}$. If $\boldsymbol{I}_i$ and $\boldsymbol{T}_i$ belong to the same semantic category, then $L_i^{(c)} = 1$; otherwise $L_i^{(j)} = 0$. Cross-modal retrieval aims to search multimedia instances, which are different from the modality of the query Q but similar enough to Q. If the query is an image, denoted as Q_I, we call this type of cross-modal as image-to-text (I2T) retrieval; otherwise text-to-image (T2I) retrieval. In the following the definition of I2T and T2I retrieval are formulated.

Definition 1. ***Cross-Modal Retrieval.*** *Given a multimedia dataset* $\mathcal{D} = \{\langle \boldsymbol{I}_i, \boldsymbol{T}_i, \boldsymbol{L}_i \rangle\}_{i=1}^{n}$ *and two queries* Q_I *and* Q_T*. The I2T retrieval is to return a set of results*

$$\mathcal{R}_{I2T} = \left\{ \boldsymbol{T}_j | Sim(\boldsymbol{T}_i, Q_I) \geq Sim(\boldsymbol{T}', Q_I), \boldsymbol{T}_i \in \mathcal{D}, \boldsymbol{T}' \in \mathcal{D} \setminus \mathcal{R}_{I2T} \right\}_{j=1}^{k}, \tag{1}$$

where $Sim(\cdot)$ *denotes the similarity function, k is the number of results.*

Apparently, Definition 1 indicates that the key problem of cross-modal retrieval is to realize the function $Sim(\cdot)$. However, due to the heterogeneity gap and the semantic gap, it is hard to measure the semantic similarity between instances of different modalities in their original space. Therefore, two non-linear mappings $\Phi_I(\cdot) : \mathbb{R}^{\lambda_I} \mapsto \mathbb{R}^{\lambda_C}$ and $\Phi_T(\cdot) : \mathbb{R}^{\lambda_T} \mapsto \mathbb{R}^{\lambda_C}$ need to be learned, which is to transform images and texts into a λ_C-dimensional common semantic subspace. Thus, the heterogeneity of different modalities can be diminished and the cross-modal representations can be described by a set of semantic concepts $\mathcal{C} = \{C\}_{l=1}^{\lambda_C}$, As a result, the cross-modal similarity can be measured accurately by the following function.

Definition 2. ***Cross-Modal Similarity Function.*** *Given a multimedia dataset* $\mathcal{D}$*, an image* $\boldsymbol{I} \in \mathcal{D}$ *and a text* $\boldsymbol{T} \in \mathcal{D}$*, the cross-modal similarity between* $\boldsymbol{I}$ *and* $\boldsymbol{T}$ *is defined as*

$$Sim(\boldsymbol{I}, \boldsymbol{T}) = \frac{\sum_{i=1}^{\lambda_C} \left(\Phi_I(\boldsymbol{I})^{(i)} \times \Phi_I(\boldsymbol{T})^{(i)} \right)}{\sqrt{\sum_{i=1}^{\lambda_C} \left(\Phi_I(\boldsymbol{I})^{(i)} \right)^2} \times \sqrt{\sum_{i=1}^{\lambda_C} \left(\Phi_I(\boldsymbol{T})^{(i)} \right)^2}}, \tag{2}$$

where $\Phi_I(\boldsymbol{I})$ *and* $\Phi_I(\boldsymbol{T})$ *denote the cross-modal representations in the common semantic subspace.* $\Phi_I(\boldsymbol{I})^{(i)}$ *and* $\Phi_I(\boldsymbol{T})^{(i)}$ *are the i-th element of representation vectors, respectively.*

To learn these two non-linear mappings, we propose a deep architecture by using adversarial learning, which generates modality-invariant representations from multi-modality data and realizes cross-modal semantic augmentation via a dual attention mechanism.

3.2. Review of Generative Adversarial Netw

As a powerful technique, generative adversarial networks (GANs) [55] have be utilized in many multimedia tasks, such as image synthesis, video generation, motion generation, face aging, etc. It consists of two components: a generator $G(\cdot; \boldsymbol{\theta}_G)$ and a discriminator $D(\cdot; \boldsymbol{\theta}_D)$, where $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_D$ are the model parameter vectors. During the training, the generator $G(\cdot; \boldsymbol{\theta}_G)$ tries to make the synthetic image more realistic to fool the discriminator $D(\cdot; \boldsymbol{\theta}_D)$. The discriminator $D(\cdot; \boldsymbol{\theta}_D)$ makes its efforts to distinguish the fake samples from real samples. In other words, $G(\cdot; \boldsymbol{\theta}_G)$ and $D(\cdot; \boldsymbol{\theta}_D)$ are diametrically against to each other.

Specifically, let $\boldsymbol{I}$ be a real image sample obey natural data distribution $P_{data}(\boldsymbol{I})$, $\boldsymbol{z} \in \mathbb{R}^{\lambda_z}$ be a random noise vector generated from distribution $P_z(\boldsymbol{z})$. After fed into the generator $G(\cdot; \boldsymbol{\theta}_G)$, $\boldsymbol{z}$ is transformed into a synthetic sample $G(\boldsymbol{z}; \boldsymbol{\theta}_G)$ that obeys the generative distribution P_G. The discriminator receives the real sample $\boldsymbol{I}$ and the synthetic sample $G(\boldsymbol{z}; \boldsymbol{\theta}_G)$ as inputs, and outputs the discriminant result $D(G(\boldsymbol{z}; \boldsymbol{\theta}_G); \boldsymbol{\theta}_D)$, a probability that $G(\boldsymbol{z}; \boldsymbol{\theta}_G)$ is produced by the generator. This adversarial process can be formulated as

$$\begin{aligned} \arg \min_{G(\cdot;\boldsymbol{\theta}_G)} \max_{D(\cdot;\boldsymbol{\theta}_D)} \mathcal{L}_{GAN}(G(\cdot; \boldsymbol{\theta}_G), D(\cdot; \boldsymbol{\theta}_D)) = \\ \mathbb{E}_{\boldsymbol{I} \sim P_{data}(\boldsymbol{I})} \left[log D(\boldsymbol{I}; \boldsymbol{\theta}_D) \right] + \\ \mathbb{E}_{\boldsymbol{z} \sim P_z(\boldsymbol{z})} \left[log(1 - D(G(\boldsymbol{z}; \boldsymbol{\theta}_G); \boldsymbol{\theta}_D)) \right], \end{aligned} \tag{3}$$

where $\mathbb{E}_{\boldsymbol{I} \sim P_{data}(\boldsymbol{I})}[\cdot]$ and $\mathbb{E}_{\boldsymbol{z} \sim P_z(\boldsymbol{z})}[\cdot]$ denote mathematical expectations:

$$\mathbb{E}_{\boldsymbol{I} \sim P_{data}(\boldsymbol{I})} \left[log D(\boldsymbol{I}; \boldsymbol{\theta}_D) \right] = \int_{\boldsymbol{I}} P_{data}(\boldsymbol{I}) log(D(\boldsymbol{I}; \boldsymbol{\theta}_D)) \, d\boldsymbol{I},$$

$$\mathbb{E}_{z\sim P_z(z)}\left[log(1-D(G(z;\theta_G);\theta_D))\right] = \int_z P_z(z)log(1-D(G(z;\theta_G);\theta_D))\,dz.$$

In the training process, the generator $G(\cdot;\theta_G)$, on one hand, synthesizes images as authentic as possible to fool the discriminator $D(\cdot;\theta_D)$ by minimizing the loss function. On the other hand, the discriminator $(\cdot;\theta_D)$ does its utmost to recognize the fake samples from real samples by maximizing the loss function, shown as follows:

$$\arg\min_{G(\cdot;\theta_G)} \mathcal{L}_{GAN}(G(\cdot;\theta_G), D(\cdot;\theta_D)) = \int_{\boldsymbol{I}} P_{data}(\boldsymbol{I})log(D(\boldsymbol{I};\theta_D))\,d\boldsymbol{I},$$

$$\arg\max_{D(\cdot;\theta_D)} \mathcal{L}_{GAN}(G(\cdot;\theta_G), D(\cdot;\theta_D)) = \int_z P_z(z)log(1-D(G(z;\theta_G);\theta_D))\,dz.$$

4. Methodology

In this section, we discuss the proposed **D**ual **A**ttention **G**enerative **A**dversarial **N**etwork (**DA-GAN**). This method is to learn cross-modal non-linear mappings in an adversarial manner, in which a dual attention mechanism is developed to mine important semantic details to bridge heterogeneity gap and semantic gap. In Section 4.1 we introduce the overview of DA-GAN, and in Sections 4.2 and 4.3 discuss the multi-modal feature learning and adversarial learning with dual attention mechanism. The implementation details are described in Section 4.4.

4.1. Overview of DA-GAN

Figure 2 illustrates the framework of DA-GAN. It consists of three layers: the input layer, generation layer, and discrimination layer.

The Input Layer. The input layer is responsible for training data preparation. To capture more semantic knowledge, two types of samples are selected from the training dataset. The one type is the image-text sample pairs $\{\langle \boldsymbol{I}_i, \boldsymbol{T}_i, \boldsymbol{L}_i\rangle\}_{i=1}^{n}$, and the other type is a group of images $\{\langle \boldsymbol{I}_j, \boldsymbol{L}_i\rangle\}_{j=1}^{m}$ and a group of texts $\{\langle \boldsymbol{T}_j, \boldsymbol{L}_i\rangle\}_{j=1}^{m}$ that have the same semantic label. They are fed into the generation layer to produce the common semantic representations.

The Generation Layer. The generation layer is a deep cross-modal generative model with intra-modal attention (intra-attention) and inter-modal attention (inter-attention). Specifically, the visual and textual features are extracted by a two-channel multi-modal feature learning model $ImgCNN(\cdot;\theta^I_{Fea})$ and $TxtCNN(\cdot;\theta^T_{Fea})$, one channel per modality, where θ^I_{Fea} and θ^T_{Fea} denote parameter vectors. For image modality, it consists of several layers of convolutional networks, which generates visual convolutional representations $\boldsymbol{\xi}_I^{(i)}$ and $\boldsymbol{\xi}_{Is}^{(i)}$ of inputs $\boldsymbol{I}_i$ and $\{\langle \boldsymbol{I}_j\rangle\}_{j=1}^{m}$, respectively. For text modality, the feature learning model consists of a word2vec model to produce word embeddings, and a combination of a bidirectional LSTM [63] (BiLSTM) and a textual CNN to output textual convolutional representations $\boldsymbol{\xi}_T^{(i)}$ and $\boldsymbol{\xi}_{Ts}^{(i)}$. A two-channel intra-attention modal is proposed to capture the important semantic details from each category (each channel per modality). It receives the convolutional representation pair $\langle \boldsymbol{\xi}_I^{(i)}, \bar{\boldsymbol{\xi}}_I^{(i)}\rangle$ and $\langle \boldsymbol{\xi}_T^{(i)}, \bar{\boldsymbol{\xi}}_T^{(i)}\rangle$ and generates the intra-attention masks for both image and text, then outputs the attention-aware representations $\boldsymbol{\xi}_I^{\prime(i)}$ and $\boldsymbol{\xi}_T^{\prime(i)}$. To narrow the heterogeneity gap, a two-channel encoder with weight-sharing strategy over two branches is used following the intra-attention model. Under weight sharing constraint, it generates λ_C-dimensional visual and textual representations $\boldsymbol{F}_I^{(i)} \in \mathbb{R}^{\lambda_C}$ and $\boldsymbol{F}_T^{(i)} \in \mathbb{R}^{\lambda_C}$, which are fed into an inter-attention model to realize cross-modal semantic feature augmentation. Besides, a two-channel decoder (one channel per modality) is employed to reconstruct the image and text representations $\boldsymbol{\zeta}_I^{(i)}$ and $\boldsymbol{\zeta}_T^{(i)}$ from distribution-consistent representations $\boldsymbol{F}_I^{(i)}$ and $\boldsymbol{F}_T^{(i)}$.

The Discrimination Layer. In discrimination layer, there are three types of discriminators, i.e., semantic category discriminator $D_S(\cdot;\theta_S)$, intra-modal discriminator $D_{Intra}(\cdot;\theta_{Intra})$ and inter-modal discriminator $D_{Inter}(\cdot;\theta_{Inter})$, to conduct semantic discrimination, intra-modality

and inter-modality discrimination. $D_S(\cdot;\boldsymbol{\theta}_S)$ and $D_{Intra}(\cdot;\boldsymbol{\theta}_{Intra})$ are two-channel models (one channel per modality). The Former is to predict the semantic labels of convolutional representations $\boldsymbol{\xi}_I'^{(i)}$ and $\boldsymbol{\xi}_T'^{(i)}$, as well as common semantic representations $\boldsymbol{F}_I'^{(i)}$ and $\boldsymbol{F}_T'^{(i)}$ by semantic discrimination loss. The latter is to distinguish the reconstructed representations $\boldsymbol{\zeta}_I^{(i)}$ and $\boldsymbol{\zeta}_T^{(i)}$ from convolutional representations $\boldsymbol{\xi}_I'^{(i)}$ and $\boldsymbol{\xi}_T'^{(i)}$ via intra-modality discrimination loss. The inter-modal discriminator $D_{Inter}(\cdot;\boldsymbol{\theta}_{Inter})$ aims to discriminate the outputs of inter-attention model, i.e., $\boldsymbol{F}_I'^{(i)}$ and $\boldsymbol{F}_T'^{(i)}$ from image and text modality.

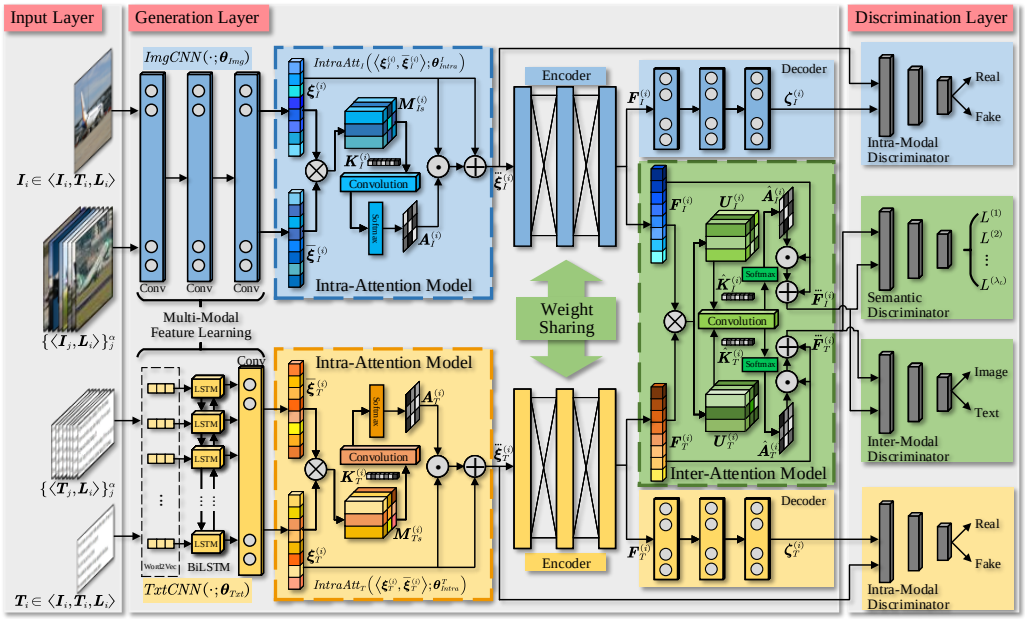


Figure 2. The framework of DA-GAN. The input layer feeds two types of samples into the generation layer: (1) the image-text sample pairs $\{\langle \boldsymbol{I}_i, \boldsymbol{T}_i, \boldsymbol{L}_i\rangle\}_{i=1}^n$ and (2) for each pair, a group of images $\{\langle \boldsymbol{I}_j, \boldsymbol{L}_i\rangle\}_{j=1}^m$ and a group of texts $\{\langle \boldsymbol{T}_j, \boldsymbol{L}_i\rangle\}_{j=1}^m$ that have the same semantic label are selected from multimedia dataset. The generation layer consists of a two-channel CNN-based multi-modal feature learning model, a two-channel intra-attention model, a two-channel encoder, a two-channel decoder, as well as an inter-attention model. The discrimination layer includes: a two-channel intra-modal discriminator to discriminate the convolutional feature representation and common semantic representation, a two-channel semantic discriminator and an inter-modal discriminator to distinguish the common semantic representations of different modalities.

4.2. Multi-Modal Feature Learning

The multi-modal feature learning model consists of two channels: visual feature learning model $ImgCNN(\cdot;\boldsymbol{\theta}_{Fea}^I)$ and textual feature learning model $TxtCNN(\cdot;\boldsymbol{\theta}_{Fea}^T)$ to generate convolutional representations of image and text samples.

4.2.1. Visual Feature Learning

The visual feature learning model is to project visual samples from original data space into convolutional feature space. Formally, $\boldsymbol{\xi}_I^{(i)} = ImgCNN(\boldsymbol{I}_i;\boldsymbol{\theta}_{Fea}^I)$, $\boldsymbol{\xi}_I^{(i)} = (\xi_I^{(i)(1)}, \xi_I^{(i)(2)}, \ldots, \xi_I^{(i)(\gamma)}) \in \mathbb{R}^\gamma$. We use a pre-trained AlexNet [64] to implement visual feature learning. We refine this model on the training dataset via squared loss. Suppose the training set $\mathcal{D} = \{\langle \boldsymbol{I}_i, \boldsymbol{T}_i, \boldsymbol{L}_i\rangle\}_{i=1}^n$ contains n image samples, the ground-true probability vector of i-th sample is denoted as $\boldsymbol{p}'(\boldsymbol{I}_i) = \boldsymbol{L}_i/||\boldsymbol{L}_i||_1$, where $||\cdot||_1$ is the L1 norm. The predictive probability vector is $\boldsymbol{p}(\boldsymbol{I}_i) = (p_i^{(1)}, p_i^{(2)}, \ldots, p_i^{(\lambda_L)})$. Thus, the objective function is

$$\arg\min_{\boldsymbol{\theta}_{Fea}^I} \mathcal{L}_{Fine}\left(\boldsymbol{\theta}_{Fea}^I\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{\lambda_L}\left(p_i^{(j)} - p_i'^{(j)}\right)^2. \tag{4}$$

4.2.2. Textual Feature Learning

The textual feature learning model is a combination of a Word2Vec model, a BiLSTM model and a textual convolutional network [65]. It generates textual convolutional representations, i.e., $\boldsymbol{\xi}_T^{(i)} = TxtCNN(\boldsymbol{T}_i; \boldsymbol{\theta}_{Fea}^T)$, $\boldsymbol{\xi}_T^{(i)} = (\xi_T^{(i)(1)}, \xi_T^{(i)(2)}, \ldots, \xi_T^{(i)(\gamma)}) \in \mathbb{R}^\gamma$. More concretely, a Word2Vec model $Word2Vec(\cdot; \boldsymbol{\theta}_{w2v})$ generates ϵ-dimensional word embedding $\boldsymbol{w}_j \in \mathbb{R}^\epsilon$ for each word in $\boldsymbol{T}_i$. Suppose the length of each text sample $\boldsymbol{T}_i \in \mathcal{D}$ is l (padded if necessary), then the embedding of it is denoted as

$$\boldsymbol{E}_i^{(1,l)} = Word2Vec(\boldsymbol{T}_i; \boldsymbol{\theta}_{w2v}) = \boldsymbol{w}_1 \bowtie \boldsymbol{w}_2 \bowtie \ldots \bowtie \boldsymbol{w}_l, \tag{5}$$

where $\bowtie$ denotes vector concatenation operator. The word embeddings are fed into a BiLSTM model to encode the contextual semantic information from both the previous and future context on forward and reverse direction, $\boldsymbol{h}(t) = BiLSTM(\boldsymbol{E}^{(1,l)}; \boldsymbol{\theta}_{Bi})$, $\boldsymbol{h}(t) \in \mathbb{R}^{\lambda_B}$.

The following textual CNN model receives $\boldsymbol{h}(t)$ at time t and encode local semantic information. Let the convolutional kernels be $\{\boldsymbol{K}_j\}_{j=1}^{\kappa}$ with size $\lambda_B \times m$, for the d-th window of the input vector covered by j-th kernel $\boldsymbol{K}_j$, namely $(\boldsymbol{h}(t), \boldsymbol{h}(t+1), \ldots, \boldsymbol{t}(t+m-1))$, the value of convolution is:

$$\hat{h}_j^{(d)}(t) = \sigma\left(\left(\sum_{i=0}^{m-1} \boldsymbol{h}(t+i-1) * \boldsymbol{K}_j\right) + \beta\right), \tag{6}$$

where $\sigma(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ denotes an activation function, $*$ denotes convolutional operator, and β is a bias term. For j-th kernel, the result of the convolution at each window on vector $\boldsymbol{h}(t)$ is

$$\hat{\boldsymbol{h}}_j(t) = \left(\hat{h}_j^{(1)}(t), \hat{h}_j^{(2)}(t), \ldots, \hat{h}_j^{(l-d+1)}(t)\right). \tag{7}$$

Then, a max pooling operation is conducted on the all the vectors $(\hat{\boldsymbol{h}}_1(t), \hat{\boldsymbol{h}}_2(t), \ldots, \hat{\boldsymbol{h}}_\kappa(t))$ as follows:

$$\begin{aligned}\left(\dot{h}_1(t), \dot{h}_2(t), \ldots, \dot{h}_\kappa(t)\right) &= MaxPooling\left(\hat{\boldsymbol{h}}_1(t), \hat{\boldsymbol{h}}_2(t), \ldots, \hat{\boldsymbol{h}}_\kappa(t)\right) \\ &= \left(max(\hat{\boldsymbol{h}}_1(t)), max(\hat{\boldsymbol{h}}_2(t)), \ldots, max(\hat{\boldsymbol{h}}_\kappa(t))\right),\end{aligned} \tag{8}$$

where $max(\cdot)$ is the function to choose the maximal element of a vector. This κ-dimensional vector is fed into the last FC layer with drop-out to restraint over-fitting:

$$\left(\xi_T^{(i)(1)}, \xi_T^{(i)(2)}, \ldots, \xi_T^{(i)(\gamma)}\right) = \boldsymbol{W}_{fc} \times \left(\dot{h}_1(t), \dot{h}_2(t), \ldots, \dot{h}_\kappa(t)\right) \odot \boldsymbol{\Omega} + \boldsymbol{\beta}, \tag{9}$$

where $\boldsymbol{W}_{fs}$ is the parameters of FC layer, $\boldsymbol{\beta}$ is the bias term, $\odot$ denotes element-wise multiplication operator, and $\boldsymbol{\Omega}$ is a mask to realize drop-out.

4.2.3. Semantic Grouping of Samples

As described in Section 4.1, for each pair $\langle \boldsymbol{I}_i, \boldsymbol{T}_i, \boldsymbol{L}_i \rangle$, the input layer produces a group of images and a group of texts, which belong to the same semantic category to $\langle \boldsymbol{I}_i, \boldsymbol{T}_i, \boldsymbol{L}_i \rangle$. In other words, it randomly samples α images $\{\langle \boldsymbol{I}_j, \boldsymbol{L}_i \rangle\}_{j=1}^{\alpha}$ and texts $\{\langle \boldsymbol{T}_j, \boldsymbol{L}_i \rangle\}_{j=1}^{\alpha}$ according to the semantic label $\boldsymbol{L}_i$ from training set $\mathcal{D}$. After that, these two groups are fed into visual and textual feature learning model, respectively, i.e.,

$$\left\{\boldsymbol{\xi}_I^{(i)(j)}\right\}_{j=1}^{\alpha} = ImgCNN\left(\left\{\boldsymbol{I}_j\right\}_{j=1}^{\alpha}; \boldsymbol{\theta}_{Fea}^I\right), \tag{10}$$

$$\left\{\boldsymbol{\xi}_T^{(i)(j)}\right\}_{j=1}^{\alpha} = TxtCNN\left(\left\{\boldsymbol{T}_j\right\}_{j=1}^{\alpha}; \boldsymbol{\theta}_{Fea}^T\right). \tag{11}$$

The final convolutional representations of the two groups are the average of each representations, i.e.,

$$\bar{\boldsymbol{\xi}}_I^{(i)} = \frac{1}{\alpha}\sum_{j=1}^{\alpha}\boldsymbol{\xi}_I^{(i)(j)}, \quad \bar{\boldsymbol{\xi}}_T^{(i)} = \frac{1}{\alpha}\sum_{j=1}^{\alpha}\boldsymbol{\xi}_T^{(i)(j)}. \tag{12}$$

In this work, $\bar{\boldsymbol{\xi}}_I^{(i)}$ and $\bar{\boldsymbol{\xi}}_T^{(i)}$ are used to represent the common semantic features of the category labeled by $\boldsymbol{L}_i$.

4.3. Adversarial Learning with Dual Attention

In DA-GAN, a novel dual attention mechanism is proposed to learn more discriminative representations via modeling intra-modal and inter-modal semantic correlations by two attention models: intra-attention and inter-attention. Besides, three types of discriminative models are integrated into the framework to achieve modality-invariant representations in an adversarial manner.

4.3.1. Intra-Attention

Intra-Attention model aims to learn more discriminative feature representations by modeling the intra-modal semantic correlations. In our method, it is a two-channel model, one channel per modality. Since the images and texts are processed in the same way, we take the image intra-attention as an example. For the feature representation pair $\langle\boldsymbol{\xi}_I^{(i)}, \bar{\boldsymbol{\xi}}_I^{(i)}\rangle$, $\boldsymbol{\xi}_I^{(i)}, \bar{\boldsymbol{\xi}}_I^{(i)} \in \mathbb{R}^{x\times y\times d}$, x, y, d denote the weight, height and depth of the tensors. For convenience of discussion, we reshape these two tensors as $\boldsymbol{\xi}_I^{(i)} = (\boldsymbol{\xi}_I^{(i)(1)}, \boldsymbol{\xi}_I^{(i)(2)}, \ldots, \boldsymbol{\xi}_I^{(i)(p)})$ and $\bar{\boldsymbol{\xi}}_I^{(i)} = (\bar{\boldsymbol{\xi}}_I^{(i)(1)}, \bar{\boldsymbol{\xi}}_I^{(i)(2)}, \ldots, \bar{\boldsymbol{\xi}}_I^{(i)(p)})$, where $p = x \times y$ is the number of spatial positions of each tensor. The semantic correlation between $\boldsymbol{\xi}_I^{(i)}$ and $\bar{\boldsymbol{\xi}}_I^{(i)}$ can be modeled by the semantic correlation matrix $\boldsymbol{M}_I^{(i)} \in \mathbb{R}^{p\times p}$:

$$\begin{aligned}
\boldsymbol{M}_I^{(i)} &= \begin{pmatrix} \boldsymbol{M}_I^{(1)(1)} & \boldsymbol{M}_I^{(1)(2)} & \ldots & \boldsymbol{M}_I^{(1)(p)} \\ \boldsymbol{M}_I^{(2)(1)} & \boldsymbol{M}_I^{(2)(2)} & \ldots & \boldsymbol{M}_I^{(2)(p)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{M}_I^{(p)(1)} & \boldsymbol{M}_I^{(p)(2)} & \ldots & \boldsymbol{M}_I^{(p)(p)} \end{pmatrix}^{(i)}, \\
\boldsymbol{M}_I^{(j)(k)} &= \boldsymbol{\xi}_I^{(i)} \bigotimes \bar{\boldsymbol{\xi}}_I^{(i)} = \left(\frac{\boldsymbol{\xi}_I^{(i)(j)}}{\left\|\boldsymbol{\xi}_I^{(i)(j)}\right\|_2}\right)^{\top}\left(\frac{\bar{\boldsymbol{\xi}}_I^{(i)(k)}}{\left\|\bar{\boldsymbol{\xi}}_I^{(i)(k)}\right\|_2}\right), \\
& j, k = 1, 2, \ldots, p.
\end{aligned} \tag{13}$$

where $\|\cdot\|_2$ is the L2 norm, notation $\bigotimes$ is called semantic correlation multiplication. Obviously, $\boldsymbol{M}_I^{(i)}$ encode the semantic correlation between the single-sample $\boldsymbol{I}_i$ and the corresponding group $\{\boldsymbol{I}_j\}_{j=1}^{\alpha}$. We reshape it in the following form:

$$\boldsymbol{M}_I^{(i)} = \left(\boldsymbol{m}_I^{(i)(1)}, \boldsymbol{m}_I^{(i)(2)}, \ldots, \boldsymbol{m}_I^{(i)(p)}\right). \tag{14}$$

where $\boldsymbol{m}_I^{(i)(j)} \in \mathbb{R}^p$ is the encoding of semantic correlation between the local single-sample feature representation $\boldsymbol{\xi}_I^{(i)(j)}$ and all the grouping-sample feature representations $\{\bar{\boldsymbol{\xi}}_I^{(i)(k)}\}_{k=1}^{p}$. Therefore, the local semantic correlation between a specific feature representation $\boldsymbol{\xi}_I^{(i)}$ and the average semantic representation $\bar{\boldsymbol{\xi}}_I^{(i)}$ of the corresponding category can be measured directly.

The intra-attention map $\boldsymbol{A}_I^{(i)}$ is generated from the semantic correlation matrices $\boldsymbol{M}_I^{(i)}$ via learning a convolutional operation to fuse the semantic correlations between local single-sample feature vector $\boldsymbol{\xi}_I^{(i)(j)}$ and all the grouping-sample features $\{\bar{\boldsymbol{\xi}}_I^{(i)(k)}\}_{k=1}^{p}$. Specifically,

let $\boldsymbol{K}_I^{(i)} \in \mathbb{R}^{p\times 1}$ be the convolutional kernel, which is learned from the inputs $\langle \boldsymbol{\xi}_I^{(i)}, \bar{\boldsymbol{\xi}}_I^{(i)} \rangle$ by meta learning as follows:

$$\boldsymbol{K}_I^{(i)} = \boldsymbol{W}_2 \times \sigma\left(\boldsymbol{W}_1 \times \left(\frac{1}{p}\sum_{j=1}^{p}\boldsymbol{M}_I^{(1)(j)}, \quad \frac{1}{p}\sum_{j=1}^{p}\boldsymbol{M}_I^{(2)(j)}, \ldots, \frac{1}{p}\sum_{j=1}^{p}\boldsymbol{M}_I^{(p)(j)}\right)^{\top}\right), \tag{15}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ denote the model parameter vectors, $\sigma(\cdot)$ is a non-linear activation function, here we employ ReLU function. Then a softmax operation is conducted on the convolution result to generate intra-attention map $\boldsymbol{A}_I^{(i)} \in \mathbb{R}^{x\times y}$:

$$\boldsymbol{A}_I^{(i)} = \left(\boldsymbol{A}_I^{(i)(1)}, \boldsymbol{A}_I^{(i)(2)}, \boldsymbol{A}_I^{(i)(p)}\right), \quad \boldsymbol{A}_I^{(i)(j)} = \frac{exp\left(\frac{1}{\Gamma}\left(\boldsymbol{K}_I^{(i)}\right)^{\top} \times \boldsymbol{m}_I^{(i)(j)}\right)}{\sum_{j=1}^{p} exp\left(\frac{1}{\Gamma}\left(\boldsymbol{K}_I^{(i)}\right)^{\top} \times \boldsymbol{m}_I^{(i)(j)}\right)}, \tag{16}$$

where Γ is the temperature hyperparameter that influences the entropy. In the same way, the intra-attention map of text modality $\boldsymbol{A}_I^{(i)} \in \mathbb{R}^{x\times y}$ is achieved. Finally, a residual attention mechanism is utilized to calculate the results for both modalities:

$$\begin{aligned} \boldsymbol{\xi}_I'^{(i)} &= \boldsymbol{\xi}_I^{(i)} \bigodot \left(1 + \boldsymbol{A}_I^{(i)}\right), \\ \boldsymbol{\xi}_T'^{(i)} &= \boldsymbol{\xi}_T^{(i);} \bigodot \left(1 + \boldsymbol{A}_T^{(i)}\right), \end{aligned} \tag{17}$$

where $\bigodot$ is the element-wise multiplication.

Following intra-attention model, a two-channel encoder $E(\cdot;\boldsymbol{\theta}_{Enc}^{I})$ and $E(\cdot;\boldsymbol{\theta}_{Enc}^{T})$ is to generate common representations $\boldsymbol{F}_I'^{(i)}$ and $\boldsymbol{F}_T'^{(i)}$. In this model, weight-sharing constraint is applied in last few layers to learn the cross-modal consistent joint distribution, which diminishes heterogeneity effectively.

4.3.2. Inter-Attention

To realize semantic augmentation in the common representation subspace, an inter-attention model is designed to learn the semantic relationship between image and text, i.e.,

$$\langle \ddot{\boldsymbol{F}}_I^{(i)}, \ddot{\boldsymbol{F}}_T^{(i)} \rangle = InterAtt\left(\langle \boldsymbol{F}_I^{(i)}, \boldsymbol{F}_T^{(i)} \rangle; \boldsymbol{\theta}_{Inter}\right). \tag{18}$$

Similar to the intra-attention mechanism, it calculates the cross-modal semantic correlation matrix $\boldsymbol{U}^{(i)}$ from $\boldsymbol{F}_I^{(i)}$ and $\boldsymbol{F}_T^{(i)}$:

$$\boldsymbol{U}^{(j)(k)} = \boldsymbol{F}_I^{(i)} \bigotimes \boldsymbol{F}_T^{(i)} = \left(\frac{\boldsymbol{F}_I^{(i)(j)}}{\left\|\boldsymbol{F}_I^{(i)(j)}\right\|_2}\right)^{\top} \left(\frac{\bar{\boldsymbol{F}}_T^{(i)(k)}}{\left\|\bar{\boldsymbol{F}}_T^{(i)(k)}\right\|_2}\right), \quad j,k = 1,2,\ldots,p. \tag{19}$$

and then generates two correlation matrices:

$$\begin{aligned} \boldsymbol{U}_I^{(i)} &= \boldsymbol{U}^{(i)} = \left(\boldsymbol{u}_I^{(i)(1)}, \boldsymbol{u}_I^{(i)(2)}, \ldots, \boldsymbol{u}_I^{(i)(p)}\right), \\ \boldsymbol{U}_T^{(i)} &= \left(\boldsymbol{U}^{(i)}\right)^{\top} = \left(\boldsymbol{u}_T^{(i)(1)}, \boldsymbol{u}_T^{(i)(2)}, \ldots, \boldsymbol{u}_T^{(i)(p)}\right). \end{aligned} \tag{20}$$

Similar to Equation (14), $\boldsymbol{u}_I^{(i)(j)}, \in \mathbb{R}^p$ encodes semantic correlation between the local image feature vector $\boldsymbol{F}_I^{(i)(j)}$ in j-th position and all the text feature vectors $\{\boldsymbol{F}_T^{(i)(k)}\}_{k=1}^{p}$. $\boldsymbol{u}_T^{(i)(2)} \in \mathbb{R}^p$ encodes the semantic correlation between the local text feature vector $\boldsymbol{F}_T^{(i)(k)}$ in j-th position

and all the image feature vectors $\{\boldsymbol{F}_I^{(i)(j)}\}_{j=1}^{p}$. Then, two convolutional kernels $\hat{\boldsymbol{K}}_I^{(i)}$ and $\hat{\boldsymbol{K}}_T^{(i)}$ are learned in the same way as Equation (15) and the inter-attention maps $\hat{\boldsymbol{A}}_I^{(i)}$ and $\hat{\boldsymbol{A}}_T^{(i)}$ for image and text are achieved by Equation (16). Thus, more discriminative cross-modal representations $\boldsymbol{F}_I^{\prime(i)}$ and $\boldsymbol{F}_T^{\prime(i)}$ can be achieved by residual attention mechanism.

4.3.3. Discriminative Model

Three types of discrimination model are integrated into DA-GAN framework: (1) a semantic discriminator $D_S(\cdot;\boldsymbol{\theta}_S)$ to realize semantic discrimination, (2) a two-channel intra-modal discriminator $D_I(\cdot;\boldsymbol{\theta}_D^I)$ and $D_T(\cdot;\boldsymbol{\theta}_D^I)$ and (3) a two-channel inter-modal discriminator $\hat{D}_I(\cdot;\hat{\boldsymbol{\theta}}_D^I)$ and $\hat{D}_T(\cdot;\hat{\boldsymbol{\theta}}_D^T)$ to realize intra-modal and inter-modal adversarial learning.

Semantic Discriminator. Semantic discriminator $D_S(\cdot;\boldsymbol{\theta}_S)$ is used to recognize the semantic category of the instance in common semantic representation subspace. To this end, a two-channel network with softmax function is added on the top of inter-attention model (one channel per modality), which takes $\boldsymbol{F}_I^{\prime(i)}$ and $\boldsymbol{F}_T^{\prime(i)}$ and inputs and outputs the predicted probability distribution $P_I(\ddot{\boldsymbol{F}}_I^{(i)})$ and $P_T(\ddot{\boldsymbol{F}}_T^{(i)})$ to calculate the semantic discrimination loss:

$$\mathcal{L}_{Sem}(\boldsymbol{\theta}_S) = -\frac{1}{m}\sum_{i=1}^{m}\left[\boldsymbol{L}_i\left(logP_I\left(\ddot{\boldsymbol{F}}_I^{(i)}\right) + logP_I\left(\ddot{\boldsymbol{F}}_T^{(i)}\right)\right)\right], \tag{21}$$

where $\boldsymbol{\theta}_S = (\boldsymbol{\theta}_G^I, \boldsymbol{\theta}_G^T, \boldsymbol{\theta}_C)$ denotes the parameter vector of this model, $\boldsymbol{\theta}_C$ is the parameter vector of the classifier. $\boldsymbol{\theta}_G^I$ and $\boldsymbol{\theta}_G^T$ denote the parameter vector of the image and text generation model respectively, i.e., $\boldsymbol{\theta}_G^I = (\boldsymbol{\theta}_{Fea}^I, \boldsymbol{\theta}_{Intra}^I, \boldsymbol{\theta}_{Enc}^I, \boldsymbol{\theta}_{Inter}^I)$ and $\boldsymbol{\theta}_G^T = (\boldsymbol{\theta}_{Fea}^T, \boldsymbol{\theta}_{Intra}^T, \boldsymbol{\theta}_{Enc}^T, \boldsymbol{\theta}_{Inter}^T)$.

Intra-Model Discriminator. The intra-modal discriminator tries to discriminate the real representations $\boldsymbol{\xi}_I^{\prime(i)}$ ($\boldsymbol{\xi}_T^{\prime(i)}$) from intra-attention model and the synthetic representations $\boldsymbol{\zeta}_I^{(i)}$ ($\boldsymbol{\zeta}_T^{(i)}$) from decoder as inputs. For simplicity, we denote this branch network as $GAN1$, whose objective function is:

$$\begin{aligned}\arg\min_{G_I,G_T}\max_{D_I,D_T}\mathcal{L}_{GAN1}\left(\boldsymbol{\theta}_G^I,\boldsymbol{\theta}_G^T,\boldsymbol{\theta}_D^I,\boldsymbol{\theta}_D^T\right) = {}& \mathbb{E}_{\boldsymbol{I}\sim P_I(\boldsymbol{I})}\left[logD_I\left(\boldsymbol{I};\boldsymbol{\theta}_D^I\right)\right] + \\ & \mathbb{E}_{\boldsymbol{I}\sim P_I(\boldsymbol{I})}\left[log\left(1-D_I\left(G_I\left(\boldsymbol{I};\boldsymbol{\theta}_G^I\right);\boldsymbol{\theta}_D^I\right)\right)\right] + \\ & \mathbb{E}_{\boldsymbol{T}\sim P_T(\boldsymbol{T})}\left[logD_T\left(\boldsymbol{T};\boldsymbol{\theta}_D^T\right)\right] + \\ & \mathbb{E}_{\boldsymbol{T}\sim P_T(\boldsymbol{T})}\left[log\left(1-D_T\left(G_T\left(\boldsymbol{T};\boldsymbol{\theta}_G^T\right);\boldsymbol{\theta}_D^T\right)\right)\right].\end{aligned} \tag{22}$$

Inter-Modal Discriminator. Similar to intra-modal discriminator, the inter-modal discriminator has two channels, the subnetwork for image modality is to recognize the visual common representation as the real sample. By contrast, the subnetwork for text modality aims to recognize the textual common representation as the real sample. This branch of the adversarial network is denoted as $GAN2$. The objective function is:

$$\begin{aligned}\arg\min_{G_I,G_T}\max_{\hat{D}_I,\hat{D}_T}\mathcal{L}_{GAN2}\left(\boldsymbol{\theta}_G^I,\boldsymbol{\theta}_G^T,\hat{\boldsymbol{\theta}}_D^I,\hat{\boldsymbol{\theta}}_D^T\right) = {}& \mathbb{E}_{\boldsymbol{I},\boldsymbol{T}\sim P_I(\boldsymbol{I}),P_T(\boldsymbol{T})}\left[log\hat{D}_I\left(G(\boldsymbol{I};\boldsymbol{\theta}_G^I);\hat{\boldsymbol{\theta}}_D^I\right) - \right. \\ & log\hat{D}_I\left(G\left(\boldsymbol{T};\boldsymbol{\theta}_G^T\right);\hat{\boldsymbol{\theta}}_D^I\right) + \\ & log\hat{D}_T\left(G_T\left(\boldsymbol{T};\boldsymbol{\theta}_G^T\right);\hat{\boldsymbol{\theta}}_D^T\right) - \\ & \left. log\hat{D}_T\left(G\left(\boldsymbol{I};\boldsymbol{\theta}_G^I\right);\hat{\boldsymbol{\theta}}_D^T\right)\right].\end{aligned} \tag{23}$$

4.3.4. Optimization

According to the above discussion, the DA-GAN model can be optimized by the following objective functions:

$$\arg\min_{G_I,G_T}\max_{D_I,D_T}\mathcal{L}_{GAN1}\left(\theta_G^I,\theta_G^T,\theta_D^I,\theta_D^T\right)+\mathcal{L}_{GAN2}\left(\theta_G^I,\theta_G^T,\hat{\theta}_D^I,\hat{\theta}_D^T\right), \tag{24}$$

$$\arg\min_{\theta_S}\mathcal{L}_{Sem}(\theta_S). \tag{25}$$

For discrimination in $GAN1$, the intra-modal discriminator takes the convolutional representation $\xi_I'^{(i)}$ ($\xi_T'^{(i)}$) and the reconstruction representation from $\zeta_I^{(i)}$ ($\zeta_T^{(i)}$) decoder as inputs. It maximizes the log-likelihood for discriminating the real data $\xi_I'^{(i)}$ ($\xi_T'^{(i)}$) and the synthetic data $\zeta_I^{(i)}$ ($\zeta_T^{(i)}$) by stochastic gradient ascending:

$$\theta_D^I\leftarrow\theta_D^I+\eta\nabla_{\theta_D^I}\frac{1}{m}\sum_{i=1}^{n}\left[log\left(D_I\left(\xi_I'^{(i)};\theta_D^I\right)\right)+log\left(1-D_I\left(\zeta_I^{(i)};\theta_D^I\right)\right)\right], \tag{26}$$

$$\theta_D^T\leftarrow\theta_D^T+\eta\nabla_{\theta_D^T}\frac{1}{m}\sum_{i=1}^{n}\left[log\left(D_T\left(\xi_T'^{(i)};\theta_D^T\right)\right)+log\left(1-D_T\left(\zeta_T^{(i)};\theta_D^T\right)\right)\right]. \tag{27}$$

For discrimination in $GAN2$, the subnetwork for image modality receives the image common representation $F_I'^{(i)}$ as the real instance and the text common representation $F_T'^{(i)}$ as the fake instance. The stochastic gradient ascending is calculated as:

$$\hat{\theta}_D^I\leftarrow\hat{\theta}_D^I+\eta\nabla_{\hat{\theta}_D^I}\frac{1}{m}\sum_{i=1}^{n}\left[log\left(\hat{D}_I\left(\ddot{F}_I^{(i)},\xi_I'^{(i)};\hat{\theta}_D^I\right)\right)+log\left(1-\hat{D}_I\left(\ddot{F}_T^{(i)},\xi_I'^{(i)};\hat{\theta}_D^I\right)\right)\right], \tag{28}$$

$$\hat{\theta}_D^T\leftarrow\hat{\theta}_D^T+\eta\nabla_{\hat{\theta}_D^T}\frac{1}{m}\sum_{i=1}^{n}\left[log\left(\hat{D}_T\left(\ddot{F}_T^{(i)},\xi_T'^{(i)};\hat{\theta}_D^T\right)\right)+log\left(1-\hat{D}_T\left(\ddot{F}_T^{(i)},\xi_I'^{(i)};\hat{\theta}_D^T\right)\right)\right]. \tag{29}$$

For the two-channel generative model, it aims to generate more authentic data from the original sample to fit the real semantic distribution by minimizing the objective function. Both of the subnetworks are optimized by stochastic gradient descent (SGD) as follows:

$$\theta_G^I\leftarrow\theta_G^I-\eta\nabla_{\theta_G^I}\frac{1}{m}\sum_{i=1}^{n}\left[log\left(\hat{D}_T\left(\ddot{F}_I^{(i)},\xi_T'^{(i)};\hat{\theta}_D^T\right)\right)+log\left(D_I\left(\zeta_T^{(i)};\theta_D^I\right)\right)\right], \tag{30}$$

$$\theta_G^T\leftarrow\theta_G^T-\eta\nabla_{\theta_G^T}\frac{1}{m}\sum_{i=1}^{n}\left[log\left(\hat{D}_I\left(\ddot{F}_T^{(i)},\xi_I'^{(i)};\hat{\theta}_D^I\right)\right)+log\left(D_T\left(\zeta_T^{(i)};\theta_D^T\right)\right)\right]. \tag{31}$$

Besides, the generative model is optimized by the semantic discrimination to learning abstract semantic concepts:

$$\theta_S\leftarrow\theta_S-\eta\nabla_{\theta_S}\left(-\frac{1}{m}\sum_{i=1}^{m}\left[L_i\left(logP_I\left(\ddot{F}_I^{(i)}\right)+logP_I\left(\ddot{F}_T^{(i)}\right)\right)\right]\right), \tag{32}$$

where η denotes the learning rate, m denotes the number of samples in each mini-batch.

The pseudocode of optimizing the proposed model is shown in Algorithm 1. Before training the $GAN1$ and $GAN2$, we pre-train the multi-modal feature learning model and intra-attention modal for both image and text on the training set, which is to prevent the instability of training $GAN1$ and $GAN2$. The minimax game is implemented by Adam [66].

Algorithm 1 Pseudocode of optimizing DA-GAN

Initialization: a training set $\mathcal{D} = \{\langle \boldsymbol{I}_i, \boldsymbol{T}_i, \boldsymbol{L}_i \rangle\}_{i=1}^{n}$, mini-batch size m, the number of generative model training steps k, learning rate η.
pre-train $ImgCNN(\cdot; \boldsymbol{\theta}^I_{Fea})$ and $IntraAtt_I(\cdot; \boldsymbol{\theta}^I_{Intra})$;
pre-train $TxtCNN(\cdot; \boldsymbol{\theta}^T_{Fea})$ and $IntraAtt_T(\cdot; \boldsymbol{\theta}^T_{Intra})$;
repeat until convergence:
for k steps **do**
    Update the parameters of generator for image $\boldsymbol{\theta}^I_G$ by Equation (30);
    Update the parameters of generator for text $\boldsymbol{\theta}^T_G$ by Equation (31);
    Update the parameters of generators for both image and text $\boldsymbol{\theta}^I_G$ and $\boldsymbol{\theta}^T_G$ by Equation (32);
end for
Update the parameters of intra-modal discriminator $\boldsymbol{\theta}^I_D$ for image by Equation (26);
Update the parameters of intra-modal discriminator $\boldsymbol{\theta}^T_D$ for text by Equation (27);
Update the parameters of inter-modal discriminator for image $\hat{\boldsymbol{\theta}}^I_D$ by Equation (28);
Update the parameters of inter-modal discriminator for text $\hat{\boldsymbol{\theta}}^T_D$ by Equation (29);
Output: the optimized DA-GAN model.

4.4. Implementation Details

Multi-Modal Feature Learning Model. The image feature learning model is implemented by the AlexNet [64] pre-trained on ImageNet dataset. Each input is resized into 256×256 without cropping and 227×227 patches are extracted randomly from the inputs. The 4096-dimensional feature maps from the fc7 layer are treated as the outputs. To improve the learning performance, we fine-tune this model on the training dataset via squared loss. The mini-batch size of 128. the learning rate of the convolutional layer and fully-connected layer are set as 0.001 and 0.002, respectively. The momentum, weight decay, and drop-out rate are set to 0.9, 0.0005, and 0.5, respectively. The convolutional kernel size is set to 3×300, following which is one layer fully-connected network. The drop-out rate is set to 0.5 to avoid over-fitting. The dimension of the last fully-connected layer is set to 4096. The Textual feature learning model includes a pre-trained word2vec model Skip-gram on Wikipedia corpus which contains over 1.8 billion words. This model outputs 300-dimensional word vectors from texts. The textual CNN contains a filter with a size of 3×300. The last fully-connected layer has 4096 dimensions and the learning rate is set to 0.01.

Encoder and Decoder. The two-channel encoder is implemented by a two-layer fully-connected network. For each channel, both of the fc layers are 1024-dimensional, and the weights of the second layer are shared over two branches to model the cross-modal joint distributions. Each branch of the decoder has two layers of fully-connected networks. The dimension of these two layers are 1024 and 4096, respectively.

Intra-modal and Inter-modal Discriminator. For intra-modal discriminator, each branch of it is constructed by one FC layer. To discriminate the convolutional representations and the reconstructed representations, the former is labeled by tag 1, and the latter is labeled by tag 0. For the inter-modal discriminator, both of the two channels are two-layer fully-connected networks. The 1ts layer has 1024 dimensions, and the 2nd layer with a sigmoid activation function calculates the predicted score for each input representation. The common representations of image modality are labeled by 1 and the representations of text modality are labeled by 0. For text modality, these two types of representations are labeled in the opposite way.

5. Experiments

5.1. Datasets

All the experiments are conducted on three widely-used benchmark datasets: Wikipedia [34], NUS-WIDE [67] and Pascal Sentences [68]. Some image and text samples of these three datasets are shown in Figure 3.

Wikipedia NUS-WIDE Pascal Sentence

Figure 3. Some image and text samples of Wikipedia, NUS-WIDE and Pascal Sentences.

5.2. Competitors

We compare the proposed DA-GAN with 13 competitors, including 6 traditional cross-modal retrieval approaches, i.e., CCA [69], KCCA [11], MCCA [70], MvDA [71], MvDA-VC [72] and JRL [73], as well as 7 deep learning-based approaches, i.e., DCCA [42], DCCAE [24], CCL [74], CMDN [75], ACMR [57], DSCMR [44], CM-GANs [76]. The brief introductions of them are listed here.

- **CCA** [69] is a statistical method that is to learn linear correlations between samples of different modalities.
- **KCCA** [11] is a non-linear extension of CCA, which employs kernel function to improve the performance of common subspace learning.
- **MCCA** [70] is a generalization of CCA to more than two views, which is used to recognize similar patterns across multiple domains.
- **MvDA** [71] jointly learns multiple view-specific linear transforms so as to construct a common subspace for multiple views.
- **MvDA-VC** [72] is an extension of MvDA with with view consistency, which utilize the structure similarity of views corresponding to the same object.
- **JRL** [73] uses sparse projection matrix and semi-supervised regularization to explore correlations of labeled and unlabeled cross-modal samples.
- **DCCA** [42] is implemented by deep neural networks to learn non-linear correlation. It has two separated DNNs, one branch per modality.
- **DCCAE** [24] is a DCCA extension that integrates CCA model and autoencoder-based model to realize multi-view representation learning.
- **CCL** [74] realizes a hierarchical network to combine multi-grained fusion and cross-modal correlation exploiting. It includes two learning stages to realize representation learning and intrinsic relevance exploiting.
- **CMDN** [75] contains two learning stages to model the complementary separate representation of different modalities, and combines cross-modal representations to generate rich cross-media correlation.
- **ACMR** [57] is a adversarial learning-based method to construct a common subspace for different modalities by generating modality-invariant representations.
- **DSCMR** [44] exploits semantic discriminative features from both label space and common representation space by supervised learning, and minimizes modality invariance loss via weight-sharing to generate modality-invariant representation.

- **CM-GANs** [76] models cross-modal joint distributions by two parallel GANs to generate modality-invariance representations

5.3. Performance Metrics

Two tasks are considered, i.e., (1) I2T retrieval and (2) T2I retrieval, both of which are defined in Definition 1. Besides, we utilize PR-curves and mAP score to measure the retrieval performance:

$$Pr = \frac{TP}{TP + FP}, \quad Re = \frac{TP}{TP + FN}, \tag{33}$$

$$mAP = \frac{1}{|\{Q\}|} \sum_{i=1}^{|\{Q\}|} AP(Q) \tag{34}$$

5.4. Experimental Results

5.4.1. Results on Wikipedia Dataset

The mAP scores of DA-GAN and the 13 competitors on the Wikipedia dataset are reported in Table 2. For both I2T and T2I tasks, the proposed DA-GAN outperforms all these state-of-the-arts by 54.3% and 63.9% respectively, higher than the two best competitors, i.e., DSCMR [44] (I2T mAP = 52.1%) and CM-GANs [76] (T2I mAP = 62.1%). Besides, the average mAP of DA-GAN is the highest, which is 3% higher than CM-GANs. The main reason is that the combination of intra- and inter-modal attention captures more single-modal and cross-modal semantic correlations. Although both DSCMR and CM-GANs extract the semantic information by supervised learning, they do not learn the inter-modal semantic correlation effectively to realize cross-modal semantic augmentation. On the other hand, except for DCCA and DCCAE whose mAPs (I2T mAP = 44.4% and 43.5%, T2I mAP = 39.6% and 38.5%) are a bit lower than JRL (I2T mAP = 44.9%, T2I mAP = 41.8%).

Table 2. The comparison results (mAP@50 in %) with 13 competitors on Wikipedia dataset. The best performance values are in bold-font.

Traditional Method	**I2T**	**T2I**	**Aver.**
CCA [69]	13.4	13.3	13.4
KCCA [11]	19.8	18.6	19.2
MCCA [70]	34.1	30.7	32.4
MvDA [71]	33.7	30.8	32.3
MvDA-VC [72]	38.8	35.8	37.3
JRL [73]	44.9	41.8	43.4
Deep Learning-Based Method	**I2T**	**T2I**	**Aver.**
DCCA [42]	44.4	39.6	42.0
DCCAE [24]	43.5	38.5	41.0
CCL [74]	50.4	45.7	48.1
CMDN [75]	48.7	42.7	45.7
ACMR [57]	47.7	43.4	45.6
DSCMR [44]	52.1	47.8	49.9
CM-GANs [76]	50.0	62.1	56.1
The Proposed Method	**I2T**	**T2I**	**Aver.**
DA-GAN	**54.3**	**63.9**	**59.1**

Figures 4 and 5 illustrate the I2T and T2I mAP scores of each category on Wikipedia dataset. The average mAP scores are shown in Figure 6. Obviously, for all these approaches, there are big differences between the retrieval precisions of different categories. Specifically, for both I2T and T2I tasks, the performances on "biology", "geography & places", "sport & recreation" and "warfare" are better than other categories. That is mainly because the samples in the above categories are semantically independent of other categories, and have more obvious distinguishing features than other categories. In contrast, the categories "art & architecture", "history" and "royalty & nobility" are relative to each other in abstract semantics. The samples of the categories have more confusing features.

From Figures 4 and 5, it is clear that DA-GAN has better semantic recognition ability. For example, the highest I2T and T2I mAP scores of DA-GAN on "biology", "sport & recreation" and "warfare" are near 83% and 85%, higher than the competitive rivals such as DSCMR (I2T mAP = 78%, T2I mAP = 73%), CCL (I2T mAP = 73%, T2I mAP = 69%) and CM-GANs (I2T mAP = 74%, T2I mAP = 82%).

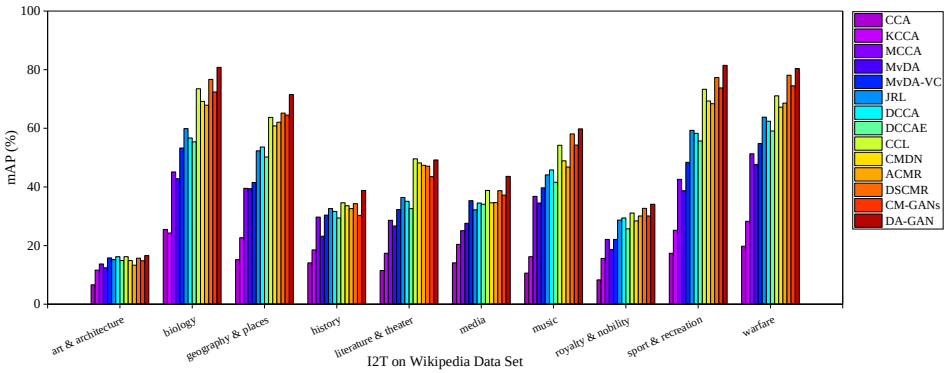


Figure 4. The mAP of I2T task of each category on Wikipedia dataset for our method DA-GAN and the competitors.

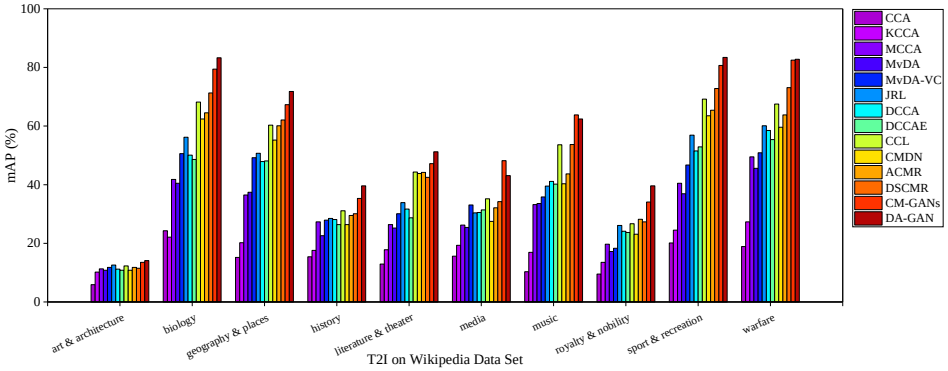


Figure 5. The mAP of T2I task of each category on Wikipedia dataset for our method DA-GAN and the competitors.

Figure 7a,d show the I2T and T2I precisions of DA-GAN and the competitors on different recalls, respectively. For both I2T and T2I tasks, DA-GAN has the highest precision at all levels of recall, which exhibits the performance improvement by adversarial learning with a dual attention mechanism. DSCMR and CM-GANs are still the most two competitive rivals, but they cannot defeat DA-GAN at any recall value.

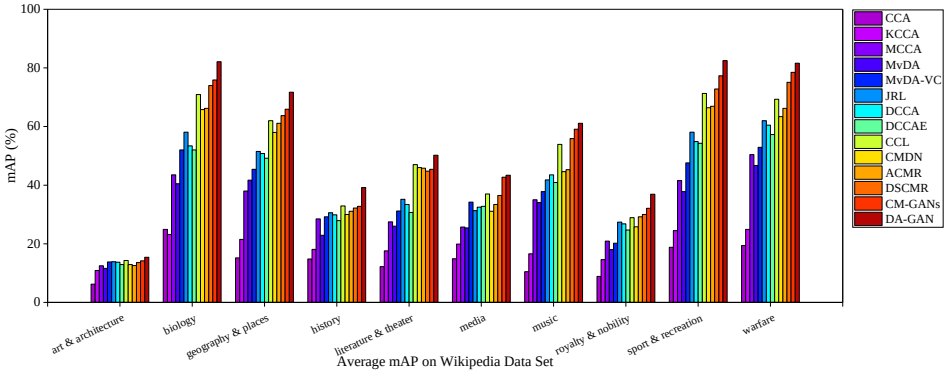
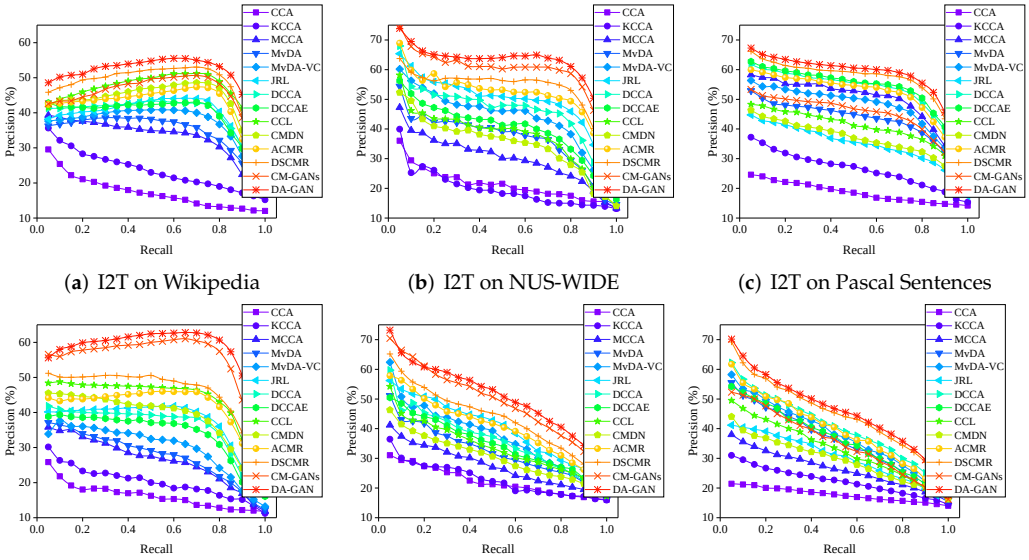


Figure 6. The average mAP of each category on Wikipedia dataset for our method DA-GAN and the competitors.



(**a**) I2T on Wikipedia (**b**) I2T on NUS-WIDE (**c**) I2T on Pascal Sentences

(**d**) T2I on Wikipedia (**e**) T2I on NUS-WIDE (**f**) T2I on Pascal Sentences

Figure 7. The PR curve of our method DA-GAN and the competitors on Wikipedia, NUS-WIDE and Pascal Sentences dataset. (**a**–**c**) show the PR curves of I2T task on Wikipedia, NUS-WIDE and Pascal Sentences; (**d**–**f**) report PR curves of I2T task.

5.4.2. Results on Nus-Wide Dataset

The mAP scores on NUS-WIDE of DA-GAN and competitors are reported in Table 3. Compared with the results on Wikipedia, the precision of all these methods are relatively higher. The proposed method performs well on this dataset, which defeats CM-GANs (I2T mAP = 78.1%, T2I mAP = 72.4%, Aver. mAP = 75.3%) and DSCMR (I2T mAP = 61.1%, T2I mAP = 61.5%, Aver. mAP = 61.3%) by I2T mAP = 79.7%, T2I mAP = 75.2%, Aver. mAP = 77.5%. It indicates that the dual attention mechanism can discover more important semantic features between different modalities to generate more discriminant representations. On the other hand, we observe that the performance of other traditional and deep learning-based approaches are far behind our method even though the precisions of them are obviously higher than the results on Wikipedia.

Table 3. The comparison results (mAP@50 in %) with 13 competitors on NUS-WIDE dataset. The best performance values are in bold-font.

Traditional Method	**I2T**	**T2I**	**Aver.**
CCA [69]	37.8	39.4	38.6
KCCA [11]	36.2	39.4	37.8
MCCA [70]	44.8	46.2	45.5
MvDA [71]	50.1	52.6	51.3
MvDA-VC [72]	52.6	55.7	54.2
JRL [73]	58.6	59.8	59.2
Deep Learning-Based Method	**I2T**	**T2I**	**Aver.**
DCCA [42]	53.2	54.9	54.0
DCCAE [24]	51.1	54.0	52.5
CCL [74]	50.6	53.5	52.1
CMDN [75]	49.2	51.5	50.4
ACMR [57]	58.8	59.9	59.3
DSCMR [44]	61.1	61.5	61.3
CM-GANs [76]	78.1	72.4	75.3
The Proposed Method	**I2T**	**T2I**	**Aver.**
DA-GAN	**79.7**	**75.2**	**77.5**

The PR curve of DA-GAN and the state-of-the-arts are presented in Figure 7b,c. We can find that the trends of the precisions on NUS-WIDE are different from the situations

on Wikipedia. For the I2T task (shown in Figure 7b), the precision of DA-GAN and the competitors decline obviously in the interval $[0.0, 0.2]$. After that, the downward trend tends to be gentle. When the recall is larger than 0.8, fast performance degradation occurs, except for three traditional methods, i.e., CCA, KCCA, and MCCA. At all levels of recall, the precision of DA-GAN is higher than all the rivals. For the T2I task (shown in Figure 7c), the performance of all these approaches shows a gradual downward trend. Although the precision of CM-GANs is slightly higher than our method in the interval $[0.1, 0.2]$, it cannot defeat DA-GAN when the recall is larger than 0.2. The retrieval accuracies of other approaches, as expected, are much lower than DA-GAN.

5.4.3. Results on Pascal Sentences Dataset

The Comparison of mAP scores of DA-GAN and the 13 state-of-the-arts on Pascal Sentences dataset are shown in Table 4. Once again, DA-GAN is the winner in this contest, which achieves I2T mAP = 72.9%, T2I mAP = 73.5% and average mAP = 73.2%, defeats the runner-up DSCMR (I2T mAP = 71.0%, T2I mAP = 72.2%, average mAP = 71.6%) by 2.5%, 1.3% and 1.6%, respectively. Different from the above comparisons, CM-GANs (I2T mAP = 61.2%, T2I mAP = 61.0%, average mAP = 61.1%) performs worse than DA-GAN and DSCMR evidently. As analyzed above, the performance improvement mainly comes from the integration of intra- and inter-modal attention as well as adversarial learning.

Table 4. The comparison results (mAP@50 in %) with 13 competitors on Pascal Sentences dataset. The best performance values are in bold-font.

Traditional Method	**I2T**	**T2I**	**Aver.**
CCA [69]	22.5	22.7	22.6
KCCA [11]	43.3	39.8	41.6
MCCA [70]	66.4	48.9	55.45
MvDA [71]	59.4	62.6	61.0
MvDA-VC [72]	64.8	67.3	66.1
JRL [73]	52.7	53.4	53.1
Deep Learning-Based Method	**I2T**	**T2I**	**Aver.**
DCCA [42]	67.8	67.7	67.8
DCCAE [24]	68.0	67.1	67.5
CCL [74]	57.6	56.1	56.9
CMDN [75]	54.4	52.6	53.5
ACMR [57]	67.1	67.6	67.3
DSCMR [44]	71.0	72.2	71.6
CM-GANs [76]	61.2	61.0	61.1
The Proposed Method	**I2T**	**T2I**	**Aver.**
DA-GAN	**72.9**	**73.5**	**73.2**

Figures 8–10 illustrate the I2T, T2I and average mAP scores of each approaches on 20 categories on Pascal Sentences dataset, respectively. For both I2T and T2I tasks, all these approaches have poor cross-modal retrieval performance in some categories, such as "bottle" and "chair". It is mainly because the objects in these categories are relatively small. By contrast, the precisions on "aeroplane", "bird", "cat", "horse", "motorbike", "sheep" and "train" are obviously higher since these samples contain much more discriminative semantic features. Specifically, for the I2T task, the mAP of DA-GAN reaches nearly 90%, 91% and 92% on "aeroplane", "cat" and "train", respectively. For the T2I task, it achieves nearly 92%, 93%, and 95% on these three categories. From Figure 10 we observe that the semantic recognition performance of DA-GAN is the best among these 14 approaches.

Figure 7c,f show the PR curves of DA-GAN and 13 state-of-the-arts for I2T and T2I tasks, respectively. On both tasks, it is clear that the changing of performance of DA-GAN and CM-GANs are very similar. Although CM-GANs show good performance, they cannot

overcome our method. For the I2T task, the precision of DA-GAN declines slowly when the recall increases from 0.2 to 0.8. After that, it drops sharply. In contrast, the performance of our method shows a significant downward trend for the T2I task, but it is still the best.

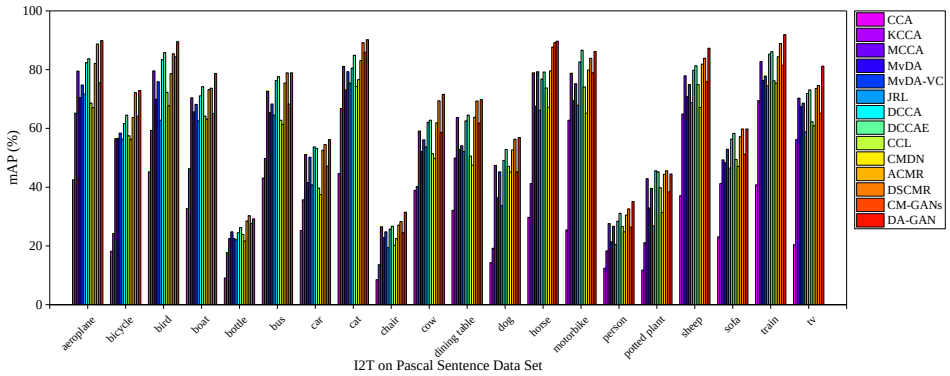


Figure 8. The mAP of I2T task of each category on Pascal Sentences dataset for our method DA-GAN and the competitors.

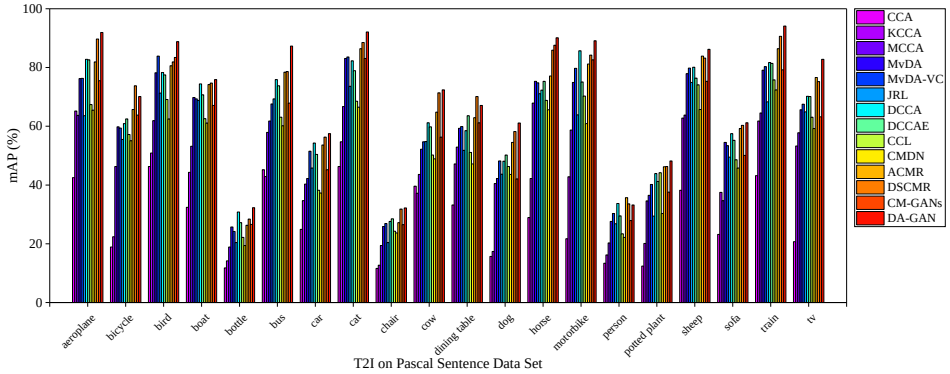


Figure 9. The mAP of T2I task of each category on Pascal Sentences dataset for our method DA-GAN and the competitors.

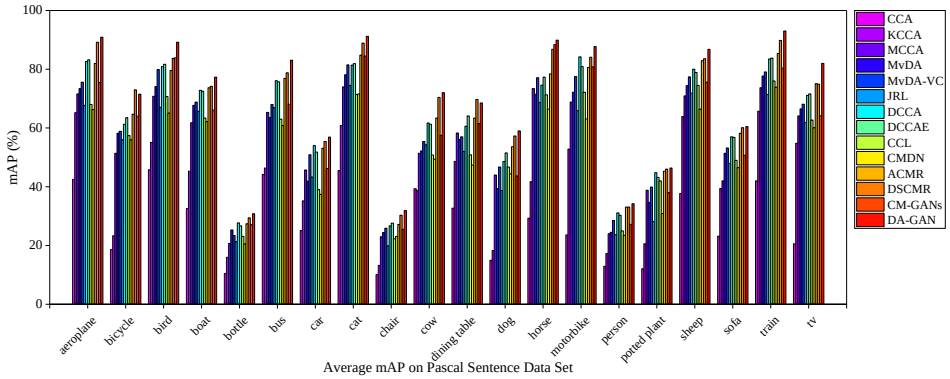


Figure 10. The average mAP of each category on Pascal Sentences dataset for the proposed method and the state-of-the-arts.

6. Conclusions

We present a new deep adversarial model for cross-modal retrieval, called Dual Attention Generative Adversarial Network (DA-GAN). This method utilizes a novel dual attention mechanism to focus on important semantic details in a uni-modal manner and a cross-modal manner, which can effectively learn high-level semantic interaction across different modalities. Besides, a dual adversarial learning method that learns modality-consistent representation is proposed to reduce the heterogeneity gap. Comprehensive experiments on four commonly used multimedia datasets indicate the great performance of the proposed method.

Author Contributions: Conceptualization, L.C. and L.Z.; methodology, L.C. and H.Z.; software, L.C. and L.Z.; validation, L.C. and X.Z.; formal analysis, H.Z.; investigation, X.Z. and L.C.; resources, H.Z. and X.Z.; data curation, L.Z.; writing—original draft preparation, L.C.; writing—review and editing, X.Z.; visualization, L.Z.; supervision, X.Z. and H.Z.; project administration, X.Z.; funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Research and Development Program of Hunan Province (2020NK2033), and the National Natural Science Foundation of China (62072166).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: This work is supported in the Key Research and Development Program of Hunan Province (2020NK2033), and the National Natural Science Foundation of China (62072166).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, Y. Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion. *ACM Trans. Multimed. Comput. Commun. Appl.* **2021**, *17*, 1–25. [CrossRef]
2. Ranjan, V.; Rasiwasia, N.; Jawahar, C.V. Multi-label cross-modal retrieval. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4094–4102.
3. Chen, Y.; Ren, P.; Wang, Y.; de Rijke, M. Bayesian personalized feature interaction selection for factorization machines. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 665–674.
4. Wu, Y.; Yang, Y. Exploring Heterogeneous Clues for Weakly-Supervised Audio-Visual Video Parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1326–1335.
5. Chen, Y.; Wang, Y.; Ren, P.; Wang, M.; de Rijke, M. Bayesian feature interaction selection for factorization machines. *Artif. Intell.* **2022**, *302*, 103589. [CrossRef]
6. Zhang, C.; Wang, Y.; Zhu, L.; Song, J.; Yin, H. Multi-graph heterogeneous interaction fusion for social recommendation. *ACM Trans. Inf. Syst.* **2021**, *40*, 1–26. [CrossRef]
7. Gu, C.; Bu, J.; Zhou, X.; Yao, C.; Ma, D.; Yu, Z.; Yan, X. Cross-modal Image Retrieval with Deep Mutual Information Maximization. *arXiv* **2021**, arXiv:2103.06032.
8. Zhang, C.; Song, J.; Zhu, X.; Zhu, L.; Zhang, S. Hcmsl: Hybrid cross-modal similarity learning for cross-modal retrieval. *ACM Trans. Multimed. Comput. Commun. Appl.* **2021**, *17*, 1–22. [CrossRef]
9. Zhang, C.; Zhong, Z.; Zhu, L.; Zhang, S.; Cao, D.; Zhang, J. M2guda: Multi-metrics graph-based unsupervised domain adaptation for cross-modal Hashing. In Proceedings of the 2021 International Conference on Multimedia Retrieval, Taipei, Taiwan, 21–24 August 2021; pp. 674–681.
10. Thomas, C.; Kovashka, A. Preserving semantic neighborhoods for robust cross-modal retrieval. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 317–335.
11. Hardoon, D.R.; Szedmák, S.; Shawe-Taylor, J. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.* **2004**, *16*, 2639–2664. [CrossRef] [PubMed]
12. Pereira, J.C.; Coviello, E.; Doyle, G.; Rasiwasia, N.; Lanckriet, G.R.G.; Levy, R.; Vasconcelos, N. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 521–535. [CrossRef]
13. Gong, Y.; Ke, Q.; Isard, M.; Lazebnik, S. A multi-view embedding space for modeling internet images, tags, and their semantics. *Int. Comput. Vis.* **2014**, *106*, 210–233. [CrossRef]
14. Sharma, A.; Kumar, A.; Daume, H.; Jacobs, D.W. Generalized multiview analysis: A discriminative latent space. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2160–2167.
15. Rasiwasia, N.; Mahajan, D.; Mahadevan, V.; Aggarwal, G. Cluster canonical correlation analysis. In Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, 22–25 April 2014.
16. Lopez-Paz, D.; Sra, S.; Smola, A.; Ghahramani, Z.; Schölkopf, B. Randomized nonlinear component analysis. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1359–1367.
17. Sun, T.; Chen, S. Locality preserving cca with applications to data visualization and pose estimation. *Image Vis. Comput.* **2007**, *25*, 531–543. [CrossRef]
18. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
19. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
20. Wang, Y.; Lin, X.; Wu, L.; Zhang, W. Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval. *IEEE Trans. Image Process.* **2017**, *26*, 1393–1404. [CrossRef]
21. Qian, B.; Wang, Y.; Hong, R.; Wang, M.; Shao, L. Diversifying inference path selection: Moving-mobile-network for landmark recognition. *IEEE Trans. Image Process.* **2021**, *30*, 4894–4904. [CrossRef]

22. Benton, A.; Khayrallah, H.; Gujral, B.; Reisinger, D.; Zhang, S.; Arora, R. Deep generalized canonical correlation analysis. *arXiv* **2017**, arXiv:1702.02519.
23. Elmadany, N.E.D.; He, Y.; Guan, L. Multiview learning via deep discriminative canonical correlation analysis. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2409–2413.
24. Wang, W.; Arora, R.; Livescu, K.; Bilmes, J.A. On deep multi-view representation learning. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; Volume 37, pp. 1083–1092.
25. Peng, Y.; Qi, J.; Yuan, Y. Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Trans. Image Process.* **2018**, *27*, 5585–5599. [CrossRef]
26. Xu, X.; Wang, T.; Yang, Y.; Zuo, L.; Shen, H.T. Cross-modal attention with semantic consistence for image-text matching. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 5412–5425. [CrossRef]
27. Fang, A.; Zhao, X.; Zhang, Y. Cross-modal image fusion theory guided by subjective visual attention. *arXiv* **2019**, arXiv:1912.10718.
28. Zhu, L.; Zhang, C.; Song, J.; Liu, L.; Zhang, S.; Li, Y. Multi-graph based hierarchical semantic fusion for cross-modal representation. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
29. Wu, L.; Wang, Y.; Gao, J.; Wang, M.; Zha, Z.-J.; Tao, D. Deep coattention-based comparator for relative representation learning in person re-identification. *IEEE Trans. Neural Netw. Learn.* **2020**, *32*, 722–735. [CrossRef]
30. Wang, K.; He, R.; Wang, L.; Wang, W.; Tan, T. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 2010–2023. [CrossRef]
31. Zhu, L.; Long, J.; Zhang, C.; Yu, W.; Yuan, X.; Sun, L. An efficient approach for geo-multimedia cross-modal retrieval. *IEEE Access* **2019**, *7*, 180571–180589. [CrossRef]
32. Zhu, L.; Song, J.; Zhu, X.; Zhang, C.; Zhang, S.; Yuan, X. Adversarial learning-based semantic correlation representation for cross-modal retrieval. *IEEE Multimed.* **2020**, *27*, 79–90. [CrossRef]
33. Wang, C.; Yang, H.; Meinel, C. Deep semantic mapping for cross-modal retrieval. In Proceedings of the 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI), Vietri sul Mare, Italy, 9–11 November 2015; pp. 234–241.
34. Rasiwasia, N.; Pereira, J.C.; Coviello, E.; Doyle, G.; Lanckriet, G.R.; Levy, R.; Vasconcelos, N. A new approach to cross-modal multimedia Retrieval. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 251–260.
35. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011.
36. Wu, L.; Wang, Y.; Shao, L. Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Trans. Image Process.* **2018**, *28*, 1602–1612. [CrossRef]
37. Zhao, L.; Chen, Z.; Yang, L.T.; Deen, M.J.; Wang, Z.J. Deep semantic mapping for heterogeneous multimedia transfer learning using co-occurrence data. *Acm Trans. Multimed. Comput. Commun. Appl.* **2019**, *15*, 1–21. [CrossRef]
38. Wang, Y.; Zhang, W.; Wu, L.; Lin, X.; Fang, M.; Pan, S. Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016; pp. 2153–2159.
39. Zhang, W.; Yao, T.; Zhu, S.; Saddik, A.E. Deep learning–based multimedia analytics: A review. *ACM Trans. Multimed. Comput. Appl.* **2019**, *15*, 1–26. [CrossRef]
40. Wei, Y.; Zhao, Y.; Lu, C.; Wei, S.; Liu, L.; Zhu, Z.; Yan, S. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE Trans. Cybern.* **2016**, *47*, 449–460. [CrossRef]
41. Zhu, L.; Song, J.; Wei, X.; Yu, H.; Long, J. Caesar: Concept augmentation based semantic representation for cross-modal retrieval. *Multimed. Tools Appl.* **2020**, *1*, 1–31. [CrossRef]
42. Andrew, G.; Arora, R.; Bilmes, J.A.; Livescu, K. Deep canonical correlation Analysis. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; Volume 28, pp. 1247–1255.
43. Gu, J.; Cai, J.; Joty, S.R.; Niu, L.; Wang, G. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7181–7189.
44. Zhen, L.; Hu, P.; Wang, X.; Peng, D. Deep supervised cross-modal Retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10394–10403.
45. Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S.C.H.; Wang, X.; Li, H. Dynamic fusion with intra- and inter-modality attention flow for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 6639–6648.
46. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
47. Liu, J.; Wang, G.; Hu, P.; Duan, L.-Y.; Kot, A.C. Global context-aware attention lstm networks for 3d action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1647–1656.

48. Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; Zhang, Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 842–850.
49. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems*; 2016; pp. 289–297.
50. Wu, L.; Wang, Y.; Li, X.; Gao, J. Deep attention-based spatially recursive networks for fine-grained visual recognition. *IEEE Trans. Cybern.* **2019**, *49*, 1791–1802. [CrossRef]
51. Sudhakaran, S.; Escalera, S.; Lanz, O. Lsta: Long short-term attention for egocentric action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9954–9963.
52. Wang, X.; Wang, Y.-F.; Wang, W.Y. Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. In Proceedings of the NAACL-HLT , New Orleans, LA, USA, 1–6 June 2018; pp. 795–801.
53. Liu, X.; Wang, Z.; Shao, J.; Wang, X.; Li, H. Improving referring expression grounding with cross-modal attention-guided erasing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1950–1959.
54. Huang, P.-Y.; Chang, X.; Hauptmann, A.G. Improving what cross-modal retrieval models learn through object-oriented inter-and intra-modal attention networks. In Proceedings of the 2019 on International Conference on Multimedia Retrieval, Ottawa, ON, Canada, 10–13 June 2019; pp. 244–252.
55. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*; 2014; pp. 2672–2680. Available online: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf (accessed on 3 January 2022).
56. Xu, X.; He, L.; Lu, H.; Gao, L.; Ji, Y. Deep adversarial metric learning for cross-modal retrieval. *World Wide Web* **2019**, *22*, 657–672. [CrossRef]
57. Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; Shen, H.T. Adversarial cross-modal Retrieval. In Proceedings of the 2017 ACM on Multimedia Conference, Mountain View, CA, USA, 23–27 October 2017; Liu, Q., Lienhart, R., Wang, H., Chen, S.K., Boll, S., Chen, Y.P., Friedland, G., Li, J., Yan, S., Eds.; ACM: New York, NY, USA, 2017; pp. 154–162.
58. Liu, R.; Zhao, Y.; Wei, S.; Zheng, L.; Yang, Y. Modality-invariant image-text embedding for image-sentence matching. *ACM Trans. Multimed. Comput. Commun. Appl.* **2019**, *15*, 1–19. [CrossRef]
59. Huang, X.; Peng, Y.; Yuan, M. MHTN: modal-adversarial hybrid transfer network for cross-modal retrieval. *IEEE Trans. Cybern.* **2020**, *50*, 1047–1059. [CrossRef]
60. Zheng, F.; Tang, Y.; Shao, L. Hetero-manifold regularisation for cross-modal hashing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1059–1071. [CrossRef] [PubMed]
61. Wang, Y.; Lin, X.; Wu, L.; Zhang, W.; Zhang, Q. LBMCH: learning bridging mapping for cross-modal hashing. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 9–13 August 2015; Baeza-Yates, R., Lalmas, M., Moffat, A., Ribeiro-Neto, B.A., Eds.; ACM: New York, NY, USA, 2015; pp. 999–1002.
62. Zhang, J.; Peng, Y.; Yuan, M. SCH-GAN: Semi-supervised cross-modal hashing by generative adversarial network. *IEEE Trans. Cybern.* **2020**, *50*, 489–502. [CrossRef] [PubMed]
63. Graves, A.; Mohamed, A.; Hinton, G.E. Speech recognition with deep recurrent neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6645–6649.
64. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1106–1114.
65. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P.P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
66. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
67. Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; Zheng, Y. Nus-wide: A real-world web image database from national university of Singapore. In Proceedings of the ACM International Conference on Image and Video Retrieval, Fira, Greece, 8–10 July 2009; ACM: New York, NY, USA, 2009; p. 48.
68. Rashtchian, C.; Young, P.; Hodosh, M.; Hockenmaier, J. Collecting image annotations using amazon's mechanical turk. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, CA, USA, 6 June 2010; Association for Computational Linguistics: Stroudsburg, PA, USA, 2010; pp. 139–147.
69. Hotelling, H. Relations between two sets of variates. *Biometrika* **1936**, *28*, 321–377. [CrossRef]
70. Rupnik, J.; Shawe-Taylor, J. Multi-view canonical correlation analysis. In Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2010), Ljubljana, Slovenia, 12 October 2010; pp. 1–4.
71. Kan, M.; Shan, S.; Zhang, H.; Lao, S.; Chen, X. Multi-view discriminant Analysis. In *Proceedings of the Computer Vision—ECCV 2012—12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Fitzgibbon, A.W., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., Eds.; Proceedings, Part I*, ser. Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7572, pp. 808–821.

72. Kan, M.; Shan, S.; Zhang, H.; Lao, S.; Chen, X. Multi-view discriminant analysis. *IEEE Trans. Pattern Anal. Machine Intell.* **2016**, *38*, 188–194. [CrossRef]
73. Zhai, X.; Peng, Y.; Xiao, J. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Trans. Circuits Syst. Video Techn.* **2014**, *24*, 965–978. [CrossRef]
74. Peng, Y.; Qi, J.; Huang, X.; Yuan, Y. CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network. *IEEE Trans. Multimed.* **2018**, *20*, 405–420. [CrossRef]
75. Peng, Y.; Huang, X.; Qi, J. Cross-media shared representation by hierarchical learning with multiple deep networks. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016; pp. 3846–3853.
76. Peng, Y.; Qi, J. Cm-gans: Cross-modal generative adversarial networks for common representation learning. *ACM Trans. Multimed. Comput. Commun. Appl.* **2019**, *15*, 1–24. [CrossRef]