Dual Adversarial Graph Neural Networks for Multi-label Cross-modal Retrieval

Shengsheng Qian,[1,2] **Dizhan Xue,**[2] **Huaiwen Zhang,**[1,2] **Quan Fang,**[1,2] **Changsheng Xu**[1,2,3]
[1]National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Peng Cheng Laboratory
{shengsheng.qian, huaiwen.zhang, qfang, csxu}@nlpr.ia.ac.cn, xuedizhan17@mails.ucas.ac.cn

Abstract

Cross-modal retrieval has become an active study field with the expanding scale of multimodal data. To date, most existing methods transform multimodal data into a common representation space where semantic similarities between items can be directly measured across different modalities. However, these methods typically suffer from following limitations: 1) They usually attempt to bridge the modality gap by designing losses in the common representation space which may not be sufficient to eliminate potential heterogeneity of different modalities in the common space. 2) They typically treat labels as independent individuals and ignore label relationships which are important for constructing semantic links between multimodal data. In this work, we propose a novel Dual Adversarial Graph Neural Networks (DAGNN) composed of the dual generative adversarial networks and the multi-hop graph neural networks, which learn modality-invariant and discriminative common representations for cross-modal retrieval. Firstly, we construct the dual generative adversarial networks to project multimodal data into a common representation space. Secondly, we leverage the multi-hop graph neural networks, in which a layer aggregation mechanism is proposed to exploit multi-hop propagation information, to capture the label correlation dependency and learn inter-dependent classifiers. Comprehensive experiments conducted on two cross-modal retrieval benchmark datasets, NUS-WIDE and MIRFlickr, indicate the superiority of DAGNN.

Introduction

Cross-modal retrieval has raised widespread interests with the objective to perform retrieval across different modalities (Wang et al. 2017; Zhen et al. 2019; Liu et al. 2020). Traditionally, modeling semantically similarity links mainly focuses on single-modality scenarios, while the cross-modal retrieval requires that semantically similar items in one modality ($e.g.$, text) can be retrieved with a query item from another modality ($e.g.$, image).

Since instances of different modality typically have inconsistent feature distributions and representations, the modality gap caused by the heterogeneous nature of data needs to be bridged. To bridge the gap among different multimodal data, the general solution is to map them into a common representation space. Traditional methods (Hotelling 1936; Yao, Mei, and Ngo 2015) typically take linear projections as basic models, which often maximize the cross-modal pairwise item correlation or classification accuracy to learn the common representation. Recently, deep neural networks (DNN) are emerging tools to automatically learn feature representations, which have been increasingly utilized in the cross-modal retrieval task (Ngiam et al. 2011; Peng and Qi 2019; Andrew et al. 2013). The DNN-based cross-modal retrieval has become an active study field to exploit nonlinear relationships and achieved great performance improvements. Most existing methods, such as DSCMR (Zhen et al. 2019) and ACMR (Wang et al. 2017), mainly consider how to transform the original representation to the common representation and design handcraft loss functions based on the common representation. However, there may exist modality-specific features in the common representation in the transform process, which may not be fully eliminated by the existing methods and lead to the performance decline. Therefore, we have to face the *Challenge 1*: How to bridge the modality gap so that modality-specific information can be eliminated well and semantic information can be fully preserved in the common representation space?

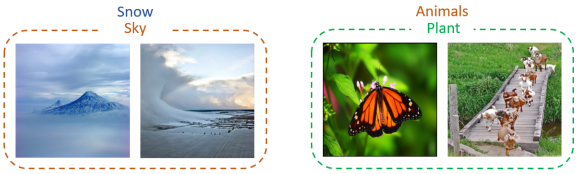


Figure 1: The left picture illustrates the co-occurrence of "Snow" and "Sky" in the NUS-WIDE dataset. The right picture demonstrates the co-occurrence of "Animals" and "Plant" in MIRFlickr dataset.

Another drawback of existing supervised methods is that labels are treated as independent individuals and the label correlations in multimodal data are ignored. For example, cross-modal datasets such as MIRFlickr (Huiskes, Thomee, and Lew 2008) and NUS-WIDE (Chua et al. 2009), often contain multiple labels for each sample, where the semantic dependencies among multiple labels are important for cross-modal retrieval and representation learning. As illustrated

in Figure 1, some labels co-occur in images with a great chance and we can observe that "Sky" usually appears together with "Snow", and "Plant" usually adjoins "Animals". Nevertheless, some pairs of labels hardly occurs in the physical world. Therefore, we need to address the *Challenge 2*: How to explore and capture the label correlation, which contains rich semantic information, to guide the model learning discriminative features?

To overcome the above challenges, we propose a novel deep cross-modal representation learning approach termed Dual Adversarial Graph Neural Networks (DAGNN) for multi-label cross-modal retrieval. For *Challenge 1*, we propose the dual generative adversarial networks (Dual GAN) to generate common representations and further reconstruct the original representations of the other modality. The generator models learn to fit the joint data distribution of different modalities, while the discriminator models learn to discriminate between original data and reconstructed data. By optimizing the adversarial loss and the modality invariance loss, the modality-specific features in the common space can be sufficiently eliminated and the semantic information can be fully preserved. For *Challenge 2*, we propose the multi-hop graph neural networks (Multi-hop GNN) to generate inter-dependent classifiers by exploring the label correlation dependency, which takes a set of labels and their corresponding prior label representations, $e.g.$, word-embedding vectors, as input. The correlation matrix of GNN is calculated according to the statistical features of labels. However, resulting from the limitation of the calculation method and different relative levels between nodes, some pairs of relative nodes may not be connected directly but be connected by some paths in the graph. Therefore, we design a novel layer aggregation mechanism, in which a linear transform is applied on the concatenation of all previous GNN layers, to hierarchically exploit multi-hop propagation information. The learned label classifiers are then employed to classify the common representation, which is generated by the Dual GAN, to perform the end-to-end training process by the classification loss and guide the model learning the underlying semantic structure. In brief, the contributions of our work are listed as follows:

- We propose a novel Dual Adversarial Graph Neural Networks (DAGNN) by designing Dual GAN to bridge the modality gap and constructing Multi-hop GNN to hierarchically explore and capture the label correlation dependency, which can learn modality-invariant and discriminative representation for multi-label cross-modal retrieval.
- We propose a novel cross-modal representation learning method by adopting the dual generative adversarial networks to generate common representations, further cross-reconstruct original representations and discriminate modality-specific features between original representations and reconstructed representations, which can explore and eliminate modality-specific information while the semantic information can be fully preserved.
- We leverage the multi-hop graph neural networks to learn inter-dependent classifiers by exploiting information in the label graph, which can utilize the label correlation to guide the model learning discriminative representations. Furthermore, a layer aggregation mechanism is proposed to flexibly leverage different neighborhood ranges for nodes, which can explicitly guide GNN utilizing multi-hop propagation information and enable better structure-aware representation.
- Extensive experiments demonstrate the superiority of DAGNN compared with ten cross-modal retrieval baselines on two public datasets NUS-WIDE and MIRFlickr.

Related Work

Traditional cross-modal retrieval methods typically take linear projections as basic models, which project multimodal data into a common representation space (Hotelling 1936; Qian et al. 2016; Qian, Zhang, and Xu 2016). Canonical Correlation Analysis (CCA) (Hotelling 1936) learns the common representation by maximizing the pairwise correlations of data between two modalities, which has many extensions, such as RCCA (Yao, Mei, and Ngo 2015), KCCA (Akaho 2001), and ml-CCA (Ranjan, Rasiwasia, and Jawahar 2015). Joint Representation Learning (JRL) (Zhai, Peng, and Xiao 2014) exploits the semantic information and the correlation jointly in a unified optimization framework.

With the development of deep learning, various deep learning based methods have been proposed. Correspondence Autoencoder (Corr-AE) (Feng, Wang, and Li 2014) correlates hidden representations of two unimodal autoencoders. The Adversarial Cross-Modal Retrieval model (ACMR) is proposed by (Wang et al. 2017), where the modality-invariant representation is generated by a feature projector and different modalities are tried to be discriminated by a modal classifier. In (Xie et al. 2020), Multi-Task Consistency-Preserving Adversarial Hashing is proposed to capture the semantic relationships between different modalities. In contrast to above two adversarial approaches, our method utilizes pre-trained true data, which is more stable and authoritative, to guide the adversarial learning. In Deep Supervised Cross-modal Retrieval (DSCMR) (Zhen et al. 2019), the discriminative learning features can be superivised through the combination of the common representation space and the label space. However, all above approaches ignore the label correlation dependency which is important for learning discriminative common representations and is exploited in our method.

Graph neural networks (GNN) have attracted high attention in various research areas, which are designed to apply deep learning architectures on graph-structured data (Yan et al. 2018; Wang et al. 2019; Wu et al. 2020). GCN (Kipf and Welling 2017) is a deep convolutional learning paradigm which integrates local node features and graph topology structure in convolutional layers. ML-GCN (Chen et al. 2019) employs GCN to propagate information among labels and merges label information with CNN features at the final classification stage. In order to tackle the previous models' disadvantages, GAT (Veličković et al. 2017) utilizes the masked self-attention layer based on graph convolutions or their approximations.

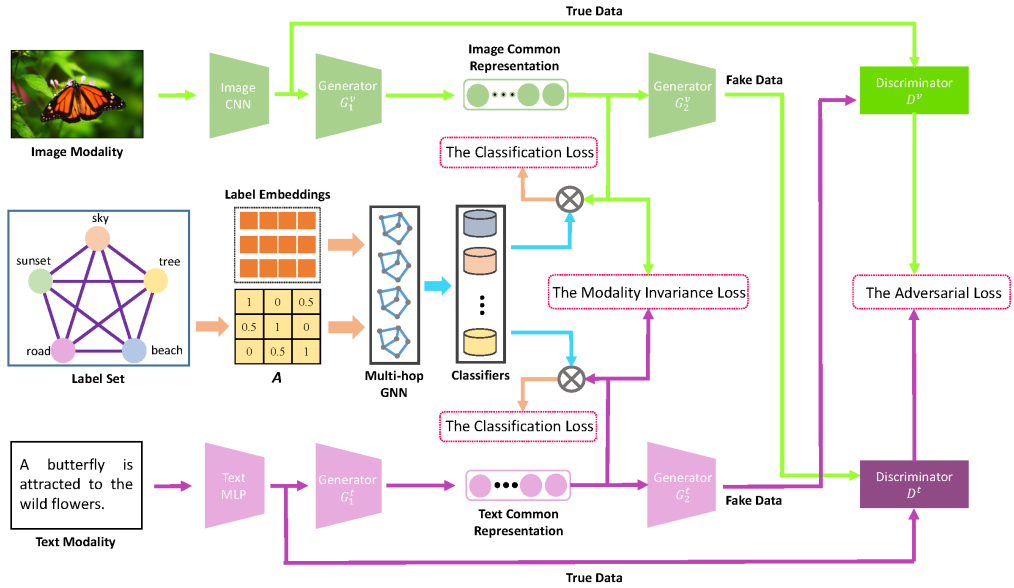


Figure 2: The overall architecture of the proposed Dual Adversarial Graph Neural Networks (DAGNN) model.

Notation and Problem Definition

We firstly introduce some notations that are used in this paper. Assuming that the training set $\mathbf{O} = \{(\mathbf{x}_i^v, \mathbf{x}_i^t)\}_{i=1}^n\}$ contains n image-text pairs. $\mathbf{x}_i^v \in \mathbb{R}^{d_v}$ and $\mathbf{x}_i^t \in \mathbb{R}^{d_t}$ are the image features and text features of the ith sample respectively, where d_v and d_t represent the image and text feature dimensions respectively. Each pair of samples $(\mathbf{x}_i^v, \mathbf{x}_i^t)$ is assigned a semantical label $\mathbf{y}_i = [y_{i1}, y_{i2}, ..., y_{ic}] \in \{0, 1\}^c$, where c denotes the category number. If the ith sample belongs to the jth category, $y_{ij} = 1$; otherwise, $y_{ij} = 0$.

The key to the cross-modal retrieval is learning modality-specific transformation functions: $\mathbf{v}_i = f(\mathbf{x}_i^v; \theta_v) \in \mathbb{R}^d$ for the image modality and $\mathbf{t}_i = g(\mathbf{x}_i^t; \theta_t) \in \mathbb{R}^d$ for the text modality, where θ_v and θ_t are trainable parameters of the two transformation functions, and d represents the number of dimensions in the common representation space. It enables the direct comparison of samples which are in the common representation space but from different modalities, and the similarity of the samples belonging to the same category would be higher than that from different categories. Therefore, the relevant samples of one modality can be retrieved by taking a query from another modality.

Methodology

As demonstrated in Figure 2, the overall architecture of our model consists of two main components, $i.e.$, the dual generative adversarial networks and the multi-hop graph neural networks.

Dual Generative Adversarial Networks

Dual generative adversarial networks (Dual GAN) include two sub-networks trained in an end-to-end style: ImgGAN for feature learning of image modality, TxtGAN for feature learning of text modality. We adopt VGGNet (Simonyan and Zisserman 2015) and Multi-Layer Perception (MLP) (Rumelhart, Hinton, and McClelland 1986) as the backbone of the ImgGAN and the TxtGAN, respectively. The inputs of ImgGAN are the raw images and the inputs of TxtGAN are the bag-of-words (BoW) features provided by datasets.

ImgGAN. We adopt the convolutional layers in 19-layer VGGNet, which is pretrained on the ImageNet (Deng et al. 2009), as convolutional layers of ImgGAN. 4096-dimensional features are generated from fc7 layer as the original representations of raw images, denoted as $\mathbf{h}_i^v \in \mathbb{R}^{d_v} (1 \leq i \leq n)$. Then, several fully-connected layers map $\mathbf{h}_i^v$ to the common representation, denoted as $\mathbf{v}_i$:

$$\mathbf{v}_i = f(\mathbf{x}_i^v; \theta_v) = G_1^v(f_{cnn}(\mathbf{x}_i^v; \theta_{cnn}); \theta_{G_1^v}) \in \mathbb{R}^d \quad (1)$$

where d is the dimensionality of the common representation space and $\theta_v = \{\theta_{cnn}, \theta_{G_1^v}\}$. Then, the common representation will pass the generator of the second GAN (denoted as G_2^v) to reconstruct the semantic representation of text modality, denoted as $\mathbf{r}_i^t$: $\mathbf{r}_i^t = G_2^v(\mathbf{v}_i; \theta_{G_2^v}) \in \mathbb{R}^{d_t}$, where d_t is the dimensionality of the text representation. For convenience, we denote $\theta_{G^v} = \{\theta_{G_1^v}, \theta_{G_2^v}\}$.

TxtGAN. The MLP in TxtGAN, which is pretrained via performing general classification task, consists of two fully-connected layers fc_{t1} and fc_{t2} ($d_t \to 4096 \to 300$) and generates features $\mathbf{h}_i^t \in \mathbb{R}^{d_t} (1 \leq i \leq n)$. Then, several fully-connected layers are followed to obtain the common representation, denoted as $\mathbf{t}_i$:

$$\mathbf{t}_i = f(\mathbf{x}_i^t; \theta_t) = G_1^t(f_{mlp}(\mathbf{x}_i^t; \theta_{mlp}); \theta_{G_1^t}) \in \mathbb{R}^d \quad (2)$$

where d is the dimensionality of the common representation space and $\theta_t = \{\theta_{mlp}, \theta_{G_1^t}\}$. Then, the common representation passes the generator of the second GAN (denoted as G_2^t) and reconstruct the image representation, denoted as $\mathbf{r}_i^v$: $\mathbf{r}_i^v = G_2^t(\mathbf{t}_i; \theta_{G_2^t}) \in \mathbb{R}^{d_v}$, where d_v is the dimensionality of the image representation and we denote $\theta_{G^t} = \{\theta_{G_1^t}, \theta_{G_2^t}\}$.

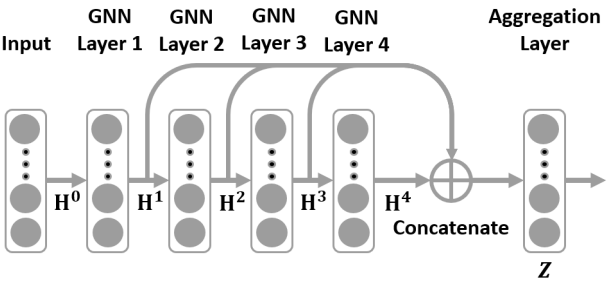


Figure 3: An example of a 4-layer multi-hop graph neural network.

After the cross-reconstruction, discriminators of Img-GAN and TxtGAN take the original representations and the reconstructed representations as input and predict the modalities of the input:

$$\begin{aligned}\hat{\boldsymbol{y}}_i^{vv} &= softmax(D^v(\mathbf{h}_i^v;\theta_{D^v})) \in \mathbb{R}^2 \\ \hat{\boldsymbol{y}}_i^{tv} &= softmax(D^v(\mathbf{r}_i^v;\theta_{D^v})) \in \mathbb{R}^2 \\ \hat{\boldsymbol{y}}_i^{vt} &= softmax(D^t(\mathbf{r}_i^t;\theta_{D^t})) \in \mathbb{R}^2 \\ \hat{\boldsymbol{y}}_i^{tt} &= softmax(D^t(\mathbf{h}_i^t;\theta_{D^t})) \in \mathbb{R}^2\end{aligned} \tag{3}$$

of which the first dimension represents the predicted probability of the input representation belonging to image modality and the other one corresponds to text modality.

Multi-hop Graph Neural Networks

Category labels in our benchmark datasets, such as $plant$, $tree$, $flower$, contain valuable semantic correlations and these labels should not be only involved in the classification loss calculation stage. Motivated by (Chen et al. 2019), we propose Multi-hop Graph Neural Networks (Multi-hop GNN) to learn inter-dependent classifiers by capturing and exploring the label correlations. Firstly, we build a graph $G = (V, E)$ where the category label set is taken as the graph vertice set V while the number of categories c is the number of vertices, and E is the edge set. Each vertex is associated with a $d_{(0)}$-dimensional feature vector. Feature matrix $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^c, \mathbf{Q} \in \mathbb{R}^{c\times d_{(0)}}$ represent the features of vertices in the graph, where $\mathbf{q}_i$ corresponds to the feature of the i-th vertex, and $d_{(0)}$ represents the dimensionality of the label-level word embedding. We also introduce a correlation matrix $\mathbf{A} \in \mathbb{R}^{c\times c}$, where $\mathbf{A}_{ij}$ is the weight of the edge between the ith vertex and the jth vertex. We employ Multi-hop GNN to map the feature matrix $\mathbf{Q} \in \mathbb{R}^{c\times d_{(0)}}$ of these vertices into the corresponding inter-dependent classifiers, $i.e.$, $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^c, \mathbf{Z} \in \mathbb{R}^{c\times d}$. Several classical GNNs, such as GAT (Veličković et al. 2017), GCN (Kipf and Welling 2017), are tested as the backbone of the Multi-hop GNN.

Layer Aggregation Mechanism. We denote the output of the lth GNN layer as $\mathbf{H}^l$. As shown in Figure 3, we apply a concatenation $[\mathbf{H}^1, \mathbf{H}^2, ..., \mathbf{H}^k]$ to combine all layers of a k-layer GNN. Then a layer aggregation mechanism is leveraged to compute aggregated features:

$$\begin{aligned}\mathbf{z}_i &= aggregation(\mathbf{H}_i^1, \mathbf{H}_i^2, ..., \mathbf{H}_i^k) \\ &= \mathbf{W}^{ag}[\mathbf{H}_i^1||\mathbf{H}_i^2||...||\mathbf{H}_i^k] + \mathbf{b}^{ag}\end{aligned} \tag{4}$$

where $\mathbf{W}^{ag} \in \mathbb{R}^{d\times(d_{(1)}+d_{(2)}+...+d_{(k)})}, \mathbf{b}^{ag} \in \mathbb{R}^d$ are parameters of the aggregation layer. Some relative nodes in Figure 2 may be connected by some paths in the graph. Therefore, the layer aggregation mechanism can hierarchically exploit multi-hop propagation information and enable better structure-aware representation. We denote all parameters of the Multi-hop GNN as θ_{GNN}. Then, by applying the learned classifiers to cross-modal representations $\{\mathbf{v}_i\}_{i=1}^n$ and $\{\mathbf{t}_i\}_{i=1}^n$, we can obtain the predicted score as:

$$\begin{aligned}\hat{\boldsymbol{y}}_{ij}^v &= cos(\mathbf{v}_i, \mathbf{z}_j) = \frac{\mathbf{v}_i\mathbf{z}_j^{\mathrm{T}}}{\|\mathbf{v}_i\|_2 \cdot \|\mathbf{z}_j\|_2}, \\ \hat{\boldsymbol{y}}_{ij}^t &= cos(\mathbf{t}_i, \mathbf{z}_j) = \frac{\mathbf{t}_i\mathbf{z}_j^{\mathrm{T}}}{\|\mathbf{t}_i\|_2 \cdot \|\mathbf{z}_j\|_2}\end{aligned} \tag{5}$$

where $\hat{\boldsymbol{y}}_{ij}^v$ is the predicted score of the ith image sample belonging to the jth category and $\hat{\boldsymbol{y}}_{ij}^t$ is the predicted score of the ith text sample belonging to the jth category.

Correlation Matrix of Multi-hop GNN

Our proposed correlation matrix $\mathbf{A}$ is based on statistical information as label co-occurrence patterns. Firstly, we compute conditional probability, $i.e.$, $P(L_i|L_j)$, which represents the occurrence probability of label L_i when label L_j appears. We count the co-occurrence frequency $\mathbf{M} \in \mathbb{N}^{c\times c}$ of label pairs and the occurrence frequency $\mathbf{N} \in \mathbb{N}^c$ of all labels in the training set. Then we calculate the conditional probability matrix:

$$\mathbf{P}_{ij} = P(L_i|L_j) = \frac{P(L_i, L_j)}{P(L_j)} = \frac{\mathbf{M}_{ij}}{\mathbf{N}_j}. \tag{6}$$

However, this matrix may be faced with two problems while performing as the correlation matrix. Firstly, some rare co-occurrence patterns between labels may be noise, which results from the existence of long-tail distributions. Secondly, the co-occurrence frequencies between labels may not be completely consistent in the training and testing scenarios, which can affect the generalization ability. Therefore, we propose a binary process with the threshold τ:

$$\mathbf{B}_{ij} = \begin{cases}0, & if\ \mathbf{P}_{ij} < \tau \\ 1, & if\ \mathbf{P}_{ij} \geq \tau\end{cases} \tag{7}$$

where $\mathbf{B}$ is the binary correlation matrix. The feature of one node is the weighted sum of features of itself and its adjacent nodes after every layer in GNN. Thus, the binary correlation matrix can result in over-smoothing, which means features of different nodes become indistinguishable. In order to solve this problem, the following re-weighted trick is applied:

$$\mathbf{A}_{ij} = \begin{cases}p \cdot \mathbf{B}_{ij}, & if\ i \neq j \\ 1, & if\ i = j\end{cases} \tag{8}$$

where $\mathbf{A}$ is the re-weighted correlation matrix, and p controls the weight assigned to a vertex itself and its neighbors. Via setting a suitable p, propagation information from different nodes can be integrated appropriately. Remarkably, when $p \to 0$, the neighboring information will be completely ignored. It is worth noting that the value of $p(> 0)$ does not influence the propagation process of GAT.

Objective Function

We can learn the instance representation via minimizing the classification error and the dissimilarity of samples from same semantic categories, while maximizing the dissimilarity of samples from different categories. Furthermore, we can bridge the modality gap while the generators try to generate modality-invariant representations and the discriminators try to discriminate between modalities.

We adopt the following objective function to measure the classification loss:

$$\mathcal{L}_{cla} = \frac{1}{n}\sum_{i=1}^{n} \|\hat{\boldsymbol{y}}_i^v - \mathbf{y}_i\|_2 + \frac{1}{n}\sum_{i=1}^{n} \|\hat{\boldsymbol{y}}_i^t - \mathbf{y}_i\|_2 \quad (9)$$

where $\|\cdot\|_2$ denotes the $l2$ norm, $\hat{\mathbf{y}}_i^v = (\hat{y}_{i1}^v, \hat{y}_{i2}^v, ..., \hat{y}_{ic}^v)$ is the predicted label of the ith image and $\hat{\mathbf{y}}_i^t = (\hat{y}_{i1}^t, \hat{y}_{i2}^t, ..., \hat{y}_{ic}^t)$ is the predicted label of the ith text.

Furthermore, we also measure the modality invariance loss that consists of inter-modal and intra-modal invariance losses in the common representation space:

$$\begin{aligned}\mathcal{L}_{mdi} =& \frac{1}{n^2}\sum_{1\le i,j\le n}(log(1+e^{\Gamma_{ij}}) - S_{ij}^{vt}\Gamma_{ij}) \\ &+\frac{1}{n^2}\sum_{1\le i,j\le n}(log(1+e^{\Phi_{ij}}) - S_{ij}^{vv}\Phi_{ij}) \\ &+\frac{1}{n^2}\sum_{1\le i,j\le n}(log(1+e^{\Theta_{ij}}) - S_{ij}^{tt}\Theta_{ij})\end{aligned} \quad (10)$$

where $\Gamma_{ij} = \frac{1}{2}cos(\mathbf{v}_i, \mathbf{t}_j)$, $\Phi_{ij} = \frac{1}{2}cos(\mathbf{v}_i, \mathbf{v}_j)$, $\Theta_{ij} = \frac{1}{2}cos(\mathbf{t}_i, \mathbf{t}_j)$, $S_{ij}^{vt} = \mathbf{y}_i\mathbf{y}_j^{\mathrm{T}}$, $S_{ij}^{vv} = \mathbf{y}_i\mathbf{y}_j^{\mathrm{T}}$, $S_{ij}^{tt} = \mathbf{y}_i\mathbf{y}_j^{\mathrm{T}}$, $cos(\cdot)$ is the cosine similarity function. The first term of Equation 10 can be written as $-\frac{1}{n^2}\sum_{1\le i,j\le n} log(\frac{e^{S_{ij}^{vt}\Gamma_{ij}}}{1+e^{\Gamma_{ij}}})$. When $S_{ij}^{vt} = 0$, minimizing this loss equals to minimizing the similarity Γ_{ij}, otherwise, it is equivalent to maximizing Γ_{ij}.

In order to sufficiently bridge the modality gap, we compute the adversarial loss of the Dual GAN:

$$\begin{aligned}\mathcal{L}_{adv} =& -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i^{vv}log(\hat{\boldsymbol{y}}_i^{vv}) - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i^{tv}log(\hat{\boldsymbol{y}}_i^{tv}) \\ &-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i^{vt}log(\hat{\boldsymbol{y}}_i^{vt}) - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i^{tt}log(\hat{\boldsymbol{y}}_i^{tt})\end{aligned} \quad (11)$$

where $\boldsymbol{y}_i^{vv} = \boldsymbol{y}_i^{vt} = [1, 0](1 \le i \le n)$ are one-hot modal labels for image instances and $\boldsymbol{y}_i^{tv} = \boldsymbol{y}_i^{tt} = [0, 1](1 \le i \le n)$ are one-hot modal labels for text instances.

Combining Equation 9, 10 and 11, we obtain the final objective function of DAGNN which can be optimized via Adam algorithm (Kingma and Ba 2015). Since the optimization goals of $\mathcal{L}_{adv}$ and $\mathcal{L}_{cla}$, $\mathcal{L}_{mdi}$ are opposite, the optimization process runs as a minimax game of the two concurrent sub-process:

$$\begin{aligned}(\theta_{G^v}, \theta_{G^t}, \theta_{GNN}) &= \arg\min_{\theta_{G^v},\theta_{G^t},\theta_{GNN}} (\mathcal{L}_{cla} + \alpha\mathcal{L}_{mdi} - \beta\mathcal{L}_{adv}) \\ (\theta_{D^v}, \theta_{D^t}) &= \arg\max_{\theta_{D^v},\theta_{D^t}} (\mathcal{L}_{cla} + \alpha\mathcal{L}_{mdi} - \beta\mathcal{L}_{adv})\end{aligned} \quad (12)$$

where the hyper-parameters α, β are trade-off factors of three components. Following (Ganin and Lempitsky 2015), we perform the minimax optimization by applying Gradient Reversal Layer (GRL), which is transparent during the forward-propagation, but multiplies the gradients by -1 during the back-propagation. Adding GRL before the first layer of the modality discriminators D^v and D^t, the minimax optimization can be performed simultaneously.

Experiments

We conduct sufficient experiments on two public multi-label cross-modal datasets NUS-WIDE and MIRFlickr to verify the superiority of DAGNN. By default, we adopt GAT-based Multi-hop GNN in DAGNN.

Datasets

NUS-WIDE (Chua et al. 2009) We construct a subset of this dataset, which contains 190,421 image-text pairs covering the 21 most commonly used concepts, through pruning samples that are either unlabeled or that have no tag information. Every image-text pair is represented by a 224x224 RGB array and an index vector of 1,000 of the most common tags. Following (Zhen et al. 2019; Su, Zhong, and Zhang 2019), we randomly pick up 2,000 image-text pairs as the testing set and the rest as the training set.

MIRFlickr (Huiskes, Thomee, and Lew 2008) is composed of 25,000 image-text pairs labeled by 24 categories. Every pair is represented by a 224x224 RGB array and a tag vector with the dimension of 500. Then 2,000 image-text pairs are randomly selected as the testing set and the rest are used for training.

Baseline Methods and Evaluation Metrics

Five traditional methods CFA (Li et al. 2003), CCA (Hotelling 1936), PLS-C2A (Tenenhaus 1998), JRL (Zhai, Peng, and Xiao 2014) and ml-CCA (Ranjan, Rasiwasia, and Jawahar 2015), and five deep learning-based methods Multimodal DBN (Srivastava and Salakhutdinov 2012), Corr-AE (Feng, Wang, and Li 2014), DCCA(Andrew et al. 2013), ACMR (Wang et al. 2017) and DSCMR (Zhen et al. 2019) are selected as the baseline methods to compare with DAGNN. For evaluating cross-modal retrieval performance on all approaches, we utilize the mean Average Precision (mAP) score (Wang et al. 2017; Zhen et al. 2019).

Implementation Details

We utilize mini-batch Adam optimizer (Kingma and Ba 2015) to optimize the whole model, and our algorithms are implemented on Pytorch deep learning framework (Paszke et al. 2017; Hu et al. 2021). The batch size m is set as 1024 for NUS-WIDE and 100 for MIRFlickr. The initial learning rates of the optimizer are 0.00005 on both datasets.

The selection of hyper-parameters. First, we choose hyper-parameters τ and p by employing the grid search and finally set $\tau = 0.3$ for the GAT-based Multi-hop GNN and $\tau = 0.4,\ p = 0.5$ for the GCN-based Multi-hop GNN on

Methods	NUS-WIDE			MIRFlickr		
	Image2Text	Text2Image	Average	Image2Text	Text2Image	Average
CFA	0.354	0.361	0.357	0.580	0.548	0.564
CCA	0.656	0.664	0.660	0.712	0.722	0.717
PLS-C2A	0.632	0.631	0.631	0.730	0.740	0.735
JRL	0.427	0.361	0.394	0.589	0.554	0.571
ml-CCA	0.669	0.668	0.668	0.734	0.742	0.738
Multimodal DBN	0.342	0.321	0.331	0.575	0.561	0.568
Corr-AE	0.632	0.629	0.630	0.708	0.727	0.717
DCCA	0.637	0.649	0.643	0.736	0.746	0.741
ACMR	0.684	0.675	0.680	0.736	0.748	0.742
DSCMR	0.706	0.739	0.722	0.752	0.799	0.775
DAGNN	**0.753**	**0.761**	**0.757**	**0.804**	**0.817**	**0.811**

Table 1: The mAP result comparisons on NUS-WIDE and MIRFlikcr datasets.

Methods	NUS-WIDE			MIRFlickr		
	Image2Text	Text2Image	Average	Image2Text	Text2Image	Average
DAGNN-1	0.469	0.572	0.520	0.731	0.715	0.723
DAGNN-2	**0.754**	0.756	0.755	0.798	0.816	0.807
DAGNN-3	0.746	0.755	0.750	0.792	**0.817**	0.805
DAGNN-4	0.740	0.755	0.747	0.787	**0.817**	0.802
DAGNN-5	0.727	0.741	0.734	0.749	0.771	0.760
DAGNN	0.753	**0.761**	**0.757**	**0.804**	**0.817**	**0.811**

Table 2: The mAP result comparisons between DAGNN variants on NUS-WIDE and MIRFlikcr.

both NUS-WIDE and MIRFlickr datasets. Second, we validate the hyper-parameters α and β in Equation 12 and finally set $\alpha = 0.2,\ \beta = 0.2$ for both datasets.

The convolutional layers of ImgGAN have the same configuration with 19-layer VGGNet and the text MLP of TxtGAN is pretrained by performing the general classification task as mentioned in the Methodology Section. Then one fully-connected layer (G_1^v and G_1^t) activated by Rectified Linear Unit (ReLU) (Nair and Hinton 2010) is followed in both ImgGAN and TxtGAN. The output dimensionality of G_1^v and G_1^t is both 1,024. In ImgGAN, the generator G_2^t consists of two fully-connected layers with the number of hidden units 512 and 300, which are activated by LeakyReLU (Maas, Hannun, and Ng 2013) with the negative slope of 0.2. In TxtGAN, the generator G_2^v consists of two fully-connected layers with the number of hidden units 2,048 and 4,096, which are activated by ReLU. The multi-hop graph neural networks consist of five GAT layers on NUS-WIDE and four GAT layers on MIRFlickr together with one aggregation layer, in which the output dimensionality of each GAT layer and aggregation layer is 1,024. For labeling representation, we adopt 300-dimensional Glove (Pennington, Socher, and Manning 2014) word embeddings as inputs.

Results and Discussions

Table 1 shows the mAP results of DAGNN and baseline methods on NUS-WIDE and MIRFlickr in two cross-modal retrieval tasks. $Image2Text$ refers the query as the image modality and the database as the text modality, while $Text2Image$ refers the situation that the query as the text modality and the database as the image modality. From the results, we have following observations: (1) Some traditional methods with deep features can also achieve high mAP scores. Specifically, linear methods CCA, PLS-C2A, and ml-CCA obtained promising results which can compete with some supervised deep learning methods (Corr-AE and DCCA). This demonstrates that our f_{cnn} and f_{mlp} have transformed input instances into approximately linear spaces, greatly reducing the difficulties in cross-modal retrieval. (2) Deep learning methods that utilize label information (ACMR and DSCMR) outperform other traditional methods, which shows that nonlinear transformation models have competitive advantages than traditional linear transformation models. (3) The proposed DAGNN model beats all baselines on the two datasets consistently. Compared to DSCMR, the state-of-the-art deep cross-modal retrieval method, our method achieves 4.7% and 2.2% improvements in mAP scores in $Image2Text$ and $Text2Image$ tasks on NUS-WIDE respectively and achieves 5.2% and 1.8% improvements on MIRFlickr. It demonstrates that DAGNN can better capture the underlying semantic correlation of labels and bridge the modality gap.

Ablation Study

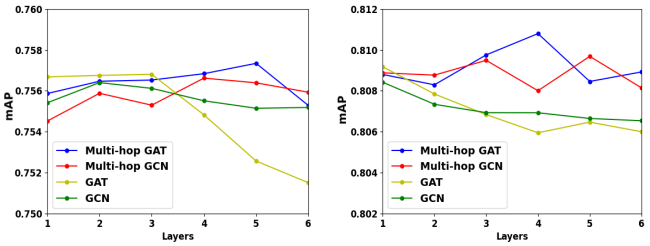
Multi-hop GNN and Components of Objective Function The objective function of the proposed DAGNN consists of three terms, aiming at optimizing the classification loss, the modality invariance loss, and the adversarial loss. To verify the effect of objective function terms and the Multi-hop GNN on the cross-modal retrieval performance, several variants are designed as following:

DAGNN-1 abandons $\mathcal{L}_{cla}$ and retains $\mathcal{L}_{mdi}$ and $\mathcal{L}_{adv}$.

DAGNN-2 abandons $\mathcal{L}_{mdi}$ and retains $\mathcal{L}_{cla}$ and $\mathcal{L}_{adv}$.

DAGNN-3 abandons $\mathcal{L}_{adv}$ and retains $\mathcal{L}_{cla}$ and $\mathcal{L}_{mdi}$.

DAGNN-4 abandons the cross-reconstruction and adopts the discriminator to discriminate between common representations of image and text modalities.



(a) NUS-WIDE (b) MIRFlickr

Figure 4: Effects of different GNN models on NUS-WIDE and MIRFlickr.

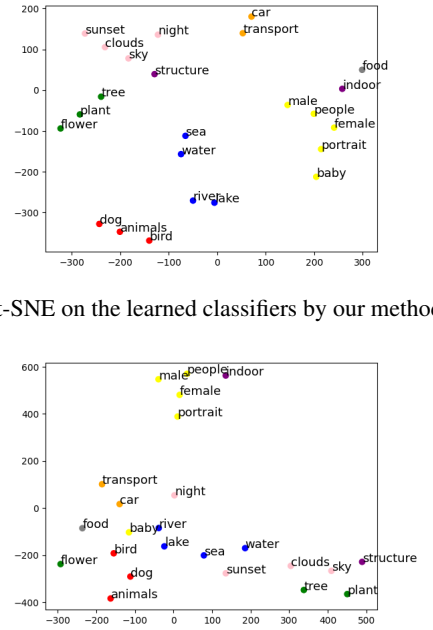
DAGNN-5 abandons the multi-hop graph neural networks and adopts a linear classifier instead.

We conduct the cross-modal retrieval task with these variants on two datasets. The optimization procedure of all variants is similar to the proposed DAGNN. The overall results are shown in Table 2. According to the results, we have following observations: (1) DAGNN outperforms DAGNN-1, DAGNN-2, and DAGNN-3, which indicates that each of the three terms in objective function contributes to the final results. (2) DAGNN outperforms DAGNN-1 with a large gap while the margin between DAGNN and DAGNN-2 or DAGNN-3 is small, which proves $\mathcal{L}_{cla}$ is the major component of the objective function. (3) DAGNN-2 outperforms DAGNN-3, which indicates that the adversarial loss is more sufficient than the modality invariance loss for bridging the modality gap. (4) DAGNN outperforms DAGNN-4, which demonstrates the effectiveness of the Dual GAN. Compared with ACMR, the adversarial reconstruction technique can sufficiently eliminate modality-specific features and the exploitation of pre-trained true data can enhance the stability of the adversarial learning. (5) DAGNN performs far superior to DAGNN-5, which indicates the effectiveness of the Multi-hop GNN. Compared with utilizing the linear classifier, our GNN-based method can learn inter-dependent classifiers which effectively exploit and preserve the semantic structure of labels.

Different GNN Models By default, we utilize Multi-hop GAT to learn inter-dependent classifiers. To verify the effectiveness of Multi-hop GAT, we try different GNN models and the number of GNN layers, which are shown in Figure 4. When using GNN models without the layer aggregation mechanism (GAT, GCN), the cross-modal retrieval mAP achieves the best performances with the layer number of one or two and declines with more layers. Compared with two traditional GNN models, Multi-hop GNN can hierarchically exploit multi-hop propagation information and get better performance as the model becomes deeper.

Classifier Visualization

To evaluate if the meaningful semantic structure can be preserved by our method, we visualize the learned classifiers on MIRFlickr dataset in different learning manners. Firstly, we adopt the t-SNE (van der Maaten and Hinton 2008) to visualize the classifiers learned by our proposed DAGNN. Secondly, we abandon the Multi-hop GNN and directly parametrize classifiers, the learned classifiers with the same training settings are also visualized by t-SNE. Figure 5 shows the results of the two methods. We can observe that classifiers in Figure 5(a) tend to be close to semantically similar classifiers, which demonstrates our method can maintain more semantic structures than another one. For example, compared with Figure 5(a), the classifier *baby* is far from its similar classifiers *people*, *male* and *female*, and the classifier *flower* stays away from its relative classifiers *plant* and *tree* in Figure 5(b). This visualization performance moreover demonstrates effectiveness of DAGNN in modeling label dependencies.



(a) t-SNE on the learned classifiers by our method.

(b) t-SNE on the learned classifiers without the Multi-hop GNN.

Figure 5: Visualization of the learned classifiers.

Conclusion

In this work, we propose a novel Dual Adversarial Graph Neural Networks (DAGNN) to learn common representations for cross-modal retrieval. To better bridge the modality gap and preserve the underlying semantic structure, we introduce an end-to-end framework composed of the dual generative adversarial networks and the multi-hop graph neural networks, which can preserve the semantical relationships among multimodal instances and capture the latent semantic structure of labels. Additionally, Multi-hop GNN is proposed to flexibly leverage different neighbor ranges for nodes, which can learn better structure-aware representation via hierarchically exploiting multi-hop propagation information. Comprehensive experimental results on two public benchmark datasets indicate that DAGNN outperform state-of-the-art methods in cross-modal retrieval.

Acknowledgments

This work was supported by National Key Research and Development Program of China (No. 2017YFB1002804), National Natural Science Foundation of China (No. 62036012, 61721004, 61720106006, 61802405, 62072456, 61832002, 61936005 and U1705262), the Key Research Program of Frontier Sciences, CAS, Grant NO. QYZDJSSWJSC039, and the K.C.Wong Education Foundation.

References

Akaho, S. 2001. A kernel method for canonical correlation analysis. In *International Meeting of Psychometric Society*, 263–269.

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep Canonical Correlation Analysis. In *ICML*, 1247–1255.

Chen, Z.; Wei, X.; Wang, P.; and Guo, Y. 2019. Multi-Label Image Recognition with Graph Convolutional Networks. In *CVPR*, 5177–5186.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *ICIVR*, 48–56.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Li, F.-F. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Feng, F.; Wang, X.; and Li, R. 2014. Cross-modal Retrieval with Correspondence Autoencoder. In *ACM MM*, 7–16.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189.

Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 28(3-4): 321–377.

Hu, J.; Qian, S.; Fang, Q.; Wang, Y.; Zhao, Q.; Zhang, H.; and Xu, C. 2021. Efficient Graph Deep Learning in TensorFlow with tf_geometric. *CoRR* abs/2101.11552.

Huiskes, M. J.; Thomee, B.; and Lew, M. S. 2008. The MIR flickr retrieval evaluation initiative. In *ICMIR*, 39–43.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Li, D.; Dimitrova, N.; Li, M.; and Sethi, I. K. 2003. Multimedia content processing through cross-modal association. In *ACM MM*, 604–611.

Liu, S.; Qian, S.; Guan, Y.; Zhan, J.; and Ying, L. 2020. Joint-modal Distribution-based Similarity Hashing for Large-scale Unsupervised Deep Cross-modal Retrieval. In *Proceedings of the 43rd International ACM SIGIR*, 1379–1388.

Maas, A. L.; Hannun, A. Y.; and Ng, A. Y. 2013. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 1–6.

Nair, V.; and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*, 807–814.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal Deep Learning. In Getoor, L.; and Scheffer, T., eds., *ICML*, 689–696.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch .

Peng, Y.; and Qi, J. 2019. CM-GANs: Cross-modal Generative Adversarial Networks for Common Representation Learning. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15(1): 22.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global vectors for word representation. In *EMNLP*, 1532–1543.

Qian, S.; Zhang, T.; and Xu, C. 2016. Multi-modal Multi-view Topic-opinion Mining for Social Event Analysis. In *ACM Conference on Multimedia Conference*, 2–11. ACM.

Qian, S.; Zhang, T.; Xu, C.; and Shao, J. 2016. Multi-Modal Event Topic Model for Social Event Analysis. *IEEE Trans. Multim.* 18(2): 233–246.

Ranjan, V.; Rasiwasia, N.; and Jawahar, C. V. 2015. Multi-Label Cross-modal Retrieval. In *ICCV*, 4094–4102.

Rumelhart, D. E.; Hinton, G. E.; and McClelland, J. L. 1986. A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition* 1(45-76): 26.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Srivastava, N.; and Salakhutdinov, R. 2012. Learning representations for multimodal data with deep belief nets. In *ICML*, 1–8.

Su, S.; Zhong, Z.; and Zhang, C. 2019. Deep Joint-Semantics Reconstructing Hashing for Large-Scale Unsupervised Cross-Modal Retrieval. In *ICCV*, 3027–3035.

Tenenhaus, M. 1998. La régression PLS: théorie et pratique. *éditions technip, paris* .

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9: 2579–2605.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* .

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017. Adversarial cross-modal retrieval. In *ACM MM*, 154–162.

Wang, Y.; He, D.; Li, F.; Long, X.; Zhou, Z.; Ma, J.; and Wen, S. 2019. Multi-Label Classification with Label Graph Superimposing. *arXiv: Computer Vision and Pattern Recognition* .

Wu, M.; Pan, S.; Zhou, C.; Chang, X.; and Zhu, X. 2020. Unsupervised Domain Adaptive Graph Convolutional Networks. In *The Web Conference 2020, 2020*, 1457–1467.

Xie, D.; Deng, C.; Li, C.; Liu, X.; and Tao, D. 2020. Multi-Task Consistency-Preserving Adversarial Hashing for Cross-Modal Retrieval. *IEEE Transactions on Image Processing* 29: 3626–3637.

Yan, S.; Xiong, Y.; Lin, D.; and Tang, X. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *AAAI*, 7444–7452.

Yao, T.; Mei, T.; and Ngo, C. 2015. Learning Query and Image Similarities with Ranking Canonical Correlation Analysis. In *ICCV*, 28–36.

Zhai, X.; Peng, Y.; and Xiao, J. 2014. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Trans. Circuits Syst. Video Technol.* 24(6): 965–978.

Zhen, L.; Hu, P.; Wang, X.; and Peng, D. 2019. Deep Supervised Cross-Modal Retrieval. In *CVPR*, 10394–10403.