Cross-modal Retrieval based on Shared Proxies

Yuxin Wei
Guangzhou University
Ligang Zheng (zlg@gzhu.edu.cn)
Guangzhou University
Guoping Qiu
University of Nottingham
Guocan Cai
Guangzhou University

Research Article

Keywords: Cross-modal, retrieval, proxy, modality gap, neighbour component analysis

Posted Date: March 10th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2667484/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at International Journal of Multimedia Information Retrieval on January 20th, 2024. See the published version at https://doi.org/10.1007/s13735-023-00316-2.

Cross-modal Retrieval based on Shared Proxies

Yuxin Wei[1], Ligang Zheng[1*], Guoping Qiu[2] and Guocan Cai[1]

[1]School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, 510006, China.

[2]School of Computer Science, University of Nottinghan, Nottingham, NG8 1BB, United Kingdom.

*Corresponding author(s). E-mail(s): zlg@gzhu.edu.cn;
Contributing authors: wyx@e.gzhu.edu.cn; guoping.qiu@nottingham.ac.uk;
guocancai@e.gzhu.edu.cn;

Abstract

Inconsistency of distribution and representation across different data modalities makes measuring cross-modal similarities a very difficult problem. Learning a common space that is semantically discriminative and modality invariant is the main challenge in cross-modal retrieval. Existing solutions usually employ pairwise or triplet data relationships to learn the common space, which can only capture the data similarity locally but would be unable to effectively characterize the global geometry of the common embedding space, and thus would limit the performance of cross-modal retrieval. In this paper, we introduce a *shared proxy* solution to cross-modal retrieval. We propose to incorporate the principles of *shared proxy* with neighbourhood component analysis to learn a common space for different modalities in which the distance between a sample's representation and its corresponding proxy is minimized while the distances between a sample's representation and the proxies not belonging to the sample are maximized. We propose the Cross-mOdal proXy learnIng (COXI) framework which integrates a cross-modal shared proxy loss, a discriminative loss and a modality invariant loss for supervised cross-modal retrieval. Extensive experiments on benchmark datasets clearly shows that COXI outperforms state of the art cross-modal retrieval techniques. Code is available on https://github.com/LigangZheng/COXI.

Keywords: Cross-modal, retrieval, proxy, modality gap,neighbour component analysis

1 Introduction

With the rapid development of mobile devices, social networks and self-media platforms, increasing amounts of data are being generated and collected at an unprecedented speed. It is common that different types of data are used to describe the same events or topics. For example, sports news can be presented via either images, videos or texts on a sports webpage. Images, videos and texts are different raw data forms (also known as modality). Data consists more than one such raw forms is referred to as multi-modal data.

The main task of cross-modal retrieval is to retrieve one modality of data based on another. Unlike traditional uni-modal retrieval which only deals with a single modality, cross-modal retrieval involves similarity search across different modalities. However, each modality usually presents

distinct distribution properties known as heterogeneous gap [1–3], and this makes it very challenging to measure the similarities between different modalities [2–8].

For addressing the above issue, an intuitive idea is to map the heterogeneous data into a semantic and modality aligned common space where the similarity can be directly measured by adopting a pre-defined distance metric such as the Euclidean or Cosine distance [6–9]. Thus the core of cross-modal retrieval is to learn a common space where feature representations from different modalities can be aligned according to their semantics and should have these properties: 1) *semantic structure preserving* - the feature representations of the same class from any modalities should be mapped as closely to each other as possible, while the feature representations of different classes should be separated as far as possible; 2) *modality alignment* - the feature representations of the same instance from different modalities should be as close as possible. An example of an "aligned common space" is shown in Fig. 1 in supplementary information.

A variety of methods have been proposed to learn the common space. Traditional methods mainly use statistical analysis technique to project the features of different modalities into a subspace to generate common representations, examples include canonical correlation analysis (CCA) [10] and its many extensions [11–13]. However, CCA and its variants only provide a very coarse association between different modalities, thus their cross-modal alignment capability is limited [6, 7, 9]. Inspired by the success of deep representation learning, researchers have proposed a large number of deep learning methods to build the common space for cross-modal retrieval, examples include but not limited to [4, 5, 8].

Although existing methods have achieved promising results, they also have limitations. One of the issues with existing cross-modal retrieval methods is that they usually use inter-/intra-modal pairwise or triplet data relationships to design a metric loss for common space learning. This means that they can only capture the data similarity locally but will fail to effectively characterize the global geometry of the common embedding space [14–16]. Another problem is that these methods rely on sampling pairs or triplets samples in the mini-batches during training, and this could lead to the sampling of prohibitively large number of tuples including many non-informative tuples thus causing slow convergence and degraded performances. Although efficient methods such as [17] that select hard/semi-hard negative samples from subsets of the whole training dataset can alleviate the problem somewhat, this kind of methods also have the weakness of failing to characterize the global geometry of the embedding space. In order to overcome these issues, we propose a proxy based solution. For a given sample, instead of comparing it with every other samples, we compare it with a small number of proxies, which avoid the problem associated with sampling. Furthermore, using proxies can capture the global geometry structure of an embedding space.

However, existing proxy-based loss functions are designed for single modality and it is unknown how the relations amongst the proxies of different constituent modalities should be handled for the multi-modality problems. What is desired is that the aligned common space of the multi-modal data must be *semantically discriminative* and *modality invariant*. That is, representations of samples with the same class label, even if they are from different modalities, should have the same distribution and close to each other. At the same time, representations of samples with different class labels should have different distributions and far apart from each other regardless of their modalities. Based on this reasoning, we introduce the concept of *shared proxy* to represent the class distribution of multi-modal data and approximate the global structure of the common embedding space where the distribution of different constituent modalities can be aligned. Different from the single modality scenario, where a proxy is only for one modality, the shared proxy, just as its name implies, is a proxy shared by all modalities. Fig. 1 gives a toy example to illustrate the difference between traditional proxy and the proposed shared proxy solution.

By combining the principle of shared proxy with neighbourhood component analysis (NCA) [18], we propose the cross-modal shared proxy loss (CMSP) to learn the globally aligned common space of multi-modal data by minimizing the distance between a representation and its corresponding shared proxy, and maximizing the distance between a representation and proxies not belonging to the representation. During the learning process, the samples with the same label are

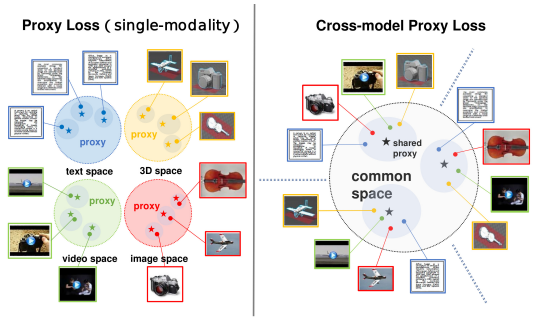


Fig. 1 Left: traditional proxy. Right: Cross modal shared proxy. The traditional proxy method learns proxies within a single modality. The proposed shared proxy method learns proxies across all modalities, and use the proxies to represent class distributions. By considering the relations among the proxies of different modalities, our method not only preserves the semantic structure, but also reduces the cross-modal discrepancy and achieves modality alignment. Combined with the NCA mechanism, the proposed method can learn a discriminative and modality-invariant common space.

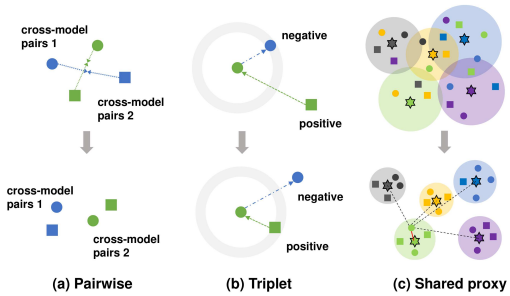


Fig. 2 Inter-modal pairwise similarity, triplet similarity and the proposed cross-modal shared proxy based similarity. Pairwise and triplet learning only considers a pair/triplet of data at a time, while our cross-modal shared proxy based similarity encourages all similar data points to collapse to the corresponding shared proxy.

pulled toward their corresponding shared proxy and are pushed away from the shared proxies of other classes. The shared proxies are learned to enable multi-modality sample representations in the shared space to maintain the *semantically discriminative* and *modality invariant* property. Fig. 2 gives a toy illustration for the difference between pair-/triplet- based loss and the CMSP loss for the multi-modal data. We further propose a novel Cross-mOdal proXy LearnIng (COXI) framework which integrates the CMSP loss, a discriminative loss and a modality invariant loss for supervised cross-modal retrieval.

To verify the effectiveness of the proposed COXI framework, we conduct extensive experiments on four widely used benchmark datasets. Results clearly show that COXI outperforms state of the art cross-modal retrieval techniques. The main contributions of this work can be summarised as follows:

- To overcome the weakness of pair-/triplet-based methods, we propose the cross-modal shared proxy loss (CMSP) to learn the globally aligned common space for multi-modal data. CMSP combines the principles of shared proxy and NCA to learn class distributions across modalities by minimizing the distance between a representation and its corresponding shared proxy and maximizing the distances between a representation and the proxies not belonging to the representation.
- We propose a Cross-mOdal proXy learnIng (COXI) framework which integrates the CMSP, a discrimination loss and a modality invariant loss for supervised cross-modal retrieval. COXI achieves state of the art cross-modal retrieval results and the code is made available.

2 Related Work

In cross-modal retrieval, users can search various modalities of data including texts, images, videos or 3D-shapes using one modality as query. Due to the heterogeneous gap, it is very difficult to measure the content similarity between different modalities. In order to bridge the heterogeneous data, cross-modal retrieval requires a common embedding space that allows the computation of the similarity between query representation and that of the retrieved data. A variety of approaches have been proposed to learn the common embedding space, which can be roughly classified as two categories, binary-representation learning [19–22] (also known as cross-modal hashing) and real-valued representation learning [4, 6–8, 23]. We review methods of real-valued representation learning as it is the most relevant to the current study.

Subspace Learning. Most traditional methods convert multi-modal data into the common subspace by using standard statistical techniques. For example, canonical correlation analysis (CCA)

[10] learns linear projection matrices by maximizing pairwise correlation of multi-modal data. Many researchers have extended CCA to cross-modal retrieval tasks [11–13, 24, 25]. Sharma *et al.* [11] extend CCA by leveraging semantic label information to guide the learning of the common subspace representations. Ranjan *et al.* [24] propose multi-label CCA (ml-CCA) by applying multi-label semantic information to build correspondence between different modalities. Fast ml-CCA [24] is an efficient version of ml-CCA, which is able to handle large-scale datasets. Gong *et al.* [12] extend CCA by incorporating a third view which is represented either by a single category or multiple non-mutually-exclusive concepts. Rasiwasia *et al.* [25] propose cluster canonical correlation analysis (cluster-CCA) to learn discriminant low dimensional representations that maximize the correlation between the two modalities while segregating the different classes on the learned space. In order to characterize the nonlinear relationship between two sets of multi-media data, Zhang *et al.* [13] also propose a kernel-based Canonical Correlation Analysis (KCCA) to learn an isomorphic common feature subspace. However, CCA and its variants only provide a very coarse association between modalities which limits their effectiveness in cross-modal retrieval.

Deep learning. Inspired by the great success of deep neural networks in representation learning, a large number of deep learning based approaches have been proposed for cross-modal retrieval. Andrew *et al.* [26] propose deep canonical correlation analysis (DCCA) to learn complex nonlinear transformations of two views of data such that the learned representations are nonlinearly correlated. Srivastava *et al.* [27] propose deep Boltzmann machines (DBM) to learn a generative model for multi-modal data, and then use the learned model to create fused representations by combining features across modalities. Feng *et al.* [28] propose the Corr-AE algorithm which incorporates representation and correlation learning into a single process by combining autoencoder cost with correlation cost. Wang *et al.* [29] find that CCA-based approaches tend to outperform unconstrained reconstruction-based approaches, and propose deep canonically correlated autoencoder (DCCAE) which combines CCA and autoencoder. However, the above methods have not considered the label information in cross-modal retrieval.

Supervised learning. To learn a more discriminative common space, researchers leverage label information to distinguish samples from different semantic categories. Specifically, label semantics can enforce different category samples to be far apart while the same category samples to be as close as possible. Wang *et al.* [30] propose a supervised deep neural network (RE-DNN) framework to explore high-level semantic correlations across modalities. To achieve the target of enforcing pairs of image and text to have similar feature representations in the common semantic space, Li *et al.* [31] utilize Euclidean loss as the cost function to optimize both image and text networks. Wei *et al.* [32] propose a deep semantic matching method for cross-modal retrieval for samples with one or more labels. Euclidean distance between ground truth probability vector and the predicted scores over classes is used to guide the neural network optimization. Castrejon *et al.* [33] use the sum of softmax loss and negative log-likelihood as optimization objective to learn shared representation that is agnostic of modality. The negative log-likelihood is used to encourage activations in the intermediate hidden layers to have similar statistics across modalities. Zhen *et al.* [6] propose a deep supervised cross-modal retrieval (DSCMR) framework which minimize the discrimination loss in both label space and the common space. The difference between representations of all image-text pairs is minimized to eliminate the cross-modal discrepancy. Jing *et al.* [8] propose an end-to-end cross-modal framework which use cross-modal center loss (CMCL) to reduce the cross-modal discrepancy. Zhang *et al.* [7] propose a hybrid cross-modal similarity learning (HCMSL) framework for cross-modal retrieval. The method learns semantic discrimination in the label space and the common subspace, and uses weight-sharing strategy to minimise the heterogeneous gap.

Adversarial learning. Many recent deep learning based methods also use generative adversary networks (GAN) [34] for cross-modal retrieval. For example, Wang *et al.* [4] propose adversary cross-modal retrieval (ACMR) framework, which seeks an effective common subspace by using adversary training. The authors use adversary

mechanism to minimise the modality gap, and use triplet constraint to enhance the semantic discrimination ability. Peng *et al.* [35] use the cross-modal generative adversary networks (CM-GANs) to learn discriminative common representation for bridging the heterogeneous gap. The authors use cross-modal adversary mechanism to simultaneously conduct intra- and inter- modality discrimination, which can effectively correlate the heterogeneous data. Xu *et al.* [36] introduce deep adversarial metric learning (DAML) to maximize the correlations between modalities. DAML non-linearly maps labeled data pairs of different modalities into a shared latent feature subspace, under which the intra-class variation is minimized and the inter-class variation is maximized. Huang *et al.* [9] propose modal-adversarial hybrid transfer network (MHTN) which adopts a modal-sharing knowledge transfer sub-network to transfer knowledge from a single-modal source domain to the cross-modal target domain, and uses a modal-adversarial semantic learning sub-network to construct an adversarial mechanism between common representation generator and modality discriminator. Wu *et al.* [23] propose to modality-specific and modality-shared generative adversary network (MS2GAN) for cross-modal retrieval.

Proxy-based metric learning. Deep metric learning (DML) aims to learn an embedding space where semantically similar instance are close together while dissimilar instances are far apart. Due to its practical significance, it has been widely used in image retrieval [37, 38] . The core of DML is the loss function which guides the learning process of the deep neural networks. Existing DML loss can be roughly categorized into two groups: pair-based and proxy-based losses. The pair-based losses including contrastive loss [39], triplet loss [40], N-pair loss [38], and lifted-structure loss [41], directly compute the loss based on pairs of samples. However, these losses rely on sampling the pairs or triplets in the mini-batch during training. Empirically, the sampling issue will lead to prohibitively large number of tuples including many non-informative tuples, which will cause slow convergence and degraded performances. To overcome this, some researchers make efforts to use efficient methods to select the hard/semi-hard negative samples from subsets of the whole training dataset [17]. However, this kind of methods also have the weakness of failing to characterize the global geometry of the embedding space.

The idea of proxy-based metric learning is to infer a small set of proxies that capture the global structure of an embedding space and relate each data point with the proxies [42–45]. Proxy-based loss generally assigns one or more proxies to each class, and enforces each raw data point to be close to its corresponding proxy and far away from the proxies of other classes. During training, all proxies are learnable parameters shared across all samples and are kept in the memory, thus avoiding the sampling issue [42, 46]. Instead of comparing samples with one another in a batch, proxy-based method compares samples against proxies, which reduces the complexity (lowers the number of possible combinations) and is more robust to noisy samples. Furthermore, proxy can approximate the space of training set and thus can characterize the global geometry of the embedding space [42–45]. A well-known proxy-based loss is ProxyNCA [46]. ProxyNCA assigns a single proxy for each class, and encourages the positive pairs to be close and negative pairs to be far apart. There also exists serveral variants of ProxyNCA such as ProxyNCA++ [42] and Proxy Anchor [47].

In general, proxy-based losses treat DML as a classification problem by using the proxies to separate samples from different classes, and it has shown powerful ability in extracting semantic discriminative representations. However, the existing proxy-based losses are only employed to tackle the single modality data. In this paper, we propose the *shared proxy* for multi-modal data and design cross-modal shared proxy loss to learn an aligned common space for cross-modal retrieval.

3 Methods

3.1 Problem Formulation

Dataset $\mathcal{S}$ contains N instances where the i-th instance s_i has M modalities with a class label y_i. x_i^m is the m^{th} modality data of the instance s_i. Formally, the notations of the data set can be formulated as follows.

$$\mathcal{S} = \{(s_i, y_i)\}_{i=1}^{N}$$

$$s_i = \{x_i^m\}_{m=1}^M$$
$$y_i \in \{1, 2, \cdots, c\} \quad (1)$$

where c is the number of classes.

Generally speaking, there is a distribution shift across different modalities for instance s_i. In other words, the representation space for modality samples $(x_i^1, x_i^2, ..., x_i^M)$ are not aligned well (refer to Fig. 1), and thus the cross-modal similarity cannot be directly measured. Therefore the goal of cross-modal retrieval is to learn a common space $\mathcal{V}$ to which each modality sample x_i^m is mapped through $v_i^m = f_m(x_i^m, \theta_m)$, where f_m is a mapping function (or deep neural network) to be learned, and θ_m is the learnable parameters for the modality m. In the common space $\mathcal{V}$, the Euclidean distance or Cosine distance can be directly used to measure the similarity of samples.

The common space $\mathcal{V}$ should be semantically discriminative and modality invariant, which means that: the distance between features of different classes needs to be as large as possible, regardless of their modalities (modality agnostic), $d(v_i^a, v_j^b) > r, \forall a, b$ when $y_i \neq y_j$, where r is a margin, and the distance between features of the same class should be as small as possible, $d(v_i^a, v_j^b) \to 0$, $\forall a, b$ when $y_i = y_j$.

3.2 Loss Function and Optimization

3.2.1 Cross-Modal Shared Proxy Loss

Existing proxy-based loss functions are mainly designed for single modality task, which usually ignore the relation among proxies of different constituent modalities. In the case of multi-modal data, when the learned common space is semantically discriminative and modality invariant, the samples with the same class label, even if they are from different modalities, should have the same cluster distribution. Motivated by this rationale, we introduce a new concept - *shared proxy*, to approximate the global structure of the common embedding space where the distribution of different constituent modalities can be aligned. Different from the single modality scenario, where a proxy is only for one modality, the shared proxy is a proxy that is shared by all modalities. The shared proxies have its innate advantage in aligning heterogeneous data and characterizing the global distribution of the data. A toy example for explaining the traditional proxy and shared cross-modal proxy is given in Fig. 1.

Given a dataset $\{x_i^m\}_{i=1}^N (m \in [1, M])$ for N instances and M modalities, $y_i \in \{1, 2, \cdots, c\}$ is the label for $s_i = \{x_i^m\}_{m=1}^M$. By introducing *one shared proxy for each class*, we propose the cross-modal shared proxy loss (CMSP) to overcome the weakness of traditional pair-/triplet based loss. A shared proxy is a k dimensional vector in the common embedding space. Let $p_i \in \mathbf{R}^k, i = 1, 2, ...c$, be the shared proxy for class y_i. CMSP is defined in (2) and it minimizes the distance between the feature representation of an input and its class's shared proxy, and maximize the distances between a feature representation and other proxies.

$$\mathcal{L}_{cmsp} = -\sum_{i=1}^{N} \log \left(\frac{1}{M} \sum_{m=1}^{M} \frac{\exp(-d(v_i^m, p_{y_i}) - \delta)}{\sum_{j \neq y_i} \exp(-d(v_i^m, p_j))} \right) \quad (2)$$

where v_i^m is the common space representation for x_i^m, $p_{y_i} \in \mathbf{R}^k$ represents the corresponding shared proxy of instance s_i, k is the dimension of the common embedding space, and $\delta > 0$ is a margin.

3.2.2 Cross-entropy Loss

To ensure that the intra-modal discrimination in data is preserved after feature projection, a classifier is deployed to predict the semantic labels of the items projected in the common subspace. We use multilayer perceptron (MLP) to predict the label of the samples projected in the common space. The weight of MLP is shared by all modalities. Given features $\{v_i^m\}_{i=1}^N$, $m \in [1, M]$ for N instance and M modalities, the predicted probability $\hat{y_i^m}$ can be calculated by $\hat{y_i^m} = MLP(v_i^m)$. Accordingly, the cross entropy loss can be defined as,

$$\mathcal{L}_d = -\frac{1}{N} \left(\sum_{i=1}^{N} \sum_{m=1}^{M} \overline{y}_i^m \cdot \log \left(\hat{y_i^m} \right) \right) \quad (3)$$

where $\overline{y}_i^m \in \{0, 1\}^c$ is the one-hot encoding of the ground truth label y_i.

Using cross-entropy training can learn the decision boundary of different types of features in the common space. The joint training of cross-entropy and cross-modal shared proxy loss can

not only optimize the feature distribution of the common space, increase the discrimination capabilities of different types of features, but also learn modal-invariant features.

3.2.3 Mean Square Error Loss

To further eliminate the discrepancy between modalities, we use the mean square error to minimize the Euclidean distance between representations from different modalities in the common space. Technically, we formulate the modality invariance loss as follows,

$$\mathcal{L}_m = \sum_{\alpha,\beta\in[1,M],\alpha\neq\beta} \left\| v_i^{\alpha} - v_i^{\beta} \right\|_2^2 \qquad (4)$$

3.2.4 Joint Training

Combining equations (2), (3) and (4), we get the total optimization objective of the proposed COXI framework.

$$\mathcal{L}oss = \alpha_{cmsp}\mathcal{L}_{cmsp} + \alpha_d\mathcal{L}_d + \alpha_m\mathcal{L}_m \qquad (5)$$

where $\alpha_{cmsp}, \alpha_d, \alpha_m$ are hyperparameters used to control the proportion of each loss function. Our proposed joint loss function can be optimized by stochastic gradient descent. Through the joint training of the three loss functions, it is possible to learn the discriminative and modal-invariant features. The details of the optimization procedure is summerized in Algorithm 1.

3.3 Framework Architecture

Our goal is to learn a common subspace under which the class-wise representations are modality invariant and discriminative. The overall architecture for our model is given in Fig. 3. As shown in Fig. 3, there are two subnetworks, the 19-layer VGGNet [48] for image feature extraction, and the sentence CNN [49] for text feature extraction. In the image part, the 4096 dimension vector generated by fc7 layer is selected as the feature of the image. The text part uses Bag of Words (BoW)/Word2Vec [50] to extract the feature vector of the text. Two fully-connected layers followed the feature extraction subnetworks: the first layer has 2048 dimensions, the second layer has 512 dimensions with weight-sharing to learn the cross-modal consistency. Three losses are jointly trained with Adam [51].

4 Experiments

In this section, we present the experiments for verfying the effectiveness of the proposed COXI. The implementation details are first introduced, and then we discuss the datasets, evaluation metrics and the 15 competing methods. Next, experimental results along with analysis are presented. In addition, we also present comprehensive experimental analysis, including the role of cross-modal shared proxy loss, the covergence analysis and the analysis of the margin paramters.

4.1 Datasets

We conduct a set of experiments on four widely used benchmark datasets: the Wikipedia [52], the Pascal Sentence [53], the NUS-WIDE-10k [54] and PKU XMedia [55].

- The *Wikipedia* dataset contains 2866 image/-text pairs corresponding to 10 categories, including history, biology and so on. Following [6], the dataset is divided into 3 parts, 2173 pairs used for training, 231 pairs for validation, and 462 pairs for testing.
- Pascal Sentence contains 1000 images which are evenly categorized into 20 categories, and image has five corresponding English sentences which make up one document. Following [5], 800 documents are selected as training, 100 documents are selected as validation, and 100 documents are selected for testing.
- The NUS-WIDE-10k dataset has a total of 10000 image/text pairs selected evenly from 10 largest categories of NUS-WIDE dataset. Following the partition [5], this dataset is also split into 3 parts: 8000 image/text for training, 1000 pairs for validation and the rest 1000 pairs for testing.
- PKU XMedia is a cross-modal dataset with five modalities, i.e., text, image, video, audio and $3D$ model. There are 20 classes, 250 image-text pairs, 25 videos, 50 audio clips and 35 $3D$ models. Following [9], this dataset is split into three parts: 9600 instances for training, 1200 instances for validation and 1200 instances for testing. In a specific cross-modal retrieval task,

Algorithm 1: The optimization procedure of COXI

Input: The number of modalities M, the training data set $\mathcal{S} = \{s_i\}_{i=1}^{N}$, the label $Y = (y_1, y_2, ..., y_c)$, the batch size n_b, the features dimension of common space d, the learning rate of the model γ, the maximal number of epochs $\mathcal{T}$, the hyperparameters of the loss function $\alpha_{cmsp}, \alpha_d, \alpha_m$, the margin δ, and the learning rate of the proxy γ_p.

Output: The optimized parameters in all sub-networks θ.

Initialization: Randomly initialize M subnetworks parameters $\theta_m, m = 1, \cdots, M$, the MLP parameters θ_p, and the proxies $P = \{p_1, p_2, ..., p_c\}$;

for $t = 1, 2, ..., \mathcal{T}$ **do**

  for $l = 1, 2, 3, ..., [\frac{n}{n_b}]$ **do**

    Randomly sampled mini-batch n_b in the dataset $\mathcal{S}$;

    Learning feature representation through each branch network, learning v_i^m through x_i, where $n \in [1, M]$, and $i \in [1, n_b]$;

    For each v_i^m, learning classification prediction $y_i^{\hat{m}}$ by linear classifier $y_i^{\hat{m}} = MLP(v_i^m)$;

    Calculate the loss of each mini-batch by the Eq.(5);

    Update the parameters of linear classifier;

    Update the parameters of all subnetworks;

    Update all proxies;

  end

end

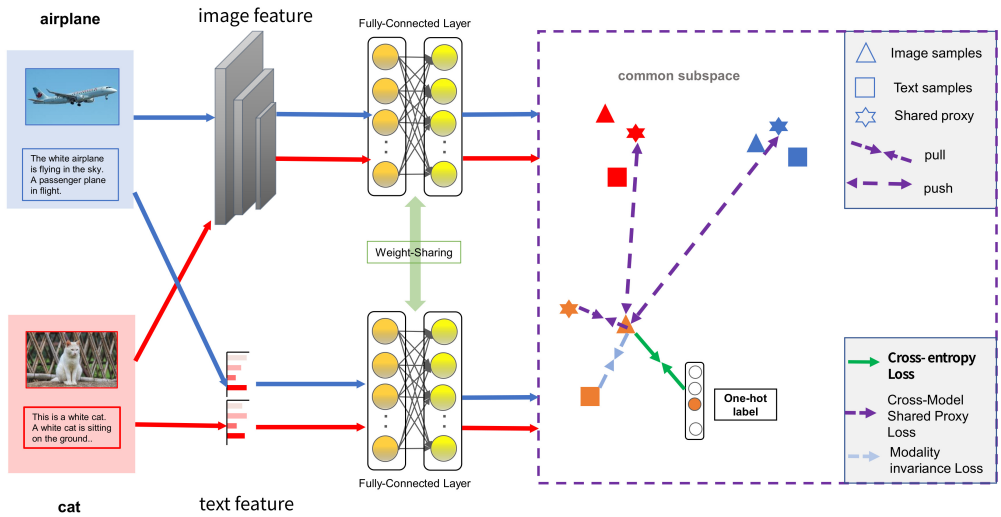


Fig. 3 Network achitecture of the COXI framework. Three losses are jointly trained to learn a modality-invariant and semantically-discriminative feature space.

following [9], the training set with 4000 image-text pairs, validation set with 500 pairs, and testing set with 500 pairs are used.

For the PKU XMedia dataset, we use the image and text features provided by [55]. As for the Pascal Sentence, Wikipedia and NUS-WIDE-10K datasets, following [5, 6], we use a 4096-demensional vector extracted by the fc7 layer of VGG-19 [56] to represent each image. For text modality, we use 1000-dimentional bag-of-words

(BoW) and 300-dimentional Word2Vec model to extact feature vectors. The statistical results of the four datasets are summarised in Table 1.

Table 1 General statistics of the four datasets used in the experiments, where "*/*/*" in the "Instance" column stands for the number of training/validation/test subsets.

Dataset	Label	Modality	Instance	Feature
Wikipedia	10	Image	2,173/231/462	4,096D VGG
		Text	2,173/231/462	300D Doc2Vec
Pascal Sentence	20	Image	800/100/100	4,096D VGG
		Text	800/100/100	300D Doc2Vec
NUS-WIDE-10K	10	Image	8,000/1000/1000	4,096D VGG
		Text	8,000/1000/1000	1,000D BoW
PKU XMedia	20	Image	4,000/500/500	4,096D VGG
		Text	4,000/500/500	3,000D BoW

4.2 Retrieval tasks and evaluation metrics

To evaluate the performance of the proposed COXI, we perform cross modal retrieval tasks as retrieving one modality by another modality query, such as retrieving text by image query (Image2Text) and retrieving image by text query (Text2Image).

We adopt the mean average precision (mAP) which is a widely used criterion to evaluate the performance of COXI. In addition, we also adopt precision-recall (PR) curve for more comprehensive evaluation, which shows the search precision at all recall levels.

4.3 Comparison with other state of the art methods

We compare COXI with state of the art methods including five traditional methods, namely CMCP [57], CCA [10], MCCA [58], JGRHML [59] and JRL [60], as well as ten DNN based methods, namely DCCA [26], CMDN [61], CCL [5], DCCAE [29], ACMR [4], DSCMR [6], MS2GAN [23], DAVAE [62], CMCL [8] and HCMSL [7]. Especially, ACMR [4] and MS2GAN [23] are GAN based methods.

Methods CMCP, JGRHML, JRL, CMDN, CCL, ACMR, DSCMR, SDML, CMCL[1], CCA, MCCA, DCCA and DCCAE were implemented based on code made available by the authors. We could not find the publicly available code of MS2GAN and HCMSL, so we implemented these methods based on Pytorch. The parameters for baselines are set according to the suggestion provided in the original papers. For a fair comparison, all the tested approaches use the same image features and the same text features in all the experiments. Table (2) reports the mAP scores (average of 30 runs with the same random seed) of two cross modal retrieval tasks (*i.e.*, Image2Texts and Text2Images) and their average results (Average) on the four datasets.

In general, we have some basic oberservations from Table (2): 1) As a classical baseline method, CCA and its variants are the most basic association algorithm for cross-modal retrieval, the mAP scores are not very high, which maybe because it doesn't leverage the class label information; 2) Deep learning based algorithms such as CDMN, CCL, DCCA, ACMR, etc, achieves better mAP performance than traditional methods. Especially, the latest algorithms such as DSCMR, MS2GAN, SDML and HSMSL have superior performances, which have a large improvements in terms of mAP scores. These oberservations are consistent with the general expectation.

On all four datasets, our proposed COXI achieves the hightest mAP scores. On Wikipedia dataset, the highest mAP score of the competing methods is MS2GAN, and COXI outperforms MS2GAN by 4.3% for Image2Text, 2.9% for Text2Image, and 3.7% for average. On the Pascal Sentence dataset, COXI outperforms the best competitor (*i.e.*, SDML) by 1.1% for Image2Text, 0.13% for Text2Image, and 0.55% for average. On NUS-WIDE-10k dataset, the best competitor are HSCML and DSCMR, and COXI achieves the improvement by 2.2% and 1.8% for Image2Text and Text2Image respectively. On PKU XMedia dataset, COXI achieves the improvement of 0.44% for Image2Texts, 0.7% for Text2Image, and 0.55% for average compared with the best results of literature (*i.e.*, SDML). In conclusion, Table (2) indicates that our COXI is an effective multi-modal representation learning approach for cross-modal retrieval across image and text.

In addition to mAP score, Fig. 4 shows the precision-recall curve of Image2Text and Text2Image for the four datasets. From the figures, we can see that the precision-recall evaluations are consistent with the mAP scores for

[1]Authors provided the code for $3D$ shape retrieval, and we adapt the code for image-text cross modal retrieval.

Table 2 The mAP score for COXI and other competing methods on the four benchmark dataset. The highest score is shown in boldface.

Method	Wikipedia			Pascal Sentence			NUS-WIDE-10k			XMedia		
	Image2Text	Text2Image	Average	Image2Text	Text2Image	Average	Image2Text	Text2Image	Average	Image2Text	Text2Image	Average
CMCP [57]	0.326	0.251	0.289	—	—	—	0.368	0.327	0.348	0.708	0.701	0.705
CCA[10]	0.249	0.195	0.222	0.225	0.227	0.226	0.378	0.394	0.386	0.510	0.499	0.505
MCCA [58]	0.341	0.307	0.324	0.664	0.689	0.677	0.448	0.462	0.455	—	—	—
JGRHML [59]	0.239	0.256	0.293	—	—	—	0.346	0.325	0.336	0.417	0.312	0.365
JRL [60]	0.449	0.418	0.434	0.527	0.534	0.531	0.586	0.598	0.592	0.751	0.741	0.746
ACMR [4]	0.477	0.434	0.456	0.671	0.676	0.673	0.603	0.581	0.592	0.898	0.889	0.894
MS2GAN [23]	0.534	0.484	0.509	0.634	0.629	0.632	0.556	0.548	0.552	0.894	0.911	0.903
CMDN [61]	0.487	0.427	0.457	0.544	0.526	0.535	0.492	0.515	0.504	0.794	0.805	0.800
DCCA [26]	0.444	0.396	0.420	0.678	0.677	0.678	0.532	0.549	0.540	0.885	0.899	0.892
DCCAE [29]	0.435	0.385	0.410	0.680	0.671	0.675	0.511	0.540	0.525	0.884	0.830	0.857
CCL [5]	0.504	0.457	0.481	0.576	0.561	0.569	0.506	0.535	0.521	0.827	0.813	0.820
DSCMR [6]	0.521	0.478	0.499	0.710	0.722	0.716	0.639	0.612	0.626	0.832	0.809	0.821
SDML [63]	0.520	0.488	0.504	0.712	0.723	0.718	0.619	0.605	0.612	0.909	0.917	0.913
CMCL [8]	0.429	0.441	0.435	0.682	0.677	0.680	0.616	0.606	0.611	0.819	0.831	0.825
HCMSL[7]	0.527	0.478	0.503	0.699	0.712	0.710	0.646	0.611	0.629	0.903	0.892	0.898
COXI($\mathcal{L}_{cmsp}$)	0.530	0.479	0.504	0.698	0.701	0.699	0.632	0.572	0.602	0.910	0.913	0.912
COXI	**0.557**	**0.498**	**0.528**	**0.720**	**0.724**	**0.722**	**0.660**	**0.623**	**0.642**	**0.913**	**0.923**	**0.918**

cross-modal retrieval tasks, where our COXI outperforms all the state of the art methods.

Our propsoed COXI framework consists of three loss functions, namely the cross-modal shared proxy loss (CMSP), discrimination loss (cross-entropy) and modality invariance loss (mean-square error), among which cross-modal shared proxy loss is the major component of the proposed framework. We also present the results of COXI($\mathcal{L}cmsp$) which only uses the cross-modal shared proxy loss (CMSP) as the optimization objective in Table (2). On Wikipedia dataset, we can see that COXI($\mathcal{L}cmsp$) achieves a bit lower mAP score than MS2GAN on Image2Text, a bit lower than MS2GAN and CMCL on Text2Image, and a little smaller mAP score than MS2GAN on Average. But COXI($\mathcal{L}cmsp$) beats all other competing algorithms. On the Pascal Sentence, NUS-WIDE-10K and PKU XMedia datasets, similar results can be found. Therefore, we can conclude that COXI($\mathcal{L}cmsp$) has an advantage over other 15 state of the art methods.

Comparison with pre-trained vision-language model: The recently pre-trained vision-language models based on the Transformer structure [64] have shown a significant performance gain for multi-modal tasks such as visual question answering (VQA), image-text retrieval, and visual entailment. We conduct cross-modal retrieval by embedding images and texts into the common space with CLIP (Contrastive Language-Image Pretraining) [65] which is a state of the art neural network pre-trained for image-text pairs. On Pascal Sentence dataset, the mAP scores for Image2Text and Text2Image are 0.553 and 0.546 respectively, which is smaller than COXI as well as many state of the art methods presented in Table 2. The reason is that the COXI and other competing methods are supervised method which can leverage the label information while pretrained model such as CLIP is a self-supervised method.

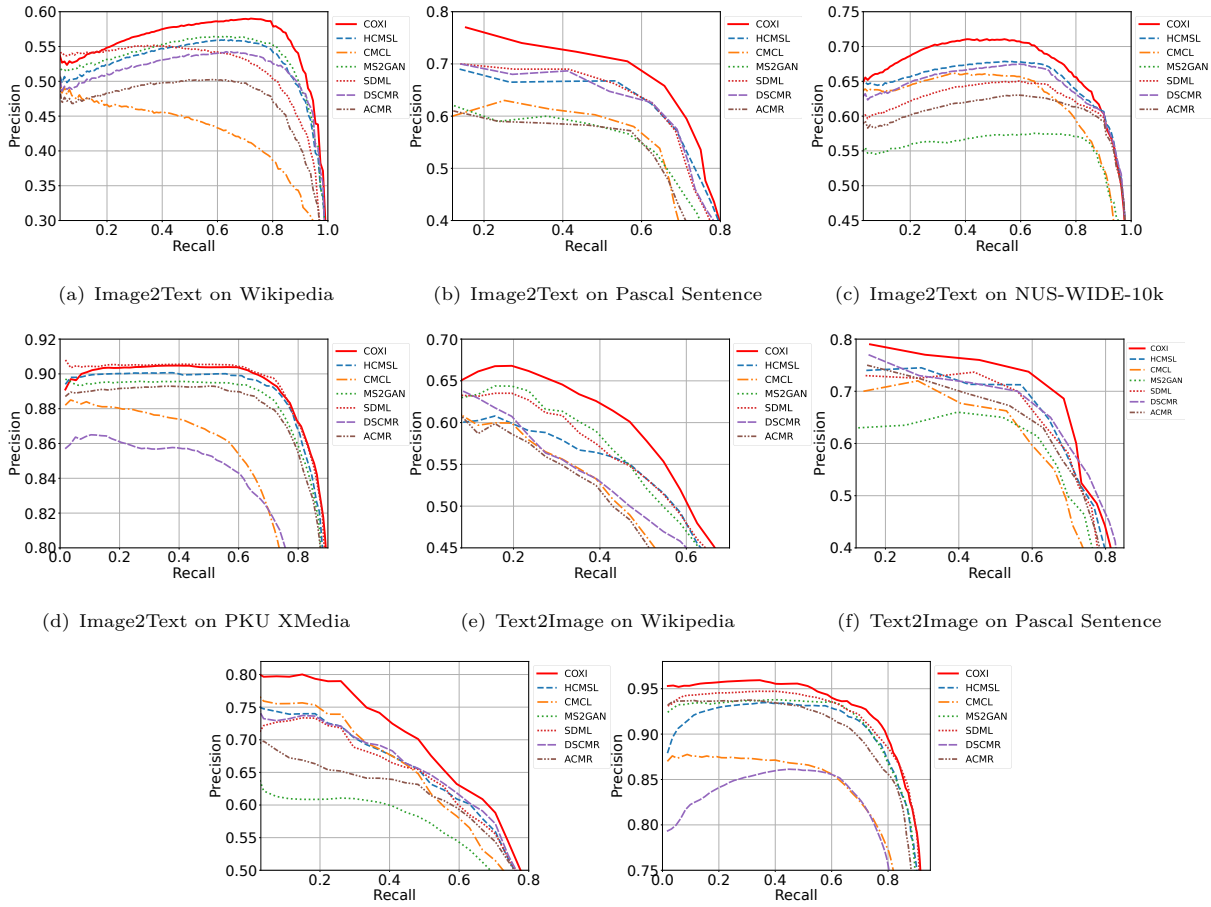
4.4 Impact of Different Components

As shown in the Eq.(5), the COXI framework is jointly trained by three loss items. $\mathcal{L}_{cmsp}$ is the cross-modal shared proxy loss, which is used to learn cross-modal similarity features and optimize the spatial structure. $\mathcal{L}_d$ is discrimination loss in common space which learns semantically-discriminative features. Mean-square loss is used to reduce the difference between modalities and learn modality-invariant representations. In the proposed COXI framework, $\mathcal{L}_{cmsp}$ is the most important component loss function. To investigate the impact of other two terms ($\mathcal{L}_d$ and $\mathcal{L}_m$) on COXI, we designed ablation experiments which are listed as follows.

- task1: optimization with only cross-modal shared proxy loss - $\mathcal{L}_{cmsp}$.
- task2: jointly optimization with $\mathcal{L}_{cmsp}+\mathcal{L}_m$ (no $\mathcal{L}_d$).
- task3: jointly optimization with $\mathcal{L}_{cmsp}+\mathcal{L}_d$. (no L_m).
- task4: jointly optimization with $\mathcal{L}_{cmsp}$, $\mathcal{L}_d$ and $\mathcal{L}_m$.

The training hyperparameters remain the same in all ablation experiments.

As previously shown in Table 2, single $\mathcal{L}_{cmsp}$ (task1) has a better mAP socre than many state of the art methods on all four benchmark datasets, which indicate the powerful ability of $\mathcal{L}_{cmsp}$. This may be due to the fact that our cross-modal proxy



(a) Image2Text on Wikipedia (b) Image2Text on Pascal Sentence (c) Image2Text on NUS-WIDE-10k

(d) Image2Text on PKU XMedia (e) Text2Image on Wikipedia (f) Text2Image on Pascal Sentence

(g) Text2Image on NUS-WIDE-10k (h) Text2Image on PKU XMedia

Fig. 4 The precision-recall curves of cross-modal retrieval tasks on the four benchmark datasets.

loss not only considers distribution alignment, but also semantic alignment.

Table (3) shows the ablation experiments on Wikipedia, Pascal Sentence, NUS-WIDE-10k and PKU XMedia. From the table, we can see that the complete version of COXI (task4) achieves 5.1%, 1.9% and 4.8% improvement over $\mathcal{L}_{cmsp}$ (task1) on Wikipedia dataset for Image2Text, Text2Image and Average, respecvitvely. Similarly, the complete version of COXI (task4) also outperforms task1 with a large margin on Pascal Sentence, NUS-WIDE-10k and PKU XMedia, respectively. The complete version of COXI (task4) also achieves better mAP score than that of task2 and task3 on all four datasets, indicating that $\mathcal{L}_d$ and $\mathcal{L}_m$ also have a certain role in learning cross-modal features. All in all, we can see that all three loss items in the objective function play positive role in improving the cross-modal retrieval quality.

4.5 Convergence performance

Fig. 5 plots the loss versus different number of epochs during the training process on four benchmark datasets. From the figure, we can see that the loss shows a downward trend and the value of the objective function decreases monotonously and converges smoothly. Especially, the COXI covnerges quickly within 150 iterations on PKU XMedia dataset, which shows its efficiency. On Wikepedia, Pascal Sentence and NUS-WIDE datasets, the value of the objective function decrease monotonously and converges in 200 $\sim$ 300 epochs.

Table 3 Ablation experiments on the test dataset. The highest score is shown in boldface.

MAP	Wikipedia			Pascal Sentence			NUS-WIDE			Xmedia		
	Image2Text	Text2Image	Average	Image2Text	Text2Image	Average	Image2Text	Text2Image	Average	Image2Text	Text2Image	Average
task1: L_{cmsp}	0.530	0.479	0.504	0.698	0.701	0.699	0.632	0.572	0.602	0.910	0.913	0.912
task2: $L_{cmsp} + L_d$	0.523	0.480	0.502	0.696	0.708	0.702	0.643	0.580	0.612	0.909	0.914	0.912
task3: $L_{cmsp} + L_m$	0.532	0.480	0.506	0.702	0.700	0.701	0.636	0.584	0.610	0.853	0.838	0.846
task4: $L_{cmsp} + L_d + L_m$	**0.557**	**0.498**	**0.528**	**0.720**	**0.724**	**0.722**	**0.653**	**0.632**	**0.642**	**0.913**	**0.923**	**0.918**

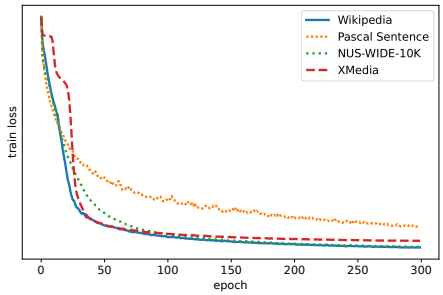
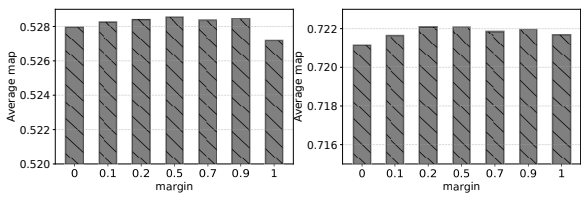


Fig. 5 The normalized loss values during the training process on four datasets.

4.6 Sensitivity study of margin

To investigate the sensibility of margin δ on the retrieval results, we conduct experiments on the Wikipedia and Pascal Sentence datasets with different margin (0, 0.1, 0.2, 0.5, 0.7, 0.9, 1). All models are trained with the same epochs and parameters. The results are shown in Fig. 6 which shows that $\delta = 0.5$ achieves the best mAP score on Wikiedia and $\delta = 0.2$ achives the best mAP on Pascal Sentence. The margin value for the two datasets has some impact on the mAP scores, but the impact is not very significant. This shows that COXI algorithm is not overly sensitive to this parameter.



(a) Wikipedia (b) Pascal Sentence

Fig. 6 Performance of different margin on Wikipedia datasets and Pascal Sentence datasets.

4.7 Impact of different dimension of the common space

To investigate the impact of the common embedding space's dimension on the retrieval performance, we conduct experiments on the four datasets with different dimension (16, 32, 64, 128, 256, 512 and 1024). The mAP results are shown in Fig. 7. From the figure, we can see that: 1) there is no significant difference when the dimension ranges from 128 to 1024; 2) the retrieval result is poor when the dimension is smaller than 64. In this paper, we set the $k = 512$ for COXI.

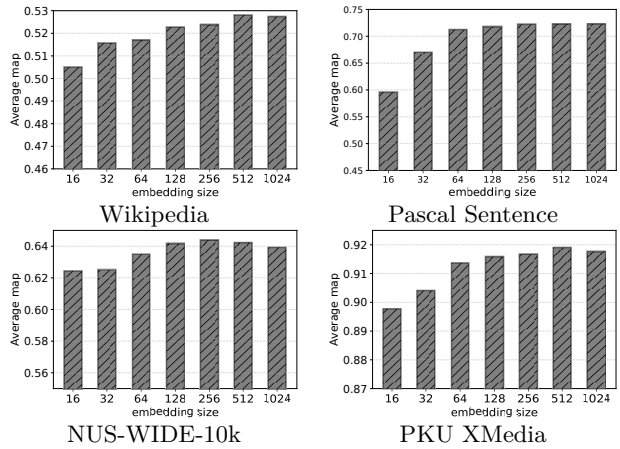
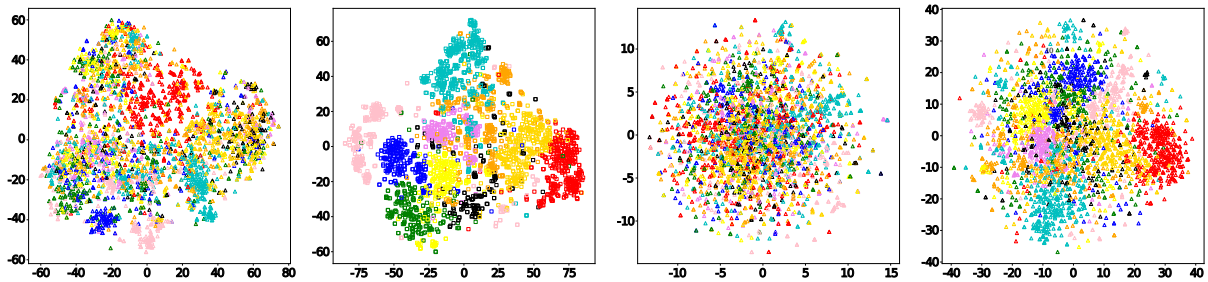


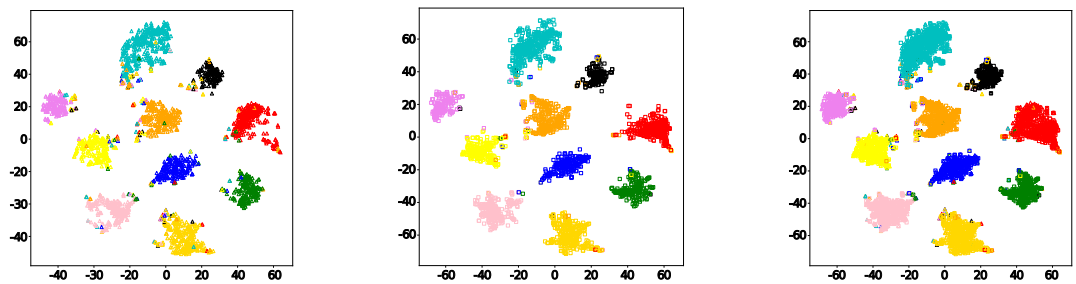
Fig. 7 Performance of different embedding dimension on four datasets.

4.8 Visualisation of the Learned Representation

To visually investigate the effectiveness of the proposed COXI, we employ the t-SNE method [66] to embed the learned features of different modalities into a 2D plane. Fig. 8(a), 8(b), 8(e), 8(f) and 8(g) separately show the distributions of the original image samples, *i.e.*, the samples represented by 4,096-dimensional VGGNet features for image, the distribution of the original text samples, *i.e.*, the samples represented by 3,000-dimensional BoW features for text, the distribution of the learned image feature representations in the common space, the distribution of the learned text feature representations in the common space and the distribution of mixed representations (image and text) in the common space, of the Wikipedia

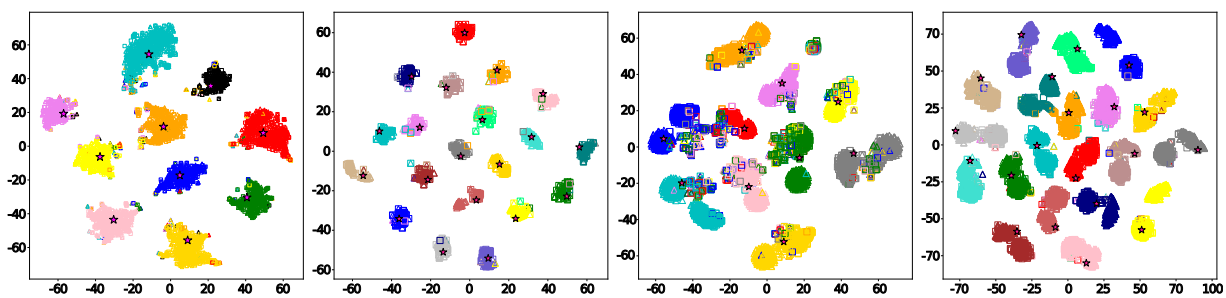


(a) Original image samples (b) Original text samples (c) Image samples in the random subspace (d) Text samples in the random subspace



(e) Image samples in the common subspace (f) Text samples in the common subspace (g) Image and text samples in the common subspace

Fig. 8 The visualisation for the Wikipedia dataset by using the t-SNE method. (a) the original image samples represented by the 4, 096-dimensional features. (b) the original text samples represented by the 300-dimensional features. (c) the image samples in the random space. (d) the text samples in the random subspace. (e) the image samples in the learned common subspace. (f) the text samples in the learned common subspace. (g) the image and text samples in the learned common subspace. Triangles and squares are image and text modalities respectively. Fig. 8(g) is obtained by superimposing Fig. 8(e) onto Fig. 8(f). Best viewed in color.



(a) Wikipedia (b) Pascal Sentence (c) NUS-WIDE-10k (d) PKU XMedia

Fig. 9 The visualisation of learned common space for the Wikipedia dataset by using the t-SNE method. The stars are learned shared proxies. Triangles and squares are image and text modalities respectively. Best viewed in color.

dataset. Meanwhile, we show the t-SNE embedding of ramdon subspace which has the same dimension as the learned subspace in Fig. 8(c) and Fig. 8(d).

From Fig. 8(a) and 8(b), we can see that the distribution of image and text features are very different and the samples from different modalities cann't be seperated in the original feature space. On the random subspace, it is impossible to find any clusters. In contrast, Fig. 8(e), 8(f) and 8(g) show that the learned representation in the common space are better semantically clustered. Especially, Fig. 8(g) shows that the distribution

of image and text samples can be semantically aligned in the common space. In summary, the COXI framework can build a common space where *multi-modal data can be aligned well according to their semantics*.

Fig. 9 shows the t-SNE visulization of learned shared proxies for all four datasets. We can see from the figure that: 1) In the learned common space, the samples with the same category label, even if they are from different modalities, have the same cluster distribution; 2) The shared proxies (red stars in each subfigure) can characterize the global structure of the common embedding space, and can be used to represent class distributions.

5 Conclusion

In cross-modal retrieval, existing methods usually use inter-/intra- modal pairwise or triplet data relationships for common space learning, which only capture the data similarity locally and fail to characterize the global geometry of the common embedding space. Based on the assumption that there exists a common space where samples with the same category label, even if they are from different modalities, have the same cluster distribution, we introduce a new terminology - *shared proxy*, to characterize the global structure of the common embedding space of multi-modal data. We have designed a cross-modal shared proxy (CMSP) loss to minimize the distance between the feaure representation and the corresopnding shared proxy, and maximize the distance between the feature representation and other proxies greater than a margin. In order to learn a semantically discriminative and modality invariant common space for cross-modal retrieval, we propose a novel supervised cross-modal retrieval framework COXI which optimizes the sum of cross-modal shared proxy loss, discrimination loss and modality invariance loss. Extensive experiments on four benchmark datasets clearly show that COXI outperforms state of the art cross-modal retrieval techniques.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

The Wikipedia [55], the Pascal Sentence [52], the NUS-WIDE-10k [53] and the PKU XMedia [54] datasets can be openly available. We give the links on https://github.com/LigangZheng/COXI.

Acknowledgments This work is supported by the Natural Science Foundation of China (U1936116), and the Science and Technology Projects in Guangzhou (202102010412).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

[1] Arya, D., Rudinac, S., Worring, M.: HyperLearn: A distributed approach for representation learning in datasets with many modalities. In: Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, pp. 2245–2253. ACM, Nice, France (2019)

[2] Lan, R., Tan, Y., Wang, X., Liu, Z., Luo, X.: Label guided discrete hashing for cross-modal retrieval. IEEE Transactions on Intelligent Transportation Systems **23**(12), 25236–25248 (2022)

[3] Cheng, Q., Tan, Z., Wen, K., Chen, C., Gu, X.: Semantic pre-alignment and ranking learning with unified framework for cross-modal retrieval. (2023)

[4] Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: Liu, Q., Lienhart, R., Wang, H., Chen, S.K., Boll, S., Chen, Y.P., Friedland, G., Li, J., Yan, S. (eds.) Proceedings of the 2017 ACM on Multimedia Conference, MM, pp. 154–162. ACM, Mountain View, CA, USA (2017). https://doi.org/10.1145/3123266.3123326

[5] Yuxin, P., Jinwei, Q., Xin, H., Yuxin, Y.: CCL: cross-modal correlation learning with multigrained fusion by hierarchical network. IEEE Transactions on Multimedia **20**(2), 405–420 (2017). https://doi.org/10.1109/TMM.2017.2742704

[6] Zhen, L., Hu, P., Wang, X., Peng, D.: Deep supervised cross-modal retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pp. 10394–10403. Computer Vision Foundation / IEEE, Long Beach, CA, USA, (2019). https://doi.org/10.1109/CVPR.2019.01064

[7] Chengyuan, Z., Jiayu, S., Xiaofeng, Z., Lei, Z., Shichao, Z.: HCMSL: Hybrid cross-modal similarity learning for cross-modal retrieval. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **17**(1s), 1–22 (2021). https://doi.org/10.1145/3412847

[8] Jing, L., Vahdani, E., Tan, J., Tian, Y.: Cross-modal center loss for 3d cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3142–3151 (2021). https://doi.org/10.1109/cvpr46437.2021.00316

[9] Huang, X., Peng, Y., Yuan, M.: MHTN: modal-adversarial hybrid transfer network for cross-modal retrieval. IEEE Trans. Cybern. **50**(3), 1047–1059 (2020)

[10] Harold, H.: Relations between two sets of variates. Breakthroughs in statistics: methodology and distribution, 162–190 (1936). https://doi.org/10.1093/biomet/28.3-4.321

[11] Sharma, A., Kumar, A., DauméIII, H., Jacobs, D.W.: Generalized multiview analysis: A discriminative latent space. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2160–2167. IEEE Computer Society, Providence, RI, USA (2012). https://doi.org/10.1109/cvpr.2012.6247923

[12] Gong, Y., Ke, Q., Isard, M., Lazebnik, S.: A multi-view embedding space for modeling internet images, tags, and their semantics. Int. J. Comput. Vis. **106**(2), 210–233 (2014)

[13] Zhang, H., Liu, Y., Ma, Z.: Fusing inherent and external knowledge with nonlinear learning for cross-media retrieval. Neurocomputing **119**, 10–16 (2013). https://doi.org/10.1016/j.neucom.2012.03.033

[14] Yuan, L., Wang, T., Zhang, X., Tay, F.E.H., Jie, Z., Liu, W., Feng, J.: Central similarity quantization for efficient image and video retrieval. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, pp. 3083–3092. Computer Vision Foundation / IEEE, Seattle, WA, USA (2020)

[15] Chen, Y., Lai, Z., Ding, Y., Lin, K., Wong, W.K.: Deep supervised hashing with anchor graph. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, pp. 9796–9804. IEEE, Seoul, Korea (South) (2019). https://doi.org/10.1109/iccv.2019.00989

[16] Carvalho, M., Cadène, R., Picard, D., Soulier, L., Thome, N., Cord, M.: Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, pp. 35–44. ACM, Ann Arbor, MI,USA (2018)

[17] Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pp. 5022–5030. Computer Vision Foundation / IEEE, Long Beach, CA, USA (2019)

[18] Goldberger, J., Roweis, S.T., Hinton, G.E., Salakhutdinov, R.: Neighbourhood components analysis. In: Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, Vancouver, British Columbia, Canada, pp. 513–520 (2004)

[19] Lin, Z., Ding, G., Hu, M., Wang, J.: Semantics-preserving hashing for cross-view retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, pp. 3864–3872. IEEE Computer Society, Boston, MA, USA (2015). https://doi.

org/10.1109/CVPR.2015.7299011

[20] Jiang, Q., Li, W.: Deep cross-modal hashing. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 3270–3278. IEEE Computer Society, Honolulu, HI, USA, (2017). https://doi.org/10.1109/CVPR.2017.348

[21] Lin, Q., Cao, W., He, Z., He, Z.: Semantic deep cross-modal hashing. Neurocomputing **396**, 113–122 (2020). https://doi.org/10.1016/j.neucom.2020.02.043

[22] Su, S., Zhong, Z., Zhang, C.: Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, pp. 3027–3035. IEEE, Seoul, Korea (South), (2019). https://doi.org/10.1109/ICCV.2019.00312

[23] Wu, F., Jing, X., Wu, Z., Ji, Y., Dong, X., Luo, X., Huang, Q., Wang, R.: Modality-specific and shared generative adversarial network for cross-modal retrieval. Pattern Recognition **104**, 107335 (2020). https://doi.org/10.1016/j.patcog.2020.107335

[24] Ranjan, V., Rasiwasia, N., Jawahar, C.V.: Multi-label cross-modal retrieval. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, pp. 4094–4102. IEEE Computer Society, Santiago, Chile (2015). https://doi.org/10.1109/iccv.2015.466

[25] Rasiwasia, N., Mahajan, D., Mahadevan, V., Aggarwal, G.: Cluster canonical correlation analysis. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, vol. 33. Reykjavik, Iceland,, pp. 823–831 (2014). PMLR

[26] Andrew, G., Arora, R., Bilmes, J.A., Livescu, K.: Deep canonical correlation analysis. In: Proceedings of the 30th International Conference on Machine Learning. JMLR Workshop and Conference Proceedings, vol. 28, pp. 1247–1255. Atlanta, GA, USA (2013)

[27] Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. In: Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, Lake Tahoe, Nevada, United States, pp. 2231–2239 (2012). Citeseer

[28] Fangxiang, F., Xiaojie, W., Ruifan, L.: Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the ACM International Conference on Multimedia, MM, pp. 7–16. ACM, Orlando, FL, USA (2014). https://doi.org/10.1145/2647868.2654902

[29] 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, ICML 2015. JMLR Workshop and Conference Proceedings, vol. 37, pp. 1083–1092. JMLR.org, Lille, France, (2015)

[30] Wang, C., Yang, H., Meinel, C.: Deep semantic mapping for cross-modal retrieval. In: 27th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2015, pp. 234–241. IEEE Computer Society, Vietri sul Mare, Italy (2015). https://doi.org/10.1109/ictai.2015.45

[31] Li, Z., Lu, W., Bao, E., Xing, W.: Learning a semantic space by deep network for cross-media retrieval. In: The 21st International Conference on Distributed Multimedia Systems, pp. 199–203. Knowledge Systems Institute, Vancouver, Canada, August (2015). https://doi.org/10.18293/dms2015-005

[32] 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. **47**(2), 449–460 (2017). https://doi.org/10.1109/tcyb.2016.2519449

[33] Castrejón, L., Aytar, Y., Vondrick, C., Pirsiavash, H., Torralba, A.: Learning aligned cross-modal representations from weakly aligned data. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, pp. 2940–2949. IEEE Computer Society, Las Vegas, NV, USA (2016). https://doi.org/10.1109/cvpr.2016.321

[34] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, Quebec, Canada, pp. 2672–2680 (2014). https://doi.org/10.3156/jsoft.29.5_177_2

[35] Peng, Y., Qi, J.: CM-GANs: Cross-modal generative adversarial networks for common representation learning. ACM Trans. Multim. Comput. Commun. Appl. **15**(1), 22–12224 (2019). https://doi.org/10.1145/3284750

[36] Xu, X., He, L., Lu, H., Gao, L., Ji, Y.: Deep adversarial metric learning for cross-modal retrieval. World Wide Web **22**(2), 657–672 (2019). https://doi.org/10.1007/s11280-018-0541-x

[37] Kim, S., Seo, M., Laptev, I., Cho, M., Kwak, S.: Deep metric learning beyond binary supervision. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pp. 2288–2297. Computer Vision Foundation / IEEE, Long Beach, CA, USA (2019)

[38] Sohn, K.: Improved deep metric learning with multi-class N-pair loss objective. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, Barcelona, Spain, pp. 1849–1857 (2016)

[39] Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), pp. 1735–1742. IEEE Computer Society, New York, NY, USA (2006). https://doi.org/10.1109/cvpr.2006.100

[40] Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, pp. 815–823. IEEE Computer Society, ,Boston, MA, USA (2015). https://doi.org/10.1109/cvpr.2015.7298682

[41] Song, H.O., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, pp. 4004–4012. IEEE Computer Society, Las Vegas, NV, USA (2016)

[42] Teh, E.W., DeVries, T., Taylor, G.W.: Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. Springer International Publishing eBooks **12369**, 448–464 (2020). https://doi.org/10.1007/978-3-030-58586-0_27

[43] Yang, Z., Bastan, M., Zhu, X., Gray, D., Samaras, D.: Hierarchical proxy-based loss for deep metric learning. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, pp. 449–458. IEEE, Waikoloa, HI, USA, (2022)

[44] Qian, Q., Shang, L., Sun, B., Hu, J., Tacoma, T., Li, H., Jin, R.: SoftTriple loss: deep metric learning without triplet sampling. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, pp. 6449–6457. IEEE, Seoul, Korea (South) (2019). https://doi.org/10.1109/iccv.2019.00655

[45] Aziere, N., Todorovic, S.: Ensemble deep manifold similarity learning using hard proxies. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pp. 7299–7307. Computer Vision Foundation / IEEE, Long Beach, CA, USA, (2019). https://doi.org/10.1109/cvpr.2019.00747

[46] Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: IEEE International Conference on Computer Vision, ICCV 2017, pp. 360–368. IEEE Computer Society, Venice, Italy (2017)

[47] Kim, S., Kim, D., Cho, M., Kwak, S.: Proxy anchor loss for deep metric learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, pp. 3235–3244. Computer Vision Foundation / IEEE, Seattle, WA, USA (2020). https://doi.org/10.1109/cvpr42600.2020.00330

[48] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA (2015)

[49] Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, pp. 1746–1751. ACL, Doha, Qatar (2014)

[50] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a Meeting Held December 5-8, 2013, Lake Tahoe, Nevada, United States, pp. 3111–3119 (2013)

[51] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) International Conference on Learning Representations, ICLR 2015,, San Diego, CA, USA (2015). http://arxiv.org/abs/1412.6980

[52] 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. **36**(3), 521–535 (2014). https://doi.org/10.1109/tpami.2013.142

[53] Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using amazon's mechanical turk. In: Proceedings of the 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 139–147. Association for Computational Linguistics, Los Angeles, USA (2010)

[54] Chua, T., 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 8th ACM International Conference on Image and Video Retrieval, CIVR 2009. ACM, Santorini Island, Greece, (2009). https://doi.org/10.1145/1646396.1646452

[55] Zhai, X., Peng, Y., Xiao, J.: Learning cross-media joint representation with sparse and semisupervised regularization. IEEE Transactions on Circuits and Systems for Video Technology (2014)

[56] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA (2015)

[57] Zhai, X., Peng, Y., Xiao, J.: Cross-modality correlation propagation for cross-media retrieval. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, pp. 2337–2340. IEEE, Kyoto, Japan (2012). https://doi.org/10.1109/icassp.2012.6288383

[58] Jan, R., John, S.: Multi-view canonical correlation analysis. In: Conference on Data Mining and Data Warehouses (SiKDD 2010), pp. 1–4 (2010)

[59] Peng, Y., Peng, Y., Xiao, J.: Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence. AAAI Press, Bellevue, Washington, USA (2013)

[60] Zhai, X., Peng, Y., Xiao, J.: Learning cross-media joint representation with sparse and semisupervised regularization. IEEE Trans. Circuits Syst. Video Technol. **24**(6), 965–978 (2014). https://doi.org/10.1109/tcsvt.

2013.2276704

[61] 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, IJCAI 2016, pp. 3846–3853. IJCAI/AAAI Press, New York, NY, USA, (2016)

[62] Jing, M., Li, J., Zhu, L., Lu, K., Yang, Y., Huang, Z.: Incomplete cross-modal retrieval with dual-aligned variational autoencoders. In: MM '20: The 28th ACM International Conference on Multimedia, Virtual Event, pp. 3283–3291. ACM, Seattle, WA, USA (2020). https://doi.org/10.1145/3394171.3413676

[63] Hu, P., Zhen, L., Peng, D., Liu, P.: Scalable deep multimodal learning for cross-modal retrieval. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, pp. 635–644. ACM, Paris, France, (2019). https://doi.org/10.1145/3331184.3331213

[64] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[65] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR

[66] Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of machine learning research **9**(11) (2008)