ISSN: 2583-9888
DOI: https://doi.org/10.51219/JAIMLD/Ming-Sheng/11

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 1 & Iss: 3

Research Article

Deep Adversarial Common Subspace Learning for Image-Text Cross-Modal Retrieval

Yuanyuan Ma[1], Ming Sheng[2]* and Jing He[1]*

[1]College of Electrical and Information Engineering, Hunan University of Technology, Zhuzhou, Hunan, China

*[2]College of Engineering, Shandong Xiehe University, Jinan, Shandong, China

Citation: Ma, Y., Sheng, M., & He, J. (2023). Deep Adversarial Common Subspace Learning for Image-Text Cross-Modal Retrieval. *J Artif Intell Mach Learn & Data Sci*, 1(3), 80-85. DOI: https://doi.org/10.51219/JAIMLD/Ming-Sheng/11

***Corresponding author:** Ming Sheng, College of Engineering, Shandong Xiehe University, Jinan, Shandong, China. Email: shengming@sdxiehe.edu.cn

Received: 25 March, 2023; **Accepted:** 12 June, 2023; **Published:** 20 July, 2023

ABSTRACT

As a result of the massive expansion of multimodal data in recent years, researchers have developed a strong interest in the problem of cross-modal retrieval. Finding instances from multimodal data that are semantically related is the aim of cross-modal retrieval. However, the distribution of data across different modalities is inconsistent and there is a heterogeneity gap. At the same time, the semantic distinction between low-level and high-level semantic information makes cross-modal retrieval work challenging. This paper proposes a deep adversarial common subspace learning approach based on real-valued representation learning for image-to-text cross-modal retrieval., which combines adversarial learning with common subspaces to get image-text feature representations of the same dimension in the common space. Secondly, it effectively utilizes label prediction and triplet loss to conduct constraint learning on the network, further improving the retrieval accuracy. Finally, the method achieves promising results on two public datasets on cross-modal retrieval tasks, proving its effective in cross-modal retrieval.

Keywords: cross-modal retrieval, heterogeneous gap, semantic gap, adversarial learning

Introduction

The rapid expansion of big data has led to an increase of multimodal data on the Internet. The multimodal data are expressed in different ways, but they have the same semantic information. As an illustration, an image may directly represent the content of a text, and the content of an image can likewise be communicated in words. Image modality and text modality enrich semantic information from different perspectives, enabling people to understand related things better. Multimodal data has become a common phenomenon nowadays, making cross-modal retrieval receive much attention. Cross-modal retrieval is a research-intensive and challenging task in information retrieval. It uses one modality sample to retrieve another modality sample, and its goal is to discover the semantic relationship between different modality samples.

This article focuses on image-text retrieval, one of the most prevalent cross-modal retrieval tasks. It mainly includes two aspects: (1) retrieve semantically similar text samples through image samples; (2) retrieve semantically similar text samples through text samples Similar image samples. By learning a common representation space, it may mine the connections between different modalities in this space to achieve the objective of cross-modal retrieval. However, cross-modal data often have different feature representations and distributions, which leads to their heterogeneity, and the similarity between two features cannot be directly calculated.

Numerous academics have undertaken substantial study and made major breakthroughs in cross-modal retrieval as a result of the rapid growth of deep learning in recent years. The data of different modalities are mapped into a subspace, and then the similarity evaluation is performed in this subspace, which is a traditional cross-modal retrieval method. Among traditional methods, canonical correlation analysis (CCA) [1] primarily teaches how to project in a linear way the maximum statistical correlation between pairs of image-text data. In further research, kernel canonical correlation analysis (KCCA) [2] learns projections on the kernel space to solve problems that cannot be solved in canonical correlation methods, a further extension of the CCA method. However, these methods simply embed data into the subspace, and the retrieval accuracy is not high.

Due to the continued improvement of deep learning techniques, cross-modal retrieval approaches have made substantial progress in the field of deep learning. As an illustration, the canonical correlation analysis approach incorporates deep neural network (DNN) [3], and deep canonical correlation analysis (DCCA) [3] and deep correlation autoencoders (DCCAE) [4] are obtained to learn common feature representations across different modalities. Peng et al. [5] proposed cross-modal multiple deep networks (CMDN) to mine the association information between complex cross-modal data through hierarchical learning and jointly model the information within and between modalities. Then Hierarchical representations are then performed to learn the correlations between them. Zhai et al. [6] proposed a joint representation learning method (JRL), which can jointly investigate the pertinent information and semantic information of various modalities in a unified optimization framework, and carry out unified optimization. Wang et al. [7] proposed an adversarial cross-modal retrieval method (ACMR), which obtains modality-invariant representations through feature projectors and uses modality classifiers to distinguish different modality feature representations. Peng et al. [8] proposed a hierarchical network-based multi-granularity fusion cross-modal correlation learning approach (CCL), which uses jointly optimized multi-layer correlations to preserve intra-modality and inter-modality correlations. Xu et al. [9] proposed a new correlation feature synthesis and alignment method (CFSA), synthesizes multimodal features with semantic correlations using a generative adversarial network, and maps the synthetic and real characteristics to a shared semantic space to capture the correlation between distinct semantic features. Shen et al. [10] propose a cluster-driven deep adversarial hashing method (CDAH), which generates modality-invariant representations by soft-clustering models and adopts modality classifiers to distinguish modality categories.

In this paper, to bridge the heterogeneity and semantic gaps between cross-modal data, we present an adversarial common subspace strategy for cross-modal retrieval. To achieve this goal, the heterogeneity gap is reduced by minimizing the adversarial loss to obtain modality-invariant representations of image samples and text samples. Then, to mine the semantic connection between image features and text features, a weight-sharing method is used to map the features into a shared space, and the semantic gap is further narrowed by combining label information and triplet constraints. The following are the primary contributions of our method:

1. Adversarial learning is implemented into a common subspace to close the heterogeneity gap and maintain modality invariance between image and text modalities.
2. Label prediction and triplet constrained loss are used to constrain the model to obtain more discriminative image-text feature representation.
3. The effectiveness of our method is demonstrated by various relevant experimental findings on two public datasets.

The rest of this article is below. Cross-modal retrieval-related research is included in Section II. The description of the problem, the structure of the model, and the objective function of the proposed method are all described in Section III. The experimental findings and analyses are in Section IV. The last section is Section V.

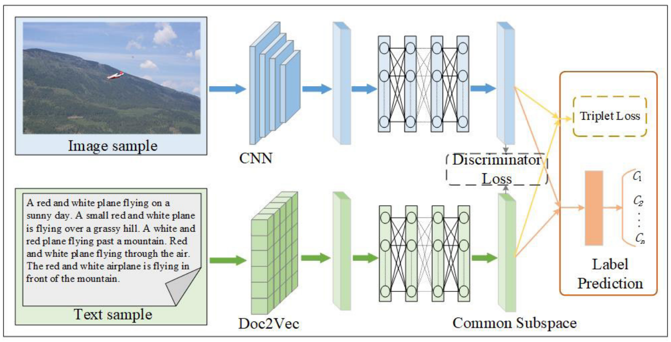


Image sample
CNN
Discriminator Loss
Triplet Loss
C_1
C_2
C_n
Label Prediction
A red and white plane flying on a sunny day. A small red and white plane is flying over a grassy hill. A white and red plane flying past a mountain. Red and white plane flying through the air. The red and white airplane is flying in front of the mountain.
Text sample
Doc2Vec
Common Subspace

Figure 1: Framework diagram of the proposed method

Related work

To overcome the heterogeneous gap in multimodal data, we aim to learn a shared feature space. Li et al. [11] proposed an unsupervised cross-modal hash retrieval method (MCMHR), which captures the potential relationship between different modalities through an auxiliary similarity matrix. Yao et al. [12] proposed a discrete semantic alignment hashing (DSAH), which mines the relationship between class labels and hash codes through collaborative filtering, and semantically aligns semantic information with text modalities using image labels. This type of method belongs to the category of hash learning, which seeks to map different modal features to a hamming space, resulting in the separation of samples from various categories and the proximity of samples within the same category.

Another method belongs to the real-valued learning category and tries to learn a real-valued common space. Hao et al. [13] proposed a cross-modal retrieval method (ACME) with adversarial cross-modal embedding, which mainly learns the shared feature space of two modalities. Zhen et al. [14] proposed a deep supervised cross-modal retrieval method (DSCMR), with the goal of finding a common real-valued representation space and learning discriminative features and learning modality invariance of features in a supervised learning setting. Li et al. [15] proposed the semantically supervised maximum correlation method (S2MC), which effectively utilizes the supervision information in the public feature space and label space based on the maximum correlation method. Xu et al. [16] propose a deep adversarial metric learning method (DAML), that feature mapping from various modalities into a common feature subspace is accomplished by a nonlinear algorithm. Hu et al. [17] proposed a cross-modal discriminative confrontation network method (CAN), the generator maps different modal data into the potential cross-modal discriminative space and reduces the heterogeneity gap in the common space through the competition between the discriminator and the generator. Compared with hash learning, real-valued learning has higher accuracy and emphasizes semantic matching across various modalities of input more.

Based on real-valued learning, the features of the two modalities are mapped into a latent common real-valued space in this paper, while combining adversarial learning to close the heterogeneity gap and obtain modality-invariant feature representations to improve retrieval accuracy.

The Proposed Method

This part primarily introduces the problem formulation, the model architecture, and the definition of the objective function.

Problem formulation

In this paper, the dataset contains image samples and text samples as X and Y. Assume $Q=\{x_i, y_i, l_i\}_{i=1}^{n}$ is n cross-modal data set of image-text pairs and their corresponding label sets, where $x_i \in \mathbb{R}^{d_x}$ stands for the image feature, $y_i \in \mathbb{R}^{d_y}$ for the text feature, d_x is the dimension of the image feature, and d_y is the dimension of the text feature. $l_i=\{l_{i1}, l_{i2}, ..., l_{ic}\}$ are the semantic labels applied to texts and images, and C is the total number of categories. If x_i and y_i belong to the same class, then $l_{ic}=1$, otherwise 0. The matrices of image features, text features, and labels are expressed as $X=[x_1, x_2, ..., x_n] \in \mathbb{R}^{n \times d_x}$, $Y=[y_1, y_2, ..., y_n] \in \mathbb{R}^{n \times d_y}$, and $L=[l_1, l_2, ..., l_n] \in \mathbb{R}^{n \times c}$ respectively. Image features and text features cannot be directly compared because they exist in different high-dimensional spaces, and they need to be mapped to a common subspace S by means of feature mapping so that images and texts have the same feature dimension S_o, the feature mapping functions for images and text are $S_X = f(X, \theta_X)$ and $S_Y = f(Y, \theta_Y)$, respectively.

Model framework

The structural architecture for our technique is displayed in Figure 1. Firstly, to extract image-text features, image samples and text samples are input into the VGG19 [18] model and Doc2vec [19] model, respectively. At the same time, the distribution alignment is carried out by using discrimination loss to close the heterogeneity gap and maintain modal invariance. Secondly, input image features and text features into two sub-networks, and perform feature representation in a common space through weight sharing. Finally, the similarity measurement is performed on the features of all modalities through the triplet constraint loss, and the label prediction loss is used to predict label to ensure that features within the modal are still recognizable.

Objective function

The final feature representation is generated after image-text features pass through shared layers. This paper employs a triplet constraint loss function to calculate the similarity between image and text modality in public space. In the triplet loss function, one modality feature is used as the anchor point, while the other modality feature is used as the positive and negative feature items [13]. Its goal is to hope that the positive samples gradually approach the anchor point, and the negative samples gradually move away from the anchor point. There are two distinct forms of triplets in this paper: one has an image feature *X* as the anchor and the other has a text feature *Y* as the anchor. The equation reads as follows:

$$\begin{aligned} L_{tri} = & \sum_{X}[d(S_{Xa}, S_{Yp}) - d(S_{Xa}, S_{Yn}) + \alpha]_+ \\ & + \sum_{X}[d(S_{Xa}, S_{Xp}) - d(S_{Xa}, S_{Xn}) + \alpha]_+ \\ & + \sum_{Y}[d(S_{Ya}, S_{Xp}) - d(S_{Ya}, S_{Xn}) + \alpha]_+ \\ & + \sum_{Y}[d(S_{Ya}, S_{Yp}) - d(S_{Ya}, S_{Yn}) + \alpha]_+ \quad (1) \end{aligned}$$

Where S_X and S_Y are the mapped image features and text features, $d(\bullet)$ is the Euclidean distance, the anchor point is indicated by the subscript *a*, while the positive sample, negative sample, and error range are indicated by the subscripts *p*, *n*, and *α*.

In the common space, to lessen the difference in feature distribution across sample pairs, it is necessary to align the encoding feature distribution of images and texts and maintain modality invariance. In this paper, using the adversarial loss of WGAN [20], and the equation reads as follows:

$$L_{dis} = E_{X \sim P_{S_X}}[\log D_M(S_X)] + E_{Y \sim P_{S_Y}}[\log(1 - D_M(S_Y))] \quad (2)$$

Where D_M is the discriminator, P_{S_X} and P_{S_Y} are the corresponding probability distributions.

And resolved by min-max optimization as follows:

$$\min_{S_X, S_Y} \max_{D_M} L_{dis}$$

Feature label information can improve cross-modal retrieval results. This study uses two category classifiers to predict labels, ensuring that image features and text features are still distinguishable within categories. The equation reads as follows:

$$L_{pre} = \frac{1}{n} \| W^T S_X - L \|_F + \frac{1}{n} \| W^T S_Y - L \|_F \quad (3)$$

Among ||•|| is the Frobenius norm, and W is the projection matrix of the feature classifier.

The overall objective function is as follows:

$$L = \alpha * L_{\text{tri}} + \beta * L_{\text{dis}} + L_{pre} \quad (4)$$

Where α and β are weight parameters. The optimization process of the method in this paper is shown in Algorithm 1:

Algorithm 1 Optimization process of the proposed method

Input: The image set *X*, the text set *Y*, the label set *L*, the learning rate *λ*, the number of epochs is *N*, and the hyperparameters *α* and *β*.

Output: The transformed image mode and text mode features represent S_X and S_Y.

update until convergence:

```
1. Randomly initialize the network parameters $\theta_X$ and $\theta_Y$.
2. for each k in Xtrain and Ytrain do
3. Calculate the image and text feature vectors $S_X$ and $S_Y$ of the common space through forward propagation.
4. Compute the result of the function in Equation (4).
5. Optimize the objective function through the Adam optimizer and update the parameters $\theta_X$ and $\theta_Y$ in the mapping network.
6. end for
7. return the final $S_X$ and $S_Y$.
```

Experiments

In this section, our content mainly includes comparative experiments, precision-recall curves, ablation experiments, and loss function change curves.

Experimental setup

Datasets and features

This paper conducts related experiments on two public datasets: Wikipedia [21] and Pascal Sentence [22]. Both datasets consist of images and texts with corresponding labels. The two datasets utilized in the experiments are briefly described in the paragraphs that follow.

The Wikipedia dataset has a total of 2866 image-text data, and each pair of image-text data contains 10 semantic category labels. There are 462 pairs in the test set, 231 pairs in the validation set, and 2173 pairs in the training set, which are three parts of the dataset.

The Pascal Sentence dataset contains 1000 image-text data, including 20 category labels. The dataset includes 800 pairs of training set, 100 pairs of verification set, and 100 pairs of test set.

For a more fair and objective comparison, this paper strictly follows the dataset division and feature extraction methods in [23,24]. For image samples, this paper mainly uses the fc7 layer of the VGG19 model in the convolutional neural network to extract 4096-dimensional features to represent. To extract text features to represent each text, a pre-trained Doc2Vec model is used to extract 300-dimensional features to represent.

Implementation Details

In the model structure, images and texts obtain 4096-dimensional and 300-dimensional features through their respective feature extraction models. Then the image features and text features are respectively input into the fully connected layer, and the 1024-dimensional features are output.

The experiment in this paper is carried out on Pytorch3.7. In the course of training, the batch size on all data sets is 50, the epoch is 100, and the learning rate is 0.0001. For parameters α and β, the appropriate parameter values were obtained by the grid search algorithm and set to 0.1 and 0.001, respectively.

Evaluation metric

The proposed method is primarily used in this paper for two retrieval tasks: image-to-text retrieval (Img2Txt) and text-to-image retrieval (Txt2Img). The evaluation criterion we utilize is mean average precision (mAP) [25], to obtain the mean average precision (mAP) value, first determine the average precision (AP) of each retrieval result for each query item, the mAP value may then be calculated. Following is how the AP and mAP are expressed:

$$AP = \frac{1}{P}\sum_{k=1}^{N}\frac{P_k}{k}\times rel_k \qquad (5)$$

$$mAP = \frac{1}{R}\sum_{k=1}^{R}AP(i) \qquad (6)$$

Where P represents the quantity of related samples in the test set, N represents the quantity of samples in the test set, and P_k represents the quantity of related samples in the first k returned results. $rel_k = 1$ if the k^{th} sample is relevant, otherwise, $rel_k = 0$. R is the quantity of query samples, and $AP(i)$ is the AP value of the i^{th} instance.

Comparison results

This study compares the suggested approach to seven commonly employed methods, both conventional and deep learning-based. Conventional methods include CCA [1], GMA [26], and LCFS [27], and deep learning methods include DCCA [3], ACMR [7], DRSL [28], and DSCMR [14].

On the two datasets, Tables 1 and 2 shown the mAP values for our method and the comparative methods. The table shows that our method performs much better than conventional methods. In the Img2Txt and Txt2Img retrieval tasks on the Pascal Sentence dataset, our method beats the DSCMR by 0.8% and 2.5%, respectively. Although the improvement on Img2Txt is not large, the average mAP score has increased by 1.6%. For the Wikipedia dataset, the two retrieval tasks of Img2Txt and Txt2Img are improved by 1.4% and 1.9%, respectively. In addition, the deep learning-based method can achieve higher mAP scores compared with conventional methods. The findings demonstrate that adversarial learning is beneficial to the model to learn modality invariance better to obtain effective cross-modal feature representations. Triplet constraints can effectively use label information, making the model able to model inter-modal similarities efficiently.

Table 1: The retrieval results on Pascal Sentence dataset.

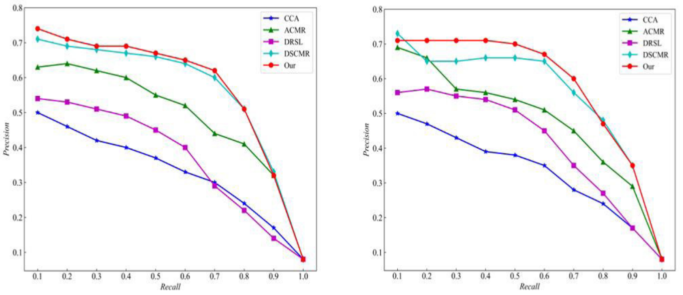
Method	Img2Txt	Txt2Img	Average
CCA	0.457	0.449	0.453
GMA	0.427	0.339	0.383
LCFS	0.344	0.267	0.306
DCCA	0.606	0.633	0.620
ACMR	0.657	0.626	0.642
DRSL	0.631	0.641	0.636
DSCMR	0.710	0.703	0.707
Our	**0.718**	**0.728**	**0.723**

Table 2: The retrieval results on Wikipedia dataset.

Method	Img2Txt	Txt2Img	Average
CCA	0.221	0.196	0.209
GMA	0.272	0.232	0.253
LCFS	0.280	0.214	0.247
DCCA	0.452	0.411	0.431
ACMR	0.416	0.392	0.404
DRSL	0.447	0.419	0.433
DSCMR	0.476	0.407	0.442
Our	**0.490**	**0.426**	**0.458**

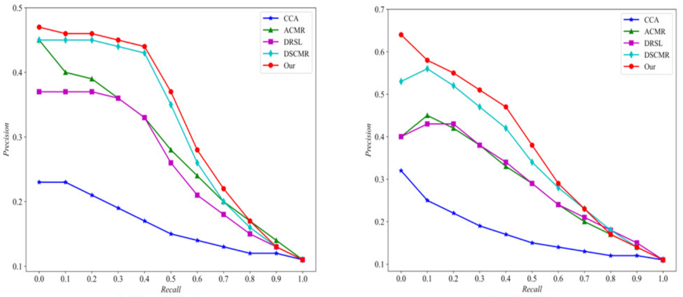
Precision-Recall curve

For a more detailed comparison, Figures 2 and 3 depict the precision-recall curves for the two datasets. This curve records the precision and recall values of CCA, ACMR, DRSL, DSCMR, and the method in this paper when performing Img2Txt retrieval and Txt2Img retrieval. From the figure, it can be seen that the curve is at a high point when the recall rate is low, and then falls as the recall rate increases. When the recall is the same, our methods obtain a better level of precision than other methods.



(a) Img2Txt (b) Txt2Img

Figure 2: Precision-recall curve on the Pascal sentence dataset



(a) Img2Txt (b) Txt2Img

Figure 3: Precision-recall curve on the Wikipedia dataset

Ablation analysis

To demonstrate the contribution of each part of the loss function to model performance, ablation learning is used for evaluation in this paper.

Table 3: Results of ablation experiment on Pascal Sentence dataset.

Model	Img2Txt	Txt2Img	Average
w/o L_{pre}	0.581	0.538	0.560
w/o L_{dis}	0.693	0.679	0.686
w/o L_{tri}	0.695	0.688	0.692
Full	**0.718**	**0.728**	**0.723**

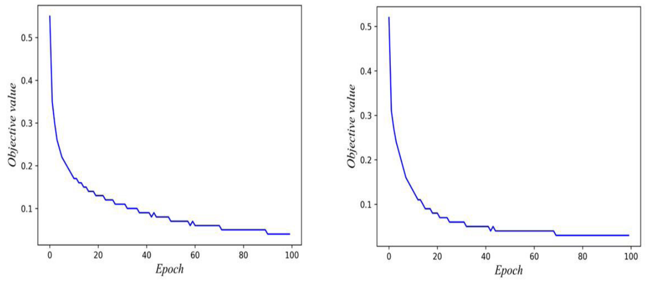
Table 4: Results of ablation experiment on Wikipedia dataset.

Model	Img2Txt	Txt2Img	Average
w/o L_{pre}	0.428	0.386	0.407
w/o L_{dis}	0.471	0.415	0.443
w/o L_{tri}	0.478	0.402	0.440
Full	**0.490**	**0.426**	**0.458**

Tables 3 and 4 display the outcomes of the ablation experiments performed on the two datasets. We can see from the data in the table that the label prediction loss, adversarial loss and triplet constraint loss in the objective function all have varying degrees of impact on the retrieval accuracy. In both datasets, the results of variants are generally better than those of the overall objective function. In addition, it can also be found that the label prediction loss has the most impact on the model.

Convergency

This paper draws a curve to record the evolution of the loss function on the two datasets. According to Figure 4, as the number of iterations rises, the loss value decreases monotonically and converges during the period of training. Although some jitters exist in the two data sets, they are close to convergence, indicating that the method has good stability.



(a) Pascal Sentence (b) Wikipedia

Figure 4: The loss function change curve on the dataset for Pascal sentences and the Wikipedia dataset

Conclusion

In this study, we propose a cross-modal image-text retrieval approach based on deep adversarial common subspaces learning, which acquires image-text feature representations of the same dimension via common spaces to reduce the heterogeneity gap. Furthermore, the semantic gap is narrowed by combining label information to obtain discriminative semantic features. The efficiency of the strategy is demonstrated through related experiments on two public datasets, they indicate that it can effectively jointly learn the adversarial loss, label prediction and triplet constraints, and achieve good results.

References

1. Hotelling, H. (1935). *Relations between two sets of variates*. Biometrika, 28, 321-377. https://doi.org/10.1007/978-1-4612-4380-9_14
2. Akaho, S. (2006). A kernel method for canonical correlation analysis. *In Proceedings of the International Meeting of Psychometric Society*, pages 263–269, https://doi.org/10.1007/s10489-013-0464-2
3. Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013). Deep Canonical Correlation Analysis. *International Conference on International Conference on Machine Learning. JMLR, 28*(3), 1247-1255. JMLR.org
4. Wang, W., Arora, R., Livescu, K., & Bilmes, J. (2015). On deep multi-view representation learning. *In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15). JMLR, 37*, 1083–1092. *JMLR.org*. Retrieved from https://proceedings.mlr.press/v37/wangb15.html
5. Peng, Y., Xin, H., & Qi, J. (2016). Cross-media shared representation by hierarchical learning with multiple deep networks. *In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16). AAAI Press*, 3846–3853. Retrieved from https://www.ijcai.org/Proceedings/16/Papers/541.pdf
6. Zhai, X., Peng, Y., & Xiao, J. (2014). Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology, 24*(6), 1-1. https://doi.org/10.1109/TCSVT.2013.2276704
7. Wang, B., Yang, Y., Xu, X., Hanjalic, A., & Shen, H. T. (2017). Adversarial Cross-Modal Retrieval. *In Proceedings of the 25th ACM international conference on Multimedia. Association for Computing Machinery*, 154–162. https://doi.org/10.1145/3123266.3123326
8. Peng, Y., Qi, J., Huang, X., & Yuan, Y. (2017). CCL: cross-modal correlation learning with multi-grained fusion by hierarchical network. *IEEE Transactions on Multimedia, 20*(2), 405-420. https://doi.org/10.1109/TMM.2017.2742704
9. Xu, X., Lin, K., Lu, H., Gao, L., & Shen, H. T. (2020). Correlated Features Synthesis and Alignment for Zero-shot Cross-modal Retrieval. *The 43rd International ACM SIGIR conference on research and development in Information Retrieval. ACM*, 1419–1428. https://doi.org/10.1145/3397271.3401149
10. Shen, X., Zhang, H., Li, L., Zhang, Z., Chen, D., & Liu, L. (2021). Clustering-driven deep adversarial hashing for scalable unsupervised cross-modal retrieval. *Neurocomputing, 459*, 152-164. https://doi.org/10.1016/j.neucom.2021.06.087
11. Li, Z., Xie, X., Ling, F., Ma, H., & Shi, Z. (2021). Matching images and texts with multi-head attention network for cross-media hashing retrieval. *Engineering Applications of Artificial Intelligence, 106*, 104475. https://doi.org/10.1016/j.engappai.2021.104475
12. Yao, T., Kong, X., Fu, H., & Tian, Q. (2019). Discrete semantic alignment hashing for cross-media retrieval. *IEEE Transactions on Cybernetics, 99*, 1-12. https://doi.org/10.1109/TCYB.2019.2912644
13. Wang, H., Sahoo, D., Liu, C., Lim, E. P., & Hoi, S. (2019). Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. *IEEE*. pp. 11564-11573. https://doi.org/10.1109/CVPR.2019.01184
14. Zhen, L., Hu, P., Wang, X., & Peng, D. (2020). Deep Supervised Cross-Modal Retrieval. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). *IEEE*. pp. 10386-10395. https://doi.org/10.1109/CVPR.2019.01064
15. Li, M., Li, Y., Huang, S. L., & Zhang, L. (2020). Semantically Supervised Maximal Correlation For Cross-Modal Retrieval. 2020 IEEE International Conference on Image Processing (ICIP). *IEEE*. pp. 2291-2295. https://doi.org/10.1109/ICIP40778.2020.9190873
16. Xu, X., He, L., Lu, H., Gao, L., & Ji, Y. (2019). Deep adversarial metric learning for cross-modal retrieval. *World Wide Web, 22*(2), 657-672. https://doi.org/10.1007/s11280-018-0541-x
17. Hu, P., Peng, X., Zhu, H., Lin, J., & Peng, D. (2020). Cross-modal discriminant adversarial network. *Pattern Recognition, 112*(1), 107734. https://doi.org/10.1016/j.patcog.2020.107734

18. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *Computer Science*. https://doi.org/10.48550/arXiv.1409.1556

19. Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *In Proceedings of the 31st International Conference on International Conference on Machine Learning*. 1188-1196. https://doi.org/10.48550/arXiv.1405.4053

20. Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. Machine Learning. https://doi.org/10.48550/arXiv.1701.07875

21. Pereira, J. C., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G., & Levy, R., et al. (2014). On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Trans Pattern Anal Mach Intell, 36*(3), 521-35. https://doi.org/10.1109/TPAMI.2013.142

22. Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010). Collecting image annotations using Amazon’s Mechanical Turk. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, pages 139–147.

23. Peng, Y., Qi, J., & Yuan, Y. (2017). Cm-gans: cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing Communications and Applications, 15*(1), Article No: 22, pp 1–24 https://doi.org/10.1145/3284750

24. Qi, J., & Peng, Y. (2018). Cross-modal Bidirectional Translation via Reinforcement Learning. *Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 2630–2636. https://doi.org/10.24963/ijcai.2018/365

25. Wang, K., Yin, Q., Wei, W., Shu, W., & Liang, W. (2016). A comprehensive survey on cross-modal retrieval. https://doi.org/10.48550/arXiv.1607.06215

26. Sharma, A., Kumar, A., Daume, H., & Jacobs, D. W. (2012). Generalized Multiview Analysis: A discriminative latent space. *IEEE*. 2160–2167. https://doi.org/10.1109/CVPR.2012.6247923

27. Wang, K., He, R., Wei, W., Liang, W., & Tan, T. (2013). Learning coupled feature spaces for cross-modal matching. *In Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV ‘13). IEEE Computer Society*, 2088–2095. https://doi.org/10.1109/ICCV.2013.261

28. Wang, X., Hu, P., Zhen, L., & Peng, D. (2021). Drsl: deep relational similarity learning for cross-modal retrieval. *Information Sciences, 546*, 298-311. https://doi.org/10.1016/j.ins.2020.08.009