An Overview of Image-text Retrieval Methods Based on Deep Learning

Han Li[1,a], Liang Bai[1,b], Songyang Lao[1,c], Zhihong Deng[2,d], Qiuyu Ren[1,e], Can Luo[1,f]

[1]College of Systems Engineering National University of Defence Technology, Changsha, China;
[2]National Defence University, Changsha, China.
a. lihan_dter@163.com b. xabpz@163.com c. laosongyang@nudt.edu.cn
d. lingyu0207@163.com e. renqiuyu_nudt@163.com f. luoc003@163.com

Keywords: Image-text retrieval, deep learning, key and difficult points, research direction.

Abstract: With the popularization and development of the Internet and digital media, the demand for cross-media retrieval based on images and texts has also increased. This paper first introduces the image-text retrieval and deep learning, and then gives the basic steps of cross retrieval. On this basis, the paper analyzes the key points and difficulties of image-text retrieval based on deep learning at present, and further discuss its research direction based on the existing work. Finally, the future development trend of image-text retrieval is discussed.

1. Introduction

With the further popularization of computers, the Internet and digital media, multimedia information with text, video, audio and images as the main body has increased dramatically, and it is possible to realize the sharing of global multimedia information through the Internet. Users are also becoming more and more acquainted with multimedia information, and various new application requirements have followed.

Deep learning has been applied in various fields and has made breakthroughs, including cross-media retrieval. Deep neural networks have the ability to characterize high-level features, and are widely used to learn from a large number of sample sets to the essential characteristics of data, which narrows the semantic gap to some extent.

Section II of this paper mainly introduces the concept and development of deep learning. Section III first expounds the image-text retrieval and gives the basic steps of search. Then we analyzed the difficulties at the current stage, and finally discussed the current research direction in this field with the difficulties. In the section IV, the future development trend of image-text retrieval is discussed.

2. Deep Learning

Deep learning, as part of machine learning, occurs in the wave of machine learning from shallow to deep learning[1]. The deep learning model and the shallow machine learning model are quite different in the means of extracting features. Shallow machine learning models do not use distributed representations[2] and require artificial extraction of features. Since the model itself only classifies or predicts tasks through the input features, the quality of the whole system depends largely on the characteristics of artificial extraction. Feature extraction engineering takes a lot of time and requires a high level of knowledge in the professional field. Deep learning is a representational learning[3] that can learn a higher level of abstract representation of data and automatically extract features from the data[4][5]. The hidden layer in deep learning is equivalent to a linear combination of input features, and the weight between the hidden layer and the input layer is equivalent to the weight of the input features in the linear combination[6]. In addition, the learning ability of the deep learning model will increase rapidly with the increase of depth[7].

Deep learning has been proposed since 1943 and has gone through three historical periods. Figure 1 shows the development of deep learning. The open circle indicates the key turning point of the rise and fall of the deep learning heat. The size of the solid circle indicates the breakthrough of deep learning in this year. The oblique upward line indicates that the deep learning heat is rising, and the oblique downward line indicates that the deep learning heat is in a falling period.

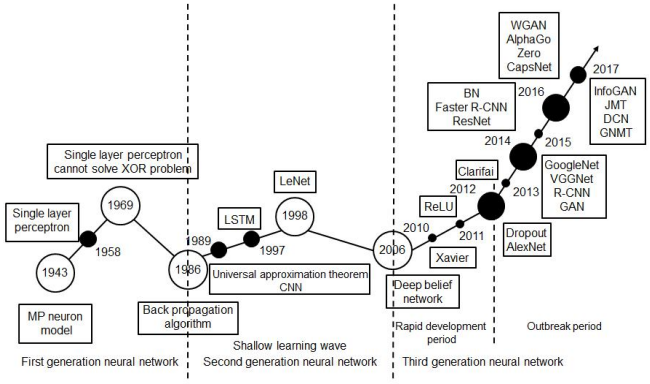

Figure 1: History of deep learning development.

3. Image-text Retrieval

Image-text retrieval is the most representative research point in cross-media search[8]-[12]. In reality, text and images are the two most common media in life. Realizing the retrieval of images and texts allows people to quickly find the required content from the massive data, reduce the workload, and effectively improve the precision and recall of information retrieval. From the technical framework, speech recognition technology is a good way to achieve the conversion of audio modality and text modality. One frame of the video modality is an image, and the modeling problems of both are better solved. Therefore, solving the modeling problem of image and text retrieval is of great significance to the cross-search of other modes.

As shown in Figure 2, when users search for the entry of the Yellow River, they will not only get text descriptions about the Yellow River, but also media information such as videos, audio, and pictures. Such search results are rich in content, and the query object can be presented to the user more stereoscopically.

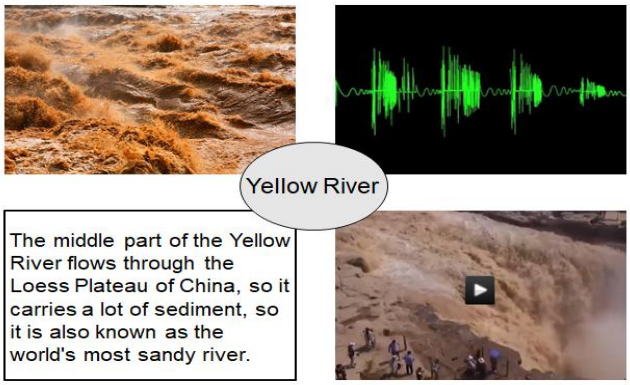

Figure 2: Different modality representations of the Yellow River.

Deep learning is mainly used in the Common space learning based methods. The image-text retrieval method based on the common space learning refers to mapping the multimedia data features in the heterogeneous space into the isomorphic space through a certain mapping mechanism, so that the two pairs of similarity comparisons can be performed between different data. Common space learning[13]-[37] aims to obtain potential subspaces shared between multiple modalities to capture complementary information between different modalities. The methods for constructing common space include traditional statistical correlation analysis[13]-[15], DNN-Based method[14][15], cross-media graph regularization[16]-[19], metric learning[20]-[22], and learning to rank methods[23]-[26], dictionary learning[27]-[30], cross-media hash[31]-[37] and so on.

3.1.Steps of Image-text Retrieval based on Deep Learning

Image-text retrieval based on deep learning is roughly divided into two stages. The first stage is feature extraction. The second stage is to cross and fuse features, eliminate differences, and map each modality to the same feature space for retrieval. As shown in Figure 3:

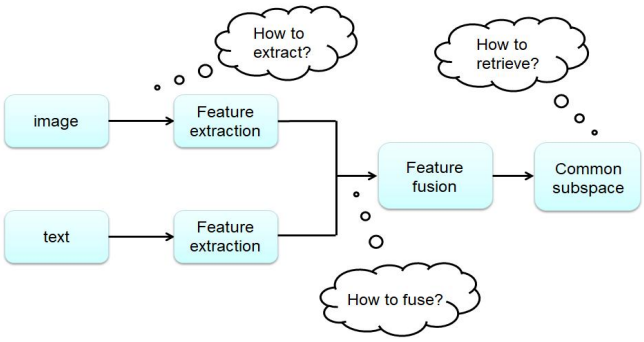

Figure 3: Image-text retrieval processes.

As can be seen from the flow chart, there are three key points of retrieval, feature extraction, feature fusion and common subspace retrieval. In the feature extraction phase, CNN is the most direct and effective method and is widely used in experiments[37]-[45]. In the latest research, LSTM has also become a feature extraction method for images[46][47]. For text, feature extraction is mainly achieved through BOW word bag, Term Frequency-Inverse Document Frequency (TF-IDF), LSTM.

The method of feature fusion is generally to design the network structure. With the LOSS constraint, the samples with similar semantics in the subspace are close to each other, and the

unrelated samples are far away as far as possible. You can also use machine learning or statistical metrics such as the maximum mean difference MMD.

After mapping features into subspaces, find similar samples in the subspace as retrieval targets by using Euclidean distance, cosine distance, and so on. If the feature is merged and converted to a hash code, it can be retrieved using the Hamming distance.

3.2. Key Points and Difficulties of Image-text Retrieval

The key points and difficulties of image-text retrieval are as follows:

Firstly, since the image data of the image and the text modality express the data by using the underlying features of different dimensions and different attributes, the correlation cannot be directly calculated according to the underlying features, so the correlation metric between the two modal media data is very difficult. This heterogeneity[48] and incomparability are also major challenges in cross-media retrieval.

Secondly, the semantic meaning between the modes is not aligned. For example, there is a text description: "There are two children playing on the lawn." In addition to the information of children, lawns, and play, there may be semantic features such as houses and sky appearing on related images. That is to say, related images and texts, their semantic information is not exactly the same, and how to deal with this additional information is also a hot and difficult point of current research.

Third, the amount of data samples is insufficient. Because cross-media correlation is very complex and diverse, high-quality labeled data is critical to training a "good" DNN model. Insufficient data can limit the training effect and easily lead to overfitting. However, labeling cross-modal retrieval dataset is very difficult. Currently widely used dataset, such as Wikipedia, NUS-WIDE and Pascal VOC, have insufficient sample size after processing.

3.3. Main Research Directions

In response to the three main issues of the previous section, the current research is mainly carried out from the following four aspects.

3.3.1. Feature Extraction

Feature extraction is the first step in image-text retrieval. Extracting the quality of features directly determines the effect of retrieval after feature fusion. In order to extract more valuable features, Qi et al.[45] used the Attention module to extract different features of images and texts at the global, local, and relationship levels, and tried to judge the similarity of features with different granularity. They split the image in[46] and input each cut part as a unit into the LSTM to find the area of the image that matches the text word.

3.3.2. Feature Fusion

Feature fusion is designed to integrate multiple modal information to achieve consistent、common model output. The fusion of multi-modal information can obtain more comprehensive features, improve the robustness of the model, and ensure that the model can still work effectively when some modes are missing.

In the early days, Ngiam et al.[49] applied the extension of RBM to public space learning and proposed a dual-mode depth autoencoder in which input from two different media types was shared by the code layer in order to learn to preserve reconstruction information across media dependencies. After that, a series of deep architectures were proposed and progress was made in cross-media retrieval. For instance, Srivastava et al.[50] used two independent deep Boltzmann

machines (DBM) to model the distribution of features of different modal samples. The additional layers at the top of the two models are used together as a joint presentation layer. Points calculate the space of the joint distribution. In addition, the work of combining CCA with DNN as a deep canonical correlation analysis has also emerged (DCCA)[51][52]. As a nonlinear extension of CCA, DCCA can learn complex nonlinear transformations of two media types. Unlike previous work[49][50], the networks they build have a shared layer on different media types. DCCA has two separate subnets that maximize media-to-media correlation through correlation constraints between code layers.

In recent years, confrontational learning has also been applied to cross-media retrieval, representative of the adversarial cross-model retrieval (ACMR) proposed by Wang et al.[41]. At the heart of the framework is a minimax game with a feature projector and a modal classifier. The feature projector generates a modal invariant representation for different modal samples. The modal classifier attempts to determine the sample modality and controls the learning of the feature projector in this way. By placing the modal classifier in the opponent's character, it is desirable to achieve modal invariance more efficiently by better coordinating the distribution of sample representations across modalities.

JIANG et al.[53] designed the DCMH network to integrate feature learning and hash learning into the same framework. In the feature learning stage, a fixed number of bits is generated for the image text, and each bit has a hash code of 1, -1. In the feature fusion stage, the author performs semantic constraints by calculating the similarity matrix S and the design loss function. At the time of retrieval, by calculating the Hamming distance (bitwise comparison), the same number of digits is the retrieval target.

3.3.3. Semantic Alignment

Semantic alignment is mainly to study how to identify the correspondence between components and elements between different modalities, so as to improve the accuracy of retrieval.

Peng et al.[47] proposed a retrieval method for constructing two semantic spaces of images and texts. The matching scores were calculated in two semantic spaces, and the similarity ranking was obtained by adaptive fusion processing.

Qi et al.[45] matched the semantics of pictures and texts at three levels to obtain the global, local and relationship information of different modalities. After discriminating separately on three levels, weighting is performed to obtain the final images retrieval result. In addition, the team of Professor Zhuang Yueting of Zhejiang University proposed a method of learning the common embedded representation space by using the maximum margin learning method combined with local alignment (visual object and lexical alignment) and global alignment (image and sentence alignment)[54].

Professor Gao Xinbo from Xi'an University proposed a cross-modal retrieval method based on discriminative dictionary learning[55]. This method learns discriminative dictionaries to interpret each modality, which not only enhances the discriminating ability of data from different classes of modalities, but also enhances the correlation of modal data in the same class. The code is then mapped to the feature space, and the distinguishing and correlation of the cross-modal data is further enhanced by the label alignment method.

3.3.4. Transfer Learning

Cross-media retrieval dataset labeling is difficult, and the current number of common dataset samples is not very large.

The featured articles in Wikipedia[56] have a total of 29 categories, generating a set of 2,866 image/text pairs based on the most popular category. The NUS-WIDE dataset[57] has a number of

data sets of 5018 after deleting no more than one hundred samples and no samples of the label. In the Pascal Visual Object Class (VOC) dataset[58], 804 keywords annotated 9963, a total of 20 categories of image samples. Currently, Peking University has built a new cross-media dataset, XMedia, which lists 20 categories on five media types. For each category, they collected data for five media types: 250 texts, 250 images, 25 videos, and a total of 12,000 media instances.

Transfer learning is an important solution to the lack of sample size. The team of Professor Peng Yuxin from Peking University has done a lot of work in this direction. In[59], a cross-modal knowledge migration network is proposed to convert trans-modal data into a common representation for retrieval. The modal shared migration sub-network uses the mode of the source domain and the target domain as a bridge to simultaneously transfer knowledge to two modalities. The layer-sharing related sub-network retains the inherent cross-modal semantic correlation to further adapt to the cross-pattern retrieval task. In[43], a progressive migration mechanism is proposed. Through the inter-domain consistency measurement of adaptive feedback, the iterative sample selection is carried out with the principle of migration difficulty from small to large, so that the migration process can gradually reduce the cross-media inter-domain differences and improve the robustness and retrieval accuracy of the model.

4. Summary and Prospect

Image text retrieval is a dynamic multidisciplinary intersection. Its main purpose is to create models that can process and correlate information from images and text. This paper reviews the major advances in the four sub-research directions of feature extraction, alignment, fusion, and migration learning in recent years. In addition, we also discuss the difficulties and development directions that need to be solved in the future for each sub-question. However, compared with tasks such as classification and detection, the accuracy of image-text retrieval is still far from the same. In the future, the introduction of memory mechanisms and reasoning mechanisms is expected to build a more complete semantic space and achieve retrieval across datasets. We hope this article can bring some new inspiration to the field of image-text retrieval, and promote the direction of image -text retrieval to flourish.

References

[1] Yu K, Jia L, Chen Y Q, et al. Deep learning: Yesterday, today, and tomorrow[J]. Journal of Computer Research and Development, 2013, 50(9) : 1799－1804.

[2] Bengio Y, Delalleau O, Roux N L． The curse of highly variable functions for local kernel machines[C] //Annual Conference on Neural Information Processing Systems． Cambridge, USA: MIT Press, 2006: 107－114.

[3] Bengio Y, Courville A, Vincent P． Representation learning: A review and new perspectives[J]． IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8):1798－1828． K. Elissa, "Title of paper if known," unpublished.

[4] LeCun Y, Bengio Y, Hinton G． Deep learning[J]． Nature, 2015, 521(7553):436－444.

[5] Goodfellow I, Bengio Y, Courville A. Deep learning[M]. Cambridge, USA: MIT Press, 2016．

[6] Bengio Y． Learning deep architectures for AI[J]． Foundations & Trends in Machine Learning, 2009, 2(1):1－127.

[7] Montufar G F, Pascanu R, Cho K. et al． On the number of linear regions of deep neural networks[C] //Annual Conference on Neural Information Processing Systems． Cambridge., USA:MIT Press, 2014: 2924－2932．

[8] Yang Yi, Wang Shengkai, Chen Guoshun, et al., Inter-media Information Technology and Applications. Beijing: Publishing House of Electronics Industry, 2014.

[9] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G.Lanckriet， Roger Levy, Nuno Vasconcelos. A New Approach to Cross-Modal Multimedia Retrieval [C]. Proc. ACM. International Conference on Multimedia, 2010， pp. 251-260.

[10] He Ning.Research on the method of obtaining cross-modal semantic information in image retrieval[D].Wuhan: Wuhan University, 2013. 36-44.

[11] Harold Hotelling. Relations between Two Sets of Variates [M]. Biometrika, 1936, 28(3/4)：321-377.

[12] David R. Hardoon, Sdndor Szedmik, John Shawe-Taylor. Canonical Correlation Analysis： an Overview with Application to Learning Methods [J]. Neural Computation, 2004, 16(12)：2639-2664.

[13] H. Hotelling, "Relations between two sets of variates," Biometrika, vol. 28, no. 3/4, pp. 321–377, 1936.

[14] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in ACM International Conference on Multimedia (ACM MM), 2010, pp. 251–260.

[15] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 36, no. 3, pp. 521–535, 2014.

[16] M. Belkin, I. Matveeva, and P. Niyogi, "Regularization and semisupervised learning on large graphs," Learning theory. Springer Berlin Heidelberg, pp. 624–638, 2004.

[17] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in AAAI Conference on Artificial Intelligence (AAAI), 2013, pp. 1198–1204.

[18] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semi-supervised regularization," IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), vol. 24, no. 6, pp. 965–978, 2014.

[19] Y. Peng, X. Zhai, Y. Zhao, and X. Huang, "Semi-supervised crossmedia feature learning with unified patch graph regularization," IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), vol. 26, no. 3, pp. 583–596, 2016.

[20] N. Quadrianto and C. H. Lamppert, "Learning multi-view neighborhood preserving projections," in International Conference on Machine Learning (ICML), 2011, pp. 425–432.

[21] B. McFee and G. Lanckriet, "Metric learning to rank," in International Conference on Machine Learning (ICML), 2010, pp. 775–782.

[22] W. Wu, J. Xu, and H. Li, "Learning similarity function between objects in heterogeneous spaces," in Microsoft Research Technique Report, 2010.

[23] D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 30, no. 8, pp. 1371–1384, 2008.

[24] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang, "Cross-media semantic representation via bi-directional learning to rank," in ACM International Conference on Multimedia (ACM MM), 2013, pp. 877– 886.

[25] X. Jiang, F. Wu, X. Li, Z. Zhao, W. Lu, S. Tang, and Y. Zhuang, "Deep compositional cross-modal learning to rank via local-global alignment," in ACM International Conference on Multimedia (ACM MM), 2015, pp. 69–78.

[26] F. Wu, X. Jiang, X. Li, S. Tang, W. Lu, Z. Zhang, and Y. Zhuang, "Cross-modal learning to rank via latent joint representation," IEEE Transactions on Image Processing (TIP), vol. 24, no. 5, pp. 1497– 1509, 2015.

[27] Y. Jia, M. Salzmann, and T. Darrell, "Factorized latent spaces with structured sparsity," in Advances in Neural Information Processing Systems (NIPS), 2010, pp. 982–990.

[28] F. Zhu, L. Shao, and M. Yu, "Cross-modality submodular dictionary learning for information retrieval," in ACM International Conference on Conference on Information and Knowledge Management (CIKM), 2014, pp. 1479–1488.

[29] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2216–2223.

[30] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in AAAI Conference on Artificial Intelligence (AAAI), 2013, pp. 1070– 1076.

[31] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang, "Sparse multi-modal hashing," IEEE Transactions on Image Processing (TIP), vol. 16, no. 2, pp. 427–439, 2014.

[32] D. Zhang, F. Wang, and L. Si, "Composite hashing with multiple information sources," in International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2011, pp. 225–234.

[33] M. Long, Y. Cao, J. Wang, and P. S. Yu, "Composite correlation quantization for efficient multimodal retrieval," in International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2016, pp. 579–588.

[34] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modalitymetric learning using similarity-sensitivehashing,"in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 3594–3601.

[35] Y. Zhen and D.-Y. Yeung, "A probabilistic model for multimodal hash function learning," in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2012, pp. 940– 948.

[36] Y. Zhuang, Z. Yu, W. Wang, F. Wu, S. Tang, and J. Shao, "Cross-media hashing with neural networks," in ACM International Conference on Multimedia (ACM MM), 2014, pp. 901–904.

[37] D. Wang, P. Cui, M. Ou, and W. Zhu, "Deep multimodal hashing with orthogonal regularization," in International Joint Conference on Artificial Intelligence (IJCAI), 2015, pp. 2291–2297.
[38] Y. Peng, J. Qi, X. Huang, and Y. Yuan. CCL: cross-modal correlation learning with multi-grained fusion by hierarchical network. IEEE Transactions on Multimedia (TMM), 20(2): 405–420, 2018.
[39] F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3441–3450, 2015.
[40] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan. Cross-modal retrieval with CNN visual features: A new baseline. IEEE Transactions on Cybernetics (TCYB), 47(2):449–460, 2017.
[41] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen. Adversarial cross-modal retrieval. In ACM International Conference on Multimedia (ACM MM), pages 154–162, 2017.
[42] Y. Peng, X. Huang, and J. Qi. Cross-media shared representation by hierarchical learning with multiple deep networks. In International Joint Conference on Artificial Intelligence (IJCAI), pages 3846–3853, 2016.
[43] Huang X, Peng Y. Deep Cross-media Knowledge Transfer[J]. 2018.
[44] Yuan M, Peng Y. Text-to-image Synthesis via Symmetrical Distillation Networks[C]// 2018 ACM Multimedia Conference. ACM, 2018.
[45] Qi J, Peng Y, Yuan Y . Cross-media Multi-level Alignment with Relation Attention Network[J]. 2018.
[46] Qi, J., Peng, Y., & Zhuo, Y. (2018). Life-long Cross-media Correlation Learning. 2018 ACM Multimedia Conference on Multimedia Conference - MM '18.
[47] Yuxin P, Jinwei Q, Yuxin Y . Modality-specific Cross-modal Similarity Measurement with Recurrent Attention Network[J]. IEEE Transactions on Image Processing, 2018:1-1.
[48] Chitra Dorai， Svetha Venkatesh. Computational Media Aesthetics： Finding Meaning Beautiful [J]， IEEE Multimedia， 2001， 8(4)：10-12.
[49] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in International Conference on Machine Learning (ICML), 2011, pp. 689–696.
[50] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in Advances in Neural Information Processing Systems (NIPS), 2012, pp. 2222–2230.
[51] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in International Conference on Machine Learning (ICML), 2010, pp. 3408–3415.
[52] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3441–3450.
[53] Jiang Q Y, Li W J. Deep Cross-Modal Hashing[C]// IEEE Conference on Computer Vision & Pattern Recognition. 2017.
[54] Xinyang Jiang, Fei Wu, Xi Li, Zhou Zhao, Weiming Lu, Siliang Tang, Yueting Zhuang: Deep Compositional Cross-modal Learning to Rank via Local-Global Alignment. ACM Multimedia 2015: 69-78.
[55] Cheng Deng, Xu Tang, Junchi Yan, Wei Liu, Xinbo Gao: Discriminative Dictionary Learning With Common Label Alignment for Cross-Modal Retrieval. IEEE Trans. Multimedia 18(2): 208-218 (2016).
[56] N. Rasiwasia et al., "A new approach to cross-modal multimedia retrieval," in Proc. ACM Int. Conf. Multimedia (ACM MM), 2010, pp. 251–260.
[57] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nuswide: A real-world Web image database from national University of Singapore," in Proc. ACM Int. Conf. Image Video Retr. (CIVR), 2009, p. 48.
[58] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, Sep. 2009.
[59] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Jiebo Luo: Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. ACM Multimedia 2017: 795-816.