Attention-Sharing Correlation Learning for Cross-Media Retrieval

Xin Huang, Zhaoda Ye, and Yuxin Peng(✉)

Institute of Computer Science and Technology, Peking University,
Beijing 100871, China
pengyuxin@pku.edu.cn

Abstract. Cross-media retrieval is a challenging research topic with wide prospect of application, aiming to retrieve among different media types by using a single-media query. The main challenge of cross-media retrieval is to learn the correlation between different media types for addressing the issue of "media gap". The close semantic correlation usually lies in specific parts of cross-media data such as image and text, which plays the key role for precious correlation mining. However, existing works usually focus on correlation learning in the level of whole media instance, or adopt patch segmentation but treat the patches indiscriminately. They ignore the fine-grained discrimination learning, which limits the retrieval accuracy. Inspired by attention mechanism, this paper proposes the attention-sharing correlation learning network, which is an end-to-end network to generate cross-media common representation for retrieval. By sharing the common attention weights, the attention of different media types can be learned coordinately. It can not only emphasize the single-media discriminative parts, but also enhance the cross-media fine-grained consistent pattern, and so learn more precious cross-media correlation to improve retrieval accuracy. Experimental results on 2 widely-used datasets with state-of-the-art methods verify the effectiveness of the proposed approach.

1 Introduction

As a key technique of information acquisition and management, multimedia retrieval has become an active research topic for decades [1], which can provide amounts of similar data with a single query. Past efforts mainly concentrate on single-media retrieval, where user query and retrieval results are of the same media type. However, with the development of multimedia and network transmission technology, multimedia data such as image, text, video and audio can be generated and found everywhere. Different media types have been merged with each other, and become the main form of big data. Under this situation, the media limitation of single-media retrieval is becoming increasingly obvious, and cross-media retrieval has become a new important retrieval paradigm.

Cross-media retrieval is proposed to retrieve data of similar semantic but different media types with a user query. Intuitively, it allows user to retrieve

Y. Zhao et al. (Eds.): ICIG 2017, Part I, LNCS 10666, pp. 477–488, 2017.
https://doi.org/10.1007/978-3-319-71607-7_42

relevant texts with an image query. Different from single-media retrieval, cross-media retrieval faces the great challenge of "media gap", which means that data of different media types have different representation forms. For example, image can be represented by features based on visual information as color and texture, while text can be represented by features based on word frequency. Representations of different media types lie in different feature space, so the similarity between them cannot be directly measured. For addressing this problem, the mainstream methods of cross-media retrieval are common representation learning. The main idea is to represent data of different media types with the same type of representation, so that cross-media similarity can be directly computed by distance measurement. Based on different models, these methods can be further divided into non-DNN based learning methods [2–4] and DNN-based methods [5–8]. They all project cross-media data into one common space by learning from their correlation.

The close semantic correlation usually lies in specific parts of cross-media data. For example, the correlation between image and text can be co-existent patterns of image patches and words. The above fine-grained correlation plays the key role for precious correlation mining. However, existing works usually focus on correlation in the level of whole media instance [2,3,9], and ignore the fine-grained information. Some recent works as [10,11] adopt patch segmentation and treat the patches indiscriminately. For example, the work of [10] takes all the patches for hypergraph construction, and all the patches are equally important. However, the importances of different parts are usually different, and they can be very noisy in semantic level, limiting the effectiveness of correlation learning.

For addressing the above problem, inspired by attention mechanism [12–14], this paper proposes the attention-sharing correlation learning network (ACLN), which is an end-to-end network to generate cross-media common representation for retrieval. ACLN first extracts local features from cross-media data, and then lets them share the common attention weights, so that the attention of different media types can be learned coordinately according to pairwise correlation and semantic information. It can not only emphasize the single-media discriminative parts, but also enhance the cross-media fine-grained consistent pattern, and so learn more precious cross-media correlation to improve retrieval accuracy. Experimental results on 2 widely-used datasets with state-of-the-art methods verify the effectiveness of the proposed approach.

2 Related Work

2.1 Cross-Media Retrieval

Cross-media retrieval is designed to retrieve among different media types. As discussed in Sect. 1, the current mainstream methods can be summarized as common representation learning, including non-DNN based methods and DNN-based methods. These methods follow the idea that although representations of different media types are different, they share the same commons on semantic

description. So in the semantic level, different media types can be represented in the same common space, leading to cross-media common representation.

Non-DNN based methods mainly learn linear projection for different media types. For example, canonical correlation analysis (CCA) [2] learns cross-media representation by maximizing the pairwise correlation, and is a classical baseline method for various cross-media problems as [15,16]. An alternative method is cross-modal factor analysis (CFA) [17], which minimizes the Frobenus norm of pairwise common representation. Beyond pairwise correlation, joint representation learning [3] is proposed to make use of semi-supervised regularization and semantic information, which can jointly learn common representation projections for up to five media types.

Instead of linear projection, DNN-based methods take deep neural network as the basic model for generating cross-media common representation. Recent years, DNN-based cross-media retrieval has become an active research topic, and many methods have been proposed [5,6,8,9]. For example, the architecture of Bimodal AE [5] takes two modalities as input, and has a middle code layer for common representation. CMDN [8] is proposed to simultaneously consider inter-modality and intra-modality information in a hierarchical multi-network architecture, which improves the retrieval accuracy. Wei et al. [9] propose to use CNN pre-trained with ImageNet as the feature extractor for images, and show the effectiveness of CNN feature in cross-media retrieval.

Existing methods usually focus on correlation in the level of whole media instance [2,3,9]. They take the whole media instances as input and learn correlation among them. However, close semantic correlation usually lies in specific parts of cross-media data, instead of whole data. The above methods ignore the fine-grained information. Note that some recent works as [10,11] first adopt instance segmentation to obtain several patches for media instances, and then use these patches indiscriminately as input. However, the importances of different parts are usually different. Taking text as example, not all words or sentences are semantically discriminative and have strong correlation with other media types. Some of them can even contain noisy information, and limit the effectiveness of correlation learning.

2.2 Attention Mechanism

Attention mechanism aims to find the "important" parts within a whole media instance, which has been applied to image and language processing. For example, visual attention models can select and focus on the regions containing discriminative information, such as the work of [12] which selects the regions that by recurrent attention model for multiple object recognition. Similarly, textual attention models are proposed to find the alignments between input and output text for helping deal with long-term dependency. Such methods have been applied to problems like question answering [13] and text generation [18].

Attention mechanism has also widely-used in problems involving multimedia, such as image caption [19] and visual QA [14,20]. For example, Lu et al. [20] propose the method of co-attention, which integrates visual and textual attention

to guide each other for better natural symmetry between image and question. Note that the attention weights of text and image are different in [20], and it needs image and text as input at the same time, so cannot support cross-media retrieval.

Our proposed ACLN approach takes an attention-sharing strategy for cross-media retrieval. That is to say, the inputs of image and text share the common attention weights to enhance the cross-media fine-grained consistent pattern, which helps learn better common representation for cross-media retrieval.

3 Attention-Sharing Correlation Learning

In this paper, we take image and text as examples to show the ACLN approach, while it can be applied for other media types. The overview of our ACLN is shown as Fig. 1, which can be viewed as an end-to-end architecture with three parts, namely (1) local feature extraction, (2) attention-sharing learning, and (3) common representation generation.

For training stage, there are two types of cross-media correlation considered by ACLN. The first is co-existence relationship (specifically pairwise correlation in this paper), which means that data of different media types exist as a whole and have close relevance; the second is common semantic information, which means that data in each pair have the same semantics, i.e., they share the same semantic label. For testing stage, image or text can serve as input independently, and ACLN can generate common representation for them to perform cross-media retrieval with distance measurement.

We denote training data as $D_{tr} = \{D^I_{tr}, D^T_{tr}\}$, where $D^I_{tr} = \{i_p, y_p\}_{p=1}^{n_{tr}}$, and $D^T_{tr} = \{t_p, y_p\}_{p=1}^{n_{tr}}$. i_p and t_p means p-th paired image and text data, y_p means their shared label, and n_{tr} denotes the number of training pairs. Testing data is denoted as $D_{te} = \{D^I_{te}, D^T_{te}\}$, where $D^I_{te} = \{i_p\}_{p=1}^{n_{te}}$, $D^T_{te} = \{t_p\}_{p=1}^{n_{te}}$, and n_{te} means the number of testing data. The aim of ACLN is to generate common representation for D^I_{te} and D^T_{te}.

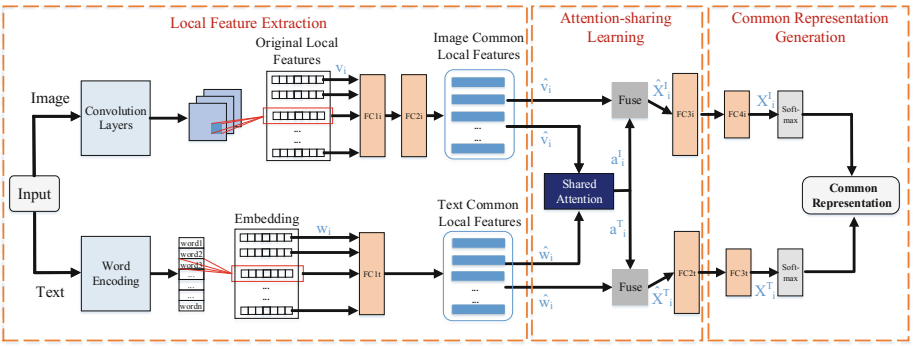

Fig. 1. An overview of our attention-sharing correlation learning network (ACLN).

3.1 Local Feature Extraction

The part of local feature extraction aims to extract the fine-grained representation for image regions and text words, and these fine-grained representations will be further fed into the attention-sharing learning part.

For a text t_p, it is encoded as 1-hot representations (vectors with only one dimension set as 1 and the others set as 0) of words and get $H(t_p) = \{w_1, ..w_T\}$ following [20], where T is the word number in t_p, $w_i \in R^{V*1}$ and V is the vocabulary size of all texts. Then we can embed each word to a representation vector as follows:

$$\hat{w}_i = W_e w_i, W_e \in R^{d_a * V} \tag{1}$$

where W_e is the weight parameters learned in the training stage of the network.

So we have the local features of text with the activation function *tanh*:

$$L(t_p) = \{tanh(\hat{w}_1), .., tanh(\hat{w}_T)\} \tag{2}$$

For an image input, we use the convolutional layers to get the image feature maps. In this paper, we take AlexNet [21] as the basic model for image. The original local features of image are denoted as $C(i_p) = \{v_1, .., v_N\}, v_i \in R^{d_i * 1}$, where the v_i is a feature vector extracted from the feature maps in spatial regions i, and N is the number of regions. Specially, we use the output of the pool5 as image local feature, and construct v_i with the value of spatial region i in each feature map. Then the local feature vectors will pass through a fully-connected layer which maps them to the same dimension as the text features (i.e., d_a here). So we have the image local features $L(i_p)$:

$$\hat{v}_i = tanh(W_{pre} v_i), W \in R^{d_a * d_i} \tag{3}$$

$$L(i_p) = \{\hat{v}_1, .., \hat{v}_N\} \tag{4}$$

where W_{pre} is the parameter of the fully-connected layer (FC1i).

At last, both the local features of images ant texts will pass through a fully-connected layer (FC2i and FC1t in Fig. 1) to convert them as *common local features*. For convenience, here we take $\hat{w}$ and $\hat{v}$ as the common local features, and have:

$$X^{\hat{w}} = tanh(W_{fc2i} \hat{w}) \tag{5}$$

$$X^{\hat{v}} = tanh(W_{fc1t} \hat{v}) \tag{6}$$

where $W_{fc2i}, W_{fc1t} \in R^{d_a * d_a}$ are the weight parameters of the fully connected layers FC2i and FC1t.

3.2 Attention-Sharing Learning

In this part, we adopt an attention-sharing structure to select the common local features which capture the correlation between two media types and then fuse to get global features.

Briefly, we adopt an attention function shared by common local features of both images and texts, and generate the attention weight vector for fusing the common local features to be the global features. Here we let $X = \{x_1, .., x_n\}$ be either the common local features of text or image for convenience, where n denotes the total number of common local features of an image or text, so we have:

$$h_i = tanh(W_a x_i) \tag{7}$$

$$a_i = \frac{e^{h_i}}{\sum_{k=1}^{n} e^{h_k}} \tag{8}$$

$$\hat{X} = \sum_{i=1}^{n} a_i x_i \tag{9}$$

where $W_a \in R^{1*d_a}$ is the weight parameter shared by all the common local features as the attention weight, which is learned to capture the fine-grained correlation between images and texts. And $\hat{X}$ is the global feature to the input X.

Note that the attention-sharing structure is to capture the fine-grained consistent patterns between different media types. Because the input is paired data, we simply assume that they share the relevant global semantics. With this in mind, we adopt the constraint that paired instances will have similar global features, letting the fusion process focus more on the local features with close correlation.

Specifically, we use the cosine similarity as the risk. For the paired global image feature $\hat{X}_i^I$ and text feature $\hat{X}_i^T$, the discrepancy of the paired feature is defined as:

$$d(\hat{X}_i^I, \hat{X}_i^T) = \frac{< \hat{X}_i^I, \hat{X}_i^T >}{\|\hat{X}_i^I\| \|\hat{X}_i^T\|} \tag{10}$$

Then we have the correlation loss as:

$$L_{corr} = \lambda \sum_{1}^{n} d(\hat{X}_i^I, \hat{X}_i^T) \tag{11}$$

where $\lambda > 0$ is a penalty parameter of the correlation loss.

3.3 Common Representation Generation

In this part, we use two fully-connected layers to obtain the final representation with the labels. Both the image and text global features will pass through two fully-connected layers and a softmax layer to generate the common representations. The semantic loss is defined as:

$$L_{Se} = \frac{1}{n} \sum_{i=1}^{n} f_s(X_i^I, L_i, \theta) + f_s(X_i^T, L_i, \theta) \tag{12}$$

where $f_s(X, L, \theta)$ is the softmax loss function:

$$f_s(X, L, \theta) = -log \frac{e^{\theta_L X}}{\sum_{j=1}^{c} e^{\theta_j X}} \tag{13}$$

where X is the output of the last fully-connected layer with an instance, L is the label of the instance, c is the total category number of the data and θ is the parameter of the network.

It should be noted that the proposed ACLN is an end-to-end network, and the correlation loss (L_{corr}) and semantic loss (L_{Se}) can be considered jointly. In training stage, by optimization with RMSProp, we can minimize the total loss to train the whole network. In the testing stage, we use the predicted probability vectors as the final common representation for performing cross-media retrieval.

4 Experiments

This section presents the experiments for verifying the effectiveness of the proposed method. We adopted 2 widely-used cross-media datasets and 7 compared methods with 2 retrieval tasks in our experiments.

4.1 Details of the Deep Architecture

In the implementation, we adopt Torch to develop our model. We use the Rnsorop optimizer with a base learning rate $4e-4$, momentum 0.99 and weight-decay $1e-8$, and set the batch size to be 20. Particularly, the learning rate of FC3i and FC4i is set to be $4e-5$ on the Wikipedia dataset. The five convolutional layers of AlexNet are pre-trained with ImageNet from the Caffe Model Zoo and fine-tuned with the images in each dataset. In the training stage, the weights of the convolutional are frozen. The text local feature (and the common local feature) is a 512-dimentional vector after embedding, i.e. $d_a = 512$. The original image local feature is a 256-dimensional vector so that $d_i = 256$. We also apply dropout with probability 0.5 on each layer. We use the CosineEmbeddingCriterion layer to calculate the correlation with the margin as 0 and penalty λ as 1.

4.2 Dataset Introduction

This section introduces the 2 datasets adopted for the experiments, namely Wikipedia dataset and NUS-WIDE-10k dataset.

Wikipedia dataset [15] is widely-used for cross-media retrieval evaluation as [3,7]. It is based on "featured articles" in Wikipedia which contains 2,866 image/text pairs with 10 high-level semantic categories. In each pair, the text describes the image with several paragraphs, so they have close correlation. Following [7], the dataset is randomly split into three parts: 2,173 pairs are selected as training set, 462 pairs are selected as testing set, and 231 pairs are used for as validation set.

NUS-WIDE-10k dataset [7] is a subset of NUS-WIDE dataset [22]. NUS-WIDE dataset contains about 270,000 images with several corresponding tags which are regarded as text in the experiments. NUS-WIDE-10k dataset is constructed with 10,000 image/text pairs which are randomly selected from 10 largest categories in NUS-WIDE dataset and each category has 1,000 pairs of images and text. Following [7], the dataset is randomly split into three parts: 8,000 pairs for training, 1,000 pairs for testing and 1000 pairs for validation.

It should be noted that although the splits of these datasets have validation sets, the ACLN and compared methods don't need validation sets as input. That is to say, validation sets will not be used in the whole experiments.

4.3 Compared Methods and Input Settings

Totally 7 state-of-the-art methods are compared in the experiments: CCA [2], CFA [17], KCCA (with Gaussian kernel) [23], JRL [3], LGCFL [4], Corr-AE [7], and Deep-SM [9]. Among these, CCA, CFA, KCCA, JRL, LGCFL are non-DNN based methods, while Corr-AE and Deep-SM are DNN-based methods. Note that Deep-SM is also an end-to-end DNN-based method.

For image, the processing is end-to-end in ACLN, and it directly takes the image pixels as input. Deep-SM also takes original images as input. However, all they other methods including CCA, CFA, KCCA, JRL, LGCFL and Corr-AE can only take feature vector as input. For them we take the same fine-tuned AlexNet adopted by ACLN, and further fine-tuned to convergence with the images. Then we extract the output of the FC7 layer in the AlexNet as the feature vector. For text, ACLN also has the end-to-end processing ability, and takes the original text as input. For all the compared methods, we train a basic ACLN network which simply averages the common local features without attention and then extract the output of the FC2t layer as the feature vector.

4.4 Evaluation Metrics

Two retrieval tasks are conducted in the experiments: text retrieval by image query, and image retrieval by text query, which are briefly denoted as Image→Text and Text→Image. We first obtain the common representation for all testing images and text with all compared methods and our ACLN. Then taking Image→Text task as example, we take each image as query, and measure the cosine distance between the common representation of the query image and all texts. Finally, we get a ranking list according to the distances and then compute the mean average precision (MAP) for it to evaluate the retrieval results.

We choose MAP score as the evaluation metric because it jointly considers the precision and ranking of results, and it can be used for fair and comprehensive evaluation. The MAP scores are computed as all queries' mean of average precision (AP), and AP is computed as:

$$AP = \frac{1}{R} \sum_{k=1}^{n} \frac{R_k}{k} \times rel_k \tag{14}$$

where R denotes relevant item number in test set (according to the label in our experiments), R_k denotes the relevant item number in top k results, n denotes the test set size, and $rel_k = 1$ means the k-th result is relevant, and 0 otherwise.

4.5 Experimental Results

Table 1 shows the MAP scores in our experiments on the 2 datasets. On Wikipedia dataset, ACLN achieves the highest MAP score of 0.430. Comparing with the best compared method Deep-SM, ACLN obtains an inspiring improvement of 0.036. Similar trends can be seen on NUS-WIDE-10k dataset, where our ACLN remains the highest MAP score of 0.487. This is because that the compared methods only focus on correlation in the level of whole media instance, and ignore the fine-grained information. ACLN can not only emphasize the single-media discriminative parts, but also enhance the cross-media fine-grained consistent pattern, and so learn more precious cross-media correlation to improve retrieval accuracy.

Table 2 shows the MAP scores of our baselines and the complete ACLN. ACLN (Baseline) means that the network is trained without the attention which simply averages the common local features. ACLN (Separate Attention) means that network is trained with separate attention which adopts independent attention weights for images and text. Except for the above differences, the rest parts of the three baselines keep the same with complete ACLN.

Table 1. MAP scores of our ACLN and compared methods.

Dataset	Method	Task		
		Image→Text	Text→Image	Average
Wikipedia dataset	CCA	0.125	0.124	0.124
	Corr-AE	0.188	0.202	0.195
	CFA	0.368	0.336	0.352
	KCCA	0.340	0.316	0.328
	JRL	0.371	0.330	0.351
	LGCFL	0.390	0.321	0.356
	Deep-SM	0.441	0.347	0.394
	Our ACLN	**0.446**	**0.415**	**0.430**
NUS-WIDE -10k dataset	CCA	0.121	0.122	0.121
	Corr-AE	0.185	0.143	0.164
	CFA	0.407	0.411	0.409
	KCCA	0.402	0.427	0.415
	JRL	0.442	0.473	0.457
	LGCFL	0.421	0.440	0.431
	Deep-SM	0.465	0.445	0.455
	Our ACLN	**0.480**	**0.495**	**0.487**

Table 2. MAP scores of our ACLN and the baselines.

Dataset	Method	Task		
		Image→Text	Text→Image	Average
Wikipedia dataset	ACLN (Baseline)	0.436	0.351	0.394
	ACLN (Separate Attention)	0.429	0.396	0.413
	Our ACLN	**0.446**	**0.415**	**0.430**
NUS-WIDE -10k dataset	ACLN (Baseline)	0.458	0.454	0.456
	ACLN (Separate Attention)	0.470	0.487	0.479
	Our ACLN	**0.480**	**0.495**	**0.487**

It can be seen that the results of ACLN (Separate Attention) are better than ACLN (Baseline), which shows that the attention mechanism helps provide fine-grained clues for improving the accuracy of cross-media retrieval. The complete ACLN is even better than ACLN (Separate Attention), which shows that the attention-sharing structure enhances the cross-media fine-grained consistent pattern for higher retrieval accuracy. The above baseline experiments show the separate contribution of our ACLN architecture, and verify its effectiveness.

5 Conclusion

This paper has proposed the attention-sharing correlation learning network (ACLN), which is designed to generate cross-media common representation with fine-grained discrimination learning for cross-media retrieval. ACLN first extracts local features from cross-media data, and then lets them share the common attention weights, so that the attention of different media types can be learned coordinately according to pairwise correlation and semantic information. It can not only emphasize the single-media discriminative parts, but also enhance the cross-media fine-grained consistent pattern, and so learn more precious cross-media correlation to improve retrieval accuracy. Experimental results on 2 widely-used datasets with state-of-the-art methods verify the effectiveness of the proposed approach. The future work lies in two aspects: first, we intend to incorporate the attention learning of more than two media types into our framework; second, we will apply ACLN to other applications like image caption to further verify its effectiveness.

Acknowledgments. This work was supported by National Natural Science Foundation of China under Grants 61371128 and 61532005.

References

1. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: state of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl. (TOMCCAP) **2**(1), 1–19 (2006)

2. Hotelling, H.: Relations between two sets of variates. Biometrika **28**(3/4), 321–377 (1936)
3. Zhai, X., Peng, Y., Xiao, J.: Learning cross-media joint representation with sparse and semi-supervised regularization. IEEE Trans. Circ. Syst. Video Technol. (TCSVT) **24**(6), 965–978 (2014)
4. Kang, C., Xiang, S., Liao, S., Xu, C., Pan, C.: Learning consistent feature representation for cross-modal multimedia retrieval. IEEE Trans. Multimedia (TMM). **17**(3), 370–381 (2015)
5. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: International Conference on Machine Learning (ICML), pp. 689–696 (2011)
6. Srivastava, N., Salakhutdinov, R.: Learning representations for multimodal data with deep belief nets. In: International Conference on Machine Learning Workshop (2012)
7. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: ACM International Conference on Multimedia (ACM MM), pp. 7–16 (2014)
8. Peng, Y., Huang, X., Qi, J.: Cross-media shared representation by hierarchical learning with multiple deep networks. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 3846–3853 (2016)
9. Wei, Y., Lu, C., Wei, S., Liu, L., Zhu, Z., Yan, S.: Cross-modal retrieval with CNN visual features: a new baseline. IEEE Trans. Cybern. (TCYB) **47**(2), 449–460 (2017)
10. Peng, Y., Zhai, X., Zhao, Y., Huang, X.: Semi-Supervised cross-media feature learning with unified patch graph regularization. IEEE Trans. Circ. Syst. Video Technol. (TCSVT) **26**(3), 583–596 (2016)
11. Peng, Y., Qi, J., Huang, X., Yuan, Y.: CCL: cross-modal correlation learning with multi-grained fusion by hierarchical network (2017). arXiv:1704.02116
12. Ba, J., Mnih, V., Kavukcuoglu, K.: Multiple object recognition with visual attention (2014). arXiv:1412.7755
13. Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R.: Ask me anything: dynamic memory networks for natural language processing. In: International Conference on Machine Learning (ICML), pp. 1378–1387 (2016)
14. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J.: Stacked attention networks for image question answering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21–29 (2016)
15. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: ACM International Conference on Multimedia (ACM MM), pp. 251–260 (2010)
16. Ranjan, V., Rasiwasia, N., Jawahar, C.V.: Multi-label cross-modal retrieval. In: IEEE International Conference on Computer Vision (ICCV), pp. 4094–4102 (2015)
17. Li, D., Dimitrova, N., Li, M., Sethi, I.K.: Multimedia content processing through cross-modal association. In: ACM International Conference on Multimedia (ACM MM), pp. 604–611 (2003)
18. Li, J., Luong, M.-T., Jurafsky, D.: A Hierarchical Neural Autoencoder for Paragraphs and Documents. In: The Association for Computer Linguistics (ACL), pp. 1106–1115 (2015)

19. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning (ICML), pp. 2048–2057 (2015)
20. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Annual Conference on Neural Information Processing Systems (NIPS), pp. 289–297 (2016)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Annual Conference on Neural Information Processing Systems (NIPS), pp. 1106–1114 (2012)
22. Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: a real-world web image database from National University of Singapore. In: ACM International Conference on Image and Video Retrieval (CIVR), p. 48 (2009)
23. Hardoon, D.R., Szedmák, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. Neural Comput. **16**(12), 2639–2664 (2004)