Deep Convolutional Neural Network for Bidirectional Image-Sentence Mapping

Tianyuan Yu(✉), Liang Bai, Jinlin Guo, Zheng Yang, and Yuxiang Xie

College of Information System and Management,
National University of Defense Technology, Changsha 410073, China
{yutianyuan92, xabpz}@163.com, gjlin99@gmail.com,
yz_nudt@hotmail.com, yxxie@nudt.edu.cn

Abstract. With the rapid development of the Internet and the explosion of data volume, it is important to access the cross-media big data including text, image, audio, and video, etc., efficiently and accurately. However, the content heterogeneity and semantic gap make it challenging to retrieve such cross-media archives. The existing approaches try to learn the connection between multiple modalities by direct utilization of hand-crafted low-level features, and the learned correlations are merely constructed with high-level feature representations without considering semantic information. To further exploit the intrinsic structures of multimodal data representations, it is essential to build up an interpretable correlation between these heterogeneous representations. In this paper, a deep model is proposed to first learn the high-level feature representation shared by different modalities like texts and images, with convolutional neural network (CNN). Moreover, the learned CNN features can reflect the salient objects as well as the details in the images and sentences. Experimental results demonstrate that proposed approach outperforms the current state-of-the-art base methods on public dataset of Flickr8K.

1 Introduction

The ubiquitous adoption of mobile Internet has made multimedia documents available everywhere in daily life in forms of web pages, images, videos, and even mobile services like interactive micro-blogs, social networks, etc., which are usually composed of multimedia formats and content descriptions. Meanwhile, the rapid increase of data volume also makes it more and more difficult for web users to access valuable and customized information for the massive information oceans. The above difficulty has triggered much attention to information retrieval approaches in research communities.

Cross-media information retrieval is challenging because of the so-called *semantic gap* problem, which means the query descriptions and returned results can hardly be corresponded accurately, especially when they belong to different modalities. As a result, one key problem in this task is how to measure the distances or similarities between multiple modalities from the view of semantics. One solution is to align the two feature spaces so that they can be comparable and such semantic mapping has attracted much research interest. However, the detection of saliency such as scenes,

L. Amsaleg et al. (Eds.): MMM 2017, Part II, LNCS 10133, pp. 136–147, 2017.
DOI: 10.1007/978-3-319-51814-5_12

objects, etc. from the visual media is not enough, the task of bidirectional multimedia retrieval also requires the machine to understand the details from images, texts, etc., and more importantly, their semantic connections with each other. As shown in Fig. 1, the detection of "house" and "window" might be noisy when providing meaningful description of the image, though they take up large part of the image. A machine needs to learn the useful correlations (such as "jump" and "trampoline") and neglect unimportant visual and textual information (such as " house" and "up on a").

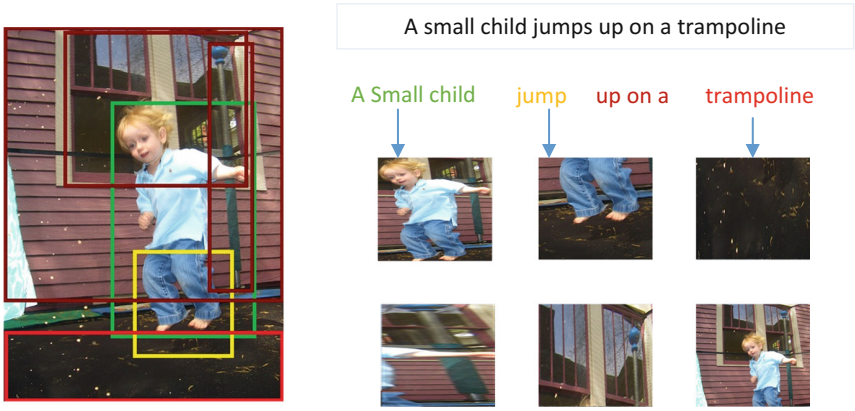

Fig. 1. Representation of the mapped image segments and a relevant sentence. The difficult is to learn the useful correlations and neglect the unimportant visual and textual information.

To deal with the above challenges, modern cross-media information retrieval approaches try to query visual media and texts alternatively, i.e. searching for relevant images with textual query, or vice versa. At the beginning of cross-media research, the task only focused on a limited number of keywords or classification tags [1]. Since one word label cannot fully represent the whole image, more recently researchers started to use long sentences or articles to search for images of interest [2, 3], and even describe a target image with appropriate captions [4]. In a more challenging task as introduced in [5], an answer can be returned through a visual Turing test when the machine is provided with an image and a corresponding textual question.

As a major breakthrough in artificial intelligence, deep learning has been successfully applied in various fields. Among the deep networks, convolutional neural network (CNN) is a typical architecture for visual feature representation [6, 7]. Compared to features extracted by traditional approaches, those derived from CNN are proved to have better performances in various computer vision tasks [8, 9] and multimedia retrieval challenges [10, 11]. Similarly in cross-media information retrieval, a large number of researchers use images labels as the targets in their networks [12, 13] aiming at classifications. Because of the limits of one word label representation, semantic details are neglected during the training process, which can definitely affect the final ranked results.

In this work, a novel deep model is introduced which learns mixed features in a common feature space from visual and textual representations respectively, and the mapped features are used to correctly determine whether the texts and images are relevant or not. Our contributions are three-fold: (1) A deep convolutional neural network which maps cross-media data into a common feature space is introduced. (2) The CNN-like model is used to analyze the textual information and extract features from textual information. (3) The attention model is combined in CNN to extract visual features from images. (4) Comprehensive evaluations in experiment demonstrate that the proposed approach differentiates from previous work in that the mixed features extracted have better representations in the common space between texts and images. In particular, the deep network achieves convincing performance on Flickr8K dataset [14] for cross-media retrieval task.

2 Related Work

Domain difference between queries and retrieval results leads to the difficulty that they are not directly comparable. This challenges cross-media information retrieval and the map of different domains to a common feature space is necessary, so that the distance between them can be measured. In this section, related work on how to model such common feature spaces is presented and discussed.

Original work in this area used low-level feature spaces to represent simple visual descriptors or linguistic keywords, separately. That is, this kind of methods are carried out in a *extract-and-combine* manner, i.e. extracting the highly correlated features in different spaces first, which are then used to construct a correlated representation in a common feature space. Though simple visual and textual features are used in these approaches, they performed well and kept the state-of-the-art results for a long time in the past. Representative methods in this category include cross-media hashing [13], canonical correlation analysis [15] and its extension [16].

The defect of above extract-and-combine approaches is obvious in that simple features cannot represent the semantic meaning correctly, leaving the semantic gap still unbridged between different modalities. As a result, advanced semantic features are proposed and extracted to construct the mid-level feature space so as to improve the performance. The most popular method in this category is multimodal topic model [17]. Similarly in [18], Latent Dirichlet Allocation (LDA) is used to build better mapping between texts and images by Blei and Jordan. However, LDA method only works well when the features are discrete, such as traditional bag-of-words features, and is not flexible enough to be adapted to other advanced features. In [19], a mutual semantic space is proposed by Pereira et al. in which texts and images are mapped to a pre-defined vocabulary of semantic concepts according to probabilities in order to utilize the underlying semantic information more directly. Based on the probabilities representation, the distance between texts and images can be measured. Because this method highly depends on manual annotations for learning the semantic concepts, it is less flexible when a new dataset is given. In such cases, a new vocabulary has to be made manually, which is undoubtedly time consuming and labor intensive.

Recently, deep learning methods are also applied in this area aiming at developing a common feature space with the learned features. In [20], a deep visual-semantic embedding model is introduced to identify visual objects using labeled images as well as semantic information gleaned from unannotated textual corpus. Similarly, Socher et al. propose a dependency tree recursive neural network (DT-RNN) to process textual information [12]. Among these methods, recurrent or recursive neural networks are used to deal with textual information and inner product is employed to strictly measure the correspondence between cross-media features to describe similarity/relevance. Except Karpathy's method, other models reason about objects only on global level. Because the information extracted from images or texts are usually represented at global level, such as background or salient objects in an image, the inner product with global features can cause inevitable mistakes, especially when the extracted keywords may not match the saliency in the image, as discussed in Fig. 1. In [21], Karpathy et al. propose a model which works on a finer level and embeds fragments of images and sentences into a common space. Though the state-of-the-art is achieved, sentence fragments are not always appropriate, especially when it comes to multiple adjectives for one noun or numeral, as they mention in [21]. Furthermore, it is hard to correspond image fragments with words or phrases in the relevant sentence. Instead, our model focuses on both local and global features in images and sentences. The proposed mixed features are demonstrated to be better compared to previous global methods.

3 Two-Stream Deep Network

The aim of this paper is to construct a deep learning model, automatically finding the semantic similar pair of images and sentences close to each other in this common space. For this purpose, a novel two-stream deep model is introduced to extract the mixed features and correctly determine the relevance relationship based on this new representation, as shown in Fig. 2.

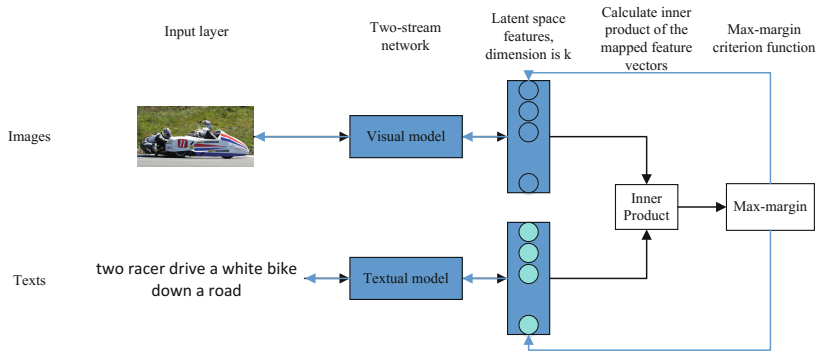

Fig. 2. Paradigm of the proposed two-stream model. Textual and visual features are extracted separately first and combined into a feature space in which max margin is used to optimize the relevance relationship.

The proposed two-stream network consists of three main components: (1) Textual Model (T-model), is responsible for training textual data with CNN and extracting the textual features. (2) Visual Model (P-model), is responsible to map the images into a common space where textual information has already been embedded. (3) Multi-Modal Embedding, involves a criterion function in order to encourage the relevant pair to have a high inner product.

The proposed model is trained on a set of images and sentences with their relationship labeled as relevant or irrelevant. In the training stage, we forward propagate the whole network to map the textual and visual information into a common space. Then inner product and max margin are used in the criterion function to backward propagate the whole network with stochastic gradient descent (SGD) method to force the semantic similar cross-media information to be close to each other in the new space. Three components of the proposed model can be described in details as follows.

3.1 Textual Model

Deep semantic similarity model (DSSM) introduced in [22] has been proved to achieve significant quality improvement on automatic highlighting of keywords and contextual entity search. One advantage of this model is that it can extract local and global features from a sentence. However, the convolutional layer in this model fixed the number of words in the group of input, which limits its function in extracting potentially relevant words. For example, for the phrase "a black and white cat", it is impossible to link the adjective "black" and the noun "cat" if the group number for relevance searching is less than four. To tackle this weakness, we extend this model and relax this constraint by searching phrases with arbitrary length. The overview of our textual model is shown in Fig. 3, which is constructed as a CNN composed of hashing layer, convolutional layer, max-pooling layer, and fully-connected layer.

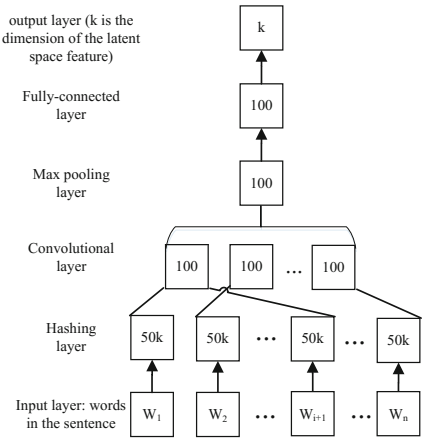

Fig. 3. Illustration of the network architecture and information flow of the textual model. The number in the rectangle represents the dimension in the layer.

As we can see from Fig. 3, the raw input of textual model includes each word in a sentence. In the hashing layer, a vector of 3-grams (tri-letter vector) is built for each word. The prominent advantage of tri-letter vector is that this representation can significantly reduce the total number of dimensions. Though English words is numerous, the number of tri-letter used to represent them can be very small. According to [23], a set of 500K-word vocabulary can boil down to only 30621 tri-letters.

After the tri-letter vectors are inputted to the convolutional layer, the local features of sentence are extracted in this layer. During the process of textual feature extraction, a sliding window is employed to concatenate words within the window to generate a new vector which is used as the input to a linear function and *tanh* activation in the last layers of the textual model. Since each word has a chance to be relevant to any other words in the sentence, the size of window is varied from one to the total number of words in the sentence. In the process, the duplicated words will increase their importance so that the extracted local features are more representative.

In the next layer of a max-pooling, the extracted feature vectors of words in sentence turns to a fixed dimension feature vector representing the sentence with the maximum operation. This is implemented by setting the i^{th} value of the output vector of max-pooling layer as the maximum value of all the i^{th} values in the input vectors. The step is to encourage the network to keep the most useful local features and form the mixed feature for each sentence. The features extracted by convolutional and max-pooling layers mainly represents the keywords and important phrases in the sentence while other useful details are kept and meaningless items are removed.

The final step of in textual model is the fully-connected layer. Like the common CNN models, there are two fully-connected layers to reduce the dimension of extracted mixed features. Going through the whole textual model, the initial sentences can be converted to vectors in a fixed-dimensional space.

3.2 Visual Model

In this section, we use the attention model originated by human visual system to extract feature from images. When people look through a picture, they usually focus on the salient parts rather than the entire image. To imitate the biological phenomenal, the attention modal is proposed to focus on different parts of the input according to different tasks.

In this work, we use the spatial transformer network introduced in [24] to focus on the visual feature in part of an image. The visual model is illustrated in Fig. 4. The input image is separated as several sections through spatial transformer network. Then, we extract the feature of each section of the input image by convolutional neural networks. Finally, the features of image sections are combined by the method of weighting. The modality of the extracted visual feature is the same with that of the extracted textual feature.

The spatial transformer network is utilized as attention model because it imitates the human visual system. The network focuses on the parts of an image, which contains much more information than others so that the useless details can be neglected. With the duplication of the more informative parts in the image, the global feature extracted

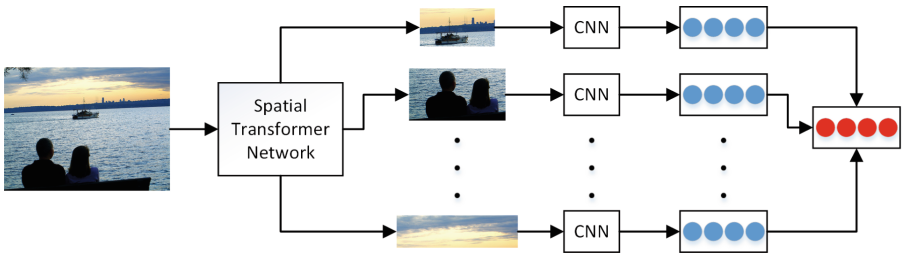

Fig. 4. Illustration of the visual model

in this paper would remain much more information delivered by the image than that extracted directly by CNN.

The spatial transformer network in our work learns four variables $\alpha_1, \alpha_2, \beta_1, \beta_2$, which makes the points (x', y') in the extracted section satisfies the Eq. (1). In the equation, (x, y) is the point in the original image. In this way, the original image can be transformed to the sections which contain much information.

$$\begin{cases} x' = \alpha_1 \bullet x + \beta_1 \\ y' = \alpha_2 \bullet y + \beta_2 \end{cases} \tag{1}$$

3.3 Multi-modal Embedding

The previous two sections of 3.1 and 3.2 have shown how the textual and visual media data can be mapped into the features with the same dimension, which means their features share a common feature space. In this section, a multi-modal objective function is defined in order to learn joint image-sentence representations. The aim of objective function is to force the corresponding pairs of images and sentences to have higher inner products than any other unrelated pairs. Since traditional classification functions such as logistic function cannot be flexibly used here to train the ranking information, we take the measure of max-margin objective function to force the difference between the inner products of correct pairs and other pairs to reach a fixed margin, which can be formalized as:

$$\begin{aligned} loss = & \sum_{(i,j)\in P} \sum_{(i,k)\notin P} \max(0, margin - v_i^T t_j + v_i^T t_k) \\ & + \sum_{(i,j)\in P} \sum_{(k,j)\notin P} \max(0, margin - v_i^T t_j + v_k^T t_j) \end{aligned} \tag{2}$$

where v_i is a column vector denoting the output of our visual model for the i-th image, t_j is a column vector representing the output of textual model for the j-th sentence. We also define P as the universal set of all the corresponding image-sentence pairs (i,j). It is obviously time-consuming if all the irrelevant cross-media information are used to optimized this model. For the purpose of efficiency, we randomly select 9 false samples

for each true sample to restrict the scale of training dataset. The hyper-parameter margin is usually set around 1. However, the range of the variable is wide, for example, it is set to 3 in [12] while 0.1 in [20]. In this paper, the margin is set to 0.5.

4 Experiment and Results

4.1 Dataset and Experiment Setup

Dataset. We use the dataset of Flickr8K [14] which consists of 8000 images, each with 5 sentences as its descriptions. Two exemplar image samples together with its sentences are shown in Fig. 5. In our experiment, we split the data into 6000 images for training, 1000 for validating, and 1000 for testing. Since there are 5 labeled description for each image, we finally obtained 30,000 training sentences and 5000 testing sentences.

1. A child in a pink dress is climbing up a set of stairs in an entry way
2. A girl going into a wooden building
3. A little girl climbing into a wooden playhouse
4. A little girl climbing the stairs to her playhouse
5. A little girl in a pink dress going into a wooden cabin

1. A black dog and a spotted dog are fighting
2. A black dog and a tri-colored dog playing with each other on the road
3. A black dog and a white dog with brown spots are staring at each other in the street
4. Two dogs of different breeds looking at each other on the road
5. Two dogs on pavement moving toward each other

Fig. 5. Two examples in the dataset of Flickr8K.

Baselines. In the comparison to other methods, several state-of-the-art methods are used as baselines including (in italics): In 2013, *Hodosh et al.* [14] introduced the dataset of Flickr8K and propose a method of bidirectional ranking on the dataset. Later, Google achieved the state-of-the-art performance on the 1000-class ImageNet using a deep visual-semantic embedding model *DeViSE* [20]. Although they focused on the potential image labels and zero-shot predictions, their model laid the foundation for the latter models. Socher et al. [12] embedded a full-frame neural network with the sentence representation from a semantic dependency tree recursive neural network (*SDT-RNN*), which has made prominent progress in the indices such as mean rank and recall at position k ($R@k$) compared to *kCCA*. Recently, deep fragment embedding proposed by *Karpathy et al.* [21] achieved a major breakthrough in the available datasets.

Evaluation Metrics. We use the popular indices recall at position k ($R@k$) and median rank scores as evaluation metrics. R@k is the percentage of ground truth among the first k returned results and is a widely used index of performance especially for search

engines and ranking systems. The median rank indicates the location of k at which the result has a recall of 50%.

Implementation Settings. In the textual model, we directly use the results of tri-letters dictionary released by the open source demo "sent2vec"[1], which includes about 50,000 tri-letters. If a new tri-letters vector occurs which is not included in the dictionary, it is then appended into the dictionary. Using this dictionary, the image captions are mapped into tri-letter vectors after punctuations are removed.

We set the number of pairs in a batch as 10, and use the 10 corresponding pairs to get 90 irrelevant pairs. Before each epoch, we shuffled the dataset in order to force the network to adapt to more irrelevant image-sentence pairs. We set the dimension of the common feature space as 20. Once the training is completed, the network model is evaluated on testing set of images and sentences. The evaluation process scores and sorts the image-sentence pair in the testing dataset. In the meantime, the locations of ground truth results are recorded.

4.2 Feature Extracted by Textual Model

In this section, the feature extracted by textual model is analyzed. Recall that all the sentences in the test dataset have been mapped into the resulted multi-modal space. From this result, we can determine which words or phrases are extracted into the final space by our network model. Typical resulted samples are shown in Fig. 6. From this figure, we can find that the global feature repeat the keywords in order to keep the features, which satisfies our needs and demands. In Fig. 6, the first underlined words (blue) are the main source of the extracted features, followed by the second underlined words (green), then the third lines (red). There are still other words existing in the final global feature, which only take up a low proportion.

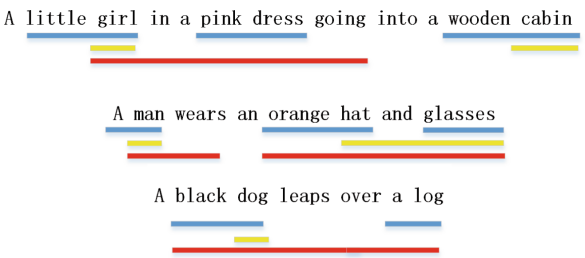

Fig. 6. Features extracted by the textual model represent keywords and key information in the sentence. (Color figure online)

[1] http://research.microsoft.com/en-us/downloads/731572aa-98e4-4c50-b99d-ae3f0c9562b9/default.aspx.

Table 1 shows the results of the average rank of the ground truth in the list. We can find that bag of words (BoW) method can achieve a good performance in the calculation of mapped sentence similarity. In Table 1, the RNN method performs the worst which is also been reported in [3]. One possible reason is the representation of RNN is dominated by the last words, which are usually not the most important words in image captions.

Table 1. Comparison of textual processing to baselines. The rank is expected to be lower because sentences describing the same image should be closer in the common feature space.

Model	Random	BoW	Bigrams	Trigrams	RNN	kCCA	CNN+tri-letter
Med *r*	998.3	24.4	22.7	21.9	38.1	20.3	19.7

4.3 Image Annotation and Searching

This experiment evaluates the performance of the proposed model finding the desired textual or visual information that is more related to the content of the given image or sentence. The results in this task are shown in Table 2. In the paper, most results listed are based on the results in [21]. When comparing with Hodosh et al. [14], we only use a subset of N sentences out of total 5N so that the two approaches can be comparable. From Table 2, we can find that our model outperforms the state-of-the-art methods on most of criteria. The main reason might lie in that [21] requires the fragments of images and sentences to be matched exactly to each other, which is a very strict constraint especially when the sentences are only focused on a part of contents in the images. Such cases tend to result in wrong matches in evaluation. Instead, in our model, the extracted textual features can effectively represent the key information in the sentence, which is more likely to match the salient objects and details of the corresponding image. Besides, the attention model used in the visual model repeats the key information in the images while the textual model repeats the key word in the texts. In the meanwhile, both of the networks can neglect the useless details in the input. Therefore, the extracted features of semantic similar pair of cross-media information can correspond more closely in our work.

Table 2. Result comparison on Flickr8K data

Flickr8K								
Model	Image Annotation				Image Search			
	R@1	R@5	R@10	Med *r*	R@1	R@5	R@10	Med *r*
Random Ranking	0.1	0.5	1.0	635	0.1	0.5	1.0	537
DeViSE [20]	4.8	16.5	27.3	28	5.9	20.1	29.6	29
SDT-RNN [12]	4.5	18.0	28.6	32	6.1	18.5	29.0	29
Karpathy et al. [21]	12.6	32.9	44.0	14	**9.7**	**29.6**	42.5	**15**
Our model	**12.8**	**34.9**	**48.7**	**12**	9.5	29.1	**43.7**	**15**
*Hodosh et al. [14]	8.3	21.6	30.3	34	7.6	20.7	30.1	38
*Karpathy et al. [21]	9.3	24.9	37.4	21	**8.8**	27.9	41.3	**17**
Our model	**9.6**	**25.3**	**37.9**	**19**	8.5	**28.4**	**42.8**	**17**

5 Conclusion

In this paper, we introduced a novel two-stream network model to fulfill the task of bidirectional cross-media information retrieval. This model first maps the textual and visual media into a common feature space. In the textual model, tri-letter vector is used to duplicate the key words and key phrases, and neglect the meaningless details. In the visual model, attention mechanism is combined in the visual model to focus on the partial salient objects in the images so that the most information can be remained and least information can be filtered. During this procedure, the cross-media pairs are judged and their relevance relationships are optimized in the proposed multi-modal embedding methods, in order to determine whether sentences or images are relevant. Comprehensive experiments on publically available dataset demonstrates that the proposed model outperforms the baselines including the state-of-the-arts and prevailing methods. The mixed features extracted by our model are also shown to be advantageous in representing the semantics in images and sentences.

References

1. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 119–126. ACM (2003)
2. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep Boltzmann machines. In: Advances in Neural Information Processing Systems, pp. 2222–2230 (2012)
3. Wu, F., Lu, X., Zhang, Z., et al.: Cross-media semantic representation via bi-directional learning to rank. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 877–886. ACM (2013)
4. Vinyals, O., Toshev, A., Bengio, S., et al.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)
5. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: a neural-based approach to answering questions about images. In: IEEE International Conference on Computer Vision, pp. 1–9. IEEE (2015)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
7. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
8. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
9. Xu, Z., Yang, Y., Hauptmann, A.G.: A discriminative CNN video representation for event detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1798–1807 (2015)
10. Paulin, M., Douze, M., Harchaoui, Z., et al.: Local convolutional features with unsupervised training for image retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 91–99 (2015)

11. Matsuo, S., Yanai, K.: CNN-based style vector for style image retrieval. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pp. 309–312. ACM (2016)
12. Socher, R., Karpathy, A., Le, Q.V., et al.: Grounded compositional semantics for finding and describing images with sentences. Trans. Assoc. Comput. Linguist. **2**, 207–218 (2014)
13. Zhuang, Y., Yu, Z., Wang, W., et al.: Cross-media hashing with neural networks. In: Proceedings of the ACM International Conference on Multimedia, pp. 901–904. ACM (2014)
14. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: data, models and evaluation metrics. J. Artif. Intell. Res. **47**, 853–899 (2013)
15. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. Neural Comput. **16**(12), 2639–2664 (2004)
16. Ballan, L., Uricchio, T., Seidenari, L., et al.: A cross-media model for automatic image annotation. In: Proceedings of International Conference on Multimedia Retrieval, p. 73. ACM (2014)
17. Wang, Y., Wu, F., Song, J., et al.: Multi-modal mutual topic reinforce modeling for cross-media retrieval. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 307–316. ACM (2014)
18. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 127–134. ACM (2003)
19. Pereira, J.C., Coviello, E., Doyle, G., et al.: On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **36**(3), 521–535 (2014)
20. Frome, A., Corrado, G.S., Shlens, J., et al.: Devise: a deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems, pp. 2121–2129 (2013)
21. Karpathy, A., Joulin, A., Li, F.F.F.: Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in Neural Information Processing Systems, pp. 1889–1897 (2014)
22. Gao, J., Deng, L., Gamon, M., et al.: Modeling interestingness with deep neural networks: U. S. Patent 20,150,363,688, 17 December 2015
23. Huang, P.S., He, X., Gao, J., et al.: Learning deep structured semantic models for web search using clickthrough data. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 2333–2338. ACM (2013)
24. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems, (NIPS 2015), pp. 2017–2025 (2015)