Deep Coordinated Textual and Visual Network for Sentiment-Oriented Cross-Modal Retrieval

Jiamei Fu[1,2], Dongyu She[2], Xingxu Yao[2], Yuxiang Zhang[1], and Jufeng Yang[2(✉)]

[1] College of Computer Science and Technology, Civil Aviation University of China, Tianjin, China
[2] College of Computer and Control Engineering, Nankai University, Tianjin, China
yangjufeng@nankai.edu.cn

Abstract. Cross-modal retrieval has attracted more and more attention recently, which enables people to retrieve desired information efficiently from a large amount of multimedia data. Most methods on cross-modal retrieval only focus on aligning the objects in image and text, while sentiment alignment is also essential for facilitating various applications, *e.g.*, entertainment, advertisement, *etc.* This paper studies the problem of retrieving visual sentiment concepts with a goal to extract sentiment-oriented information from social multimedia content, *i.e.*, sentiment oriented cross-media retrieval. Such problem is inherently challenging due to the subjective and ambiguity characteristics of the adjectives like "sad" and "awesome". Thus, we focus on modeling visual sentiment concepts with adjective-noun pairs, *e.g.*, "sad dog" and "awesome flower", where associating adjectives with concrete objects makes the concepts more tractable. This paper proposes a deep coordinated textural and visual network with two branches to learn a joint semantic embedding space for both images and texts. The visual branch is based on a convolutional neural network (CNN) pre-trained on a large dataset, which is optimized with the classification loss. The textual branch is added on the fully-connected layer providing supervision of the textual semantic space. In order to learn the coordinated representation for different modalities, the multi-task loss function is optimized during the end-to-end training process. We have conducted extensive experiments on a subset of the large-scale VSO dataset. The results show that the proposed model is able to retrieval sentiment-oriented data, which performs favorably against the state-of-the-art methods.

Keywords: Cross-modal retrieval · Visual sentiment analysis
Convolutional neural network

J. Fu and D. She—The two authors contributed equally to this paper.

X. Geng and B.-H. Kang (Eds.): PRICAI 2018, LNAI 11012, pp. 684–696, 2018.
https://doi.org/10.1007/978-3-319-97304-3_52

Fig. 1. Some examples from the VSO dataset [7]. The traditional cross-modal retrieval task focuses on the object-level alignment (a), while this work focuses the sentiment-oriented retrieval task (b).

1 Introduction

With the advance of social media, more and more people tend to share experiences and opinions on the social networks, *e.g.*, Flickr, Twitter and Instagram, *etc.* Generally, user-generated contents include multimedia data, *e.g.*, texts, images and videos, which have experienced tremendous growth in recent years. Cross-modal retrieval therefore emerges as the research topic and has attracted much attention [1–3], which aims to retrieve relevant data of another modality given one modality of data as the query. The existing methods mainly focus on aligning the objects in image and text, while computational sentiment understanding of such media data is of great importance for applications, *e.g.*, affective computing [4], opinion mining [5], entertainment [6], *etc.* This paper studies the problem of retrieving visual sentiment concepts with a goal to extract sentiment-oriented information from social multimedia content, referred as *sentiment oriented cross-modal retrieval*.

Numerous methods mainly focus on learning separate representation and projecting different modalities to a common space for similarity measure, which can be divided to two types: traditional and deep learning based methods. The first is to learn projections in the traditional frameworks, *e.g.*, canonical correlation analysis (CCA) [1], which cannot capture the complex cross-modal correlation. Recently, deep learning has achieved great progress in single-modal scenario, *e.g.*, image classification [8], object detection [9]. There are several methods [10,11] that utilize the ability of deep neural network to model complex cross-modal correlation of the object. While modeling the concrete visual concepts (*i.e.*, nouns) has been widely studied, there is little work that model the affective visual concepts. Such problem is inherently challenging due to the subjective and ambiguity characteristics of the adjectives like "sad" and "awesome". Thus, we focus on modeling visual sentiment concepts with adjective-noun pairs (ANP), *e.g.*, "sad

dog" and "awesome flower", where associating adjectives with concrete objects makes the concepts more tractable. Thus, different from traditional object-level retrieval task, *sentiment-oriented cross-modal retrieval task* aims to retrieval images when given ANP queries with sentiment alignment and vice versa as shown in Fig. 1.

In order to address this problem, this paper proposes deep coordinated textual and visual network to learn a sentiment semantic embedding space for both images and texts. In specific, the visual branch is based on a convolutional neural network (CNN) pre-trained on a large dataset, *i.e.*, ImageNet [12], which is optimized with the classification loss. The textual branch is added on the fully-connected layer providing supervision of the textual semantic space by employing the semantic regularization constraint. In order to learn the coordinated representation for different modalities, the multi-task loss function is optimized during the end-to-end training process. Extensive experiments have been conducted on a subset of the large-scale VSO dataset [7]. The results show that the proposed model is able to retrieval sentiment-oriented data, which performs favorable against the state-of-the-art methods.

2 Related Work

In this section, we focus on reviewing the related approaches for cross-modal retrieval ranging from traditional methods [3,13,14] to deep learning based methods [15,34].

2.1 Traditional Methods

Canonical correlation analysis is widely used for model multimodal data and extended to variable types [3,13,14], which aims to learn a common space along with two modalities of data. For example, Rasiwasia *et al.* [1] propose a two-stage method to combine category information with CCA and learn a semantic space to measure the similarity of different features between two modalities. Li *et al.* [16] propose cross-modal factor analysis (CFA) to learn the pairwise cross-modal correlation, which attempts to minimize the Frobenius norm between pairwise data. CCA is unsupervised that does not need the category labels, while some researchers try to apply semi-supervised learning and graph regularization into cross-modal common representation learning. For example, Zhai *et al.* [17] propose joint representation learning (JRL) to utilize graph to model each modality of data to a common space and further learn semantic representation in an semi-supervised model. Joint graph regularized heterogeneous metric Learning (JGRHML) is proposed to learn the project matrices by adopting metric learning and graph regularization [19]. Wang *et al.* [18] utilize multimodal graph regularization on the projected data with an iterative algorithm to preserve inter-modality and intra-modality relationships. In addition, multi-view CCA [13] is proposed to project visual and textual features to the common space and combine a third view tolearn high-level semantic and multi-label CCA [20] is developed

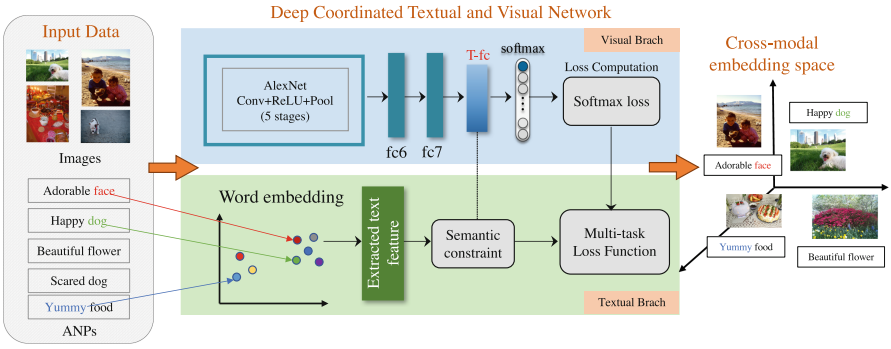

Fig. 2. Illustration of the proposed deep coordinated textual and visual network. Given the images and the corresponding ANP, we employ two branches and represent the different modalities in the joint feature embedding space. The multi-task loss is optimized for the cross-modal representation learning.

to address the problem of cross-modal retrieval when data has multiple labels and incorporate such information to learn the subspace.

2.2 Deep Learning Based Methods

Deep learning has shown the strong ability to extract the representation of texts and images, which has achieved state-of-the-art performance in various related tasks, *e.g.*, object detection [16,32,33] and visual sentiment prediction [29,30] and image retrieval [31]. Srivastava *et al.* [34] propose multimodal deep belief network (DBN) to utilize two separate DBNs to model image and text and then employ a joint RBM to learn common representation. Ngiam *et al.* [15] attempt to extend RBM to model data of multiple modalities based on deep auto-encoder network, named bimodal autoencoders. Correspondence autoencoder (Corr-AE) [21] is proposed to learn separate representation and then jointly learn the correlation with two subnetworks by minimizing error of combination of representation learning between different modalities. Peng *et al.* [2,22] propose cross-media multiple deep networks (CMDN) to jointly learn the intra-modality and inter-modality correlation with two separate networks and then apply hierarchical learning to model correlation between different modalities. Liu *et al.* [27] propose a framework to map features of cross-media data to an common semantic space and model the similarity of data. Some researchers also try to combine DNN with CCA, namely deep canonical correlation analysis (DCCA) [10,23], which use two separate subnetworks to learn non-linear representation and design the correlation constraints to maximize the total correlation.

3 Methodology

As illustrated in Fig. 2, we develop a deep framework including two branches, *i.e.*, visual branch and textual branch, which can be used to learn non-linear and complex representation for images and texts. In addition, we employ the classification loss and semantic constraint as the multi-task optimization function, which learns the cross-modal representation during the end-to-end training. The detailed construction is illustrated in the following section.

3.1 Visual Branch

Different from the traditional object recognition task that images of the same categorization often share highly-similar appearance patterns, visual sentiment analysis are usually involved with great intra-class variance. With the sentiment supervision, it is challenging to learn a mapping function from the low-level image features to high-level semantic space due to the "affective gap" [24]. Recently, CNN has the ability to extract high-level semantic representation from images, which has achieved state-of-the-art performance in the visual recognition tasks. To utilize the high-level features to guide the sentiment representation learning, we employ the AlexNet [12] as the basic model, which is composed of five convolutional layers and three fully-connected layers. The adopted AlexNet architecture consists of more than 60 million parameters, which is hard to optimize from scratch with the limited amount of affective data. Considering the good results achieved by previous work with transfer learning [25], we initialize the weights in each layer but the last one with the parameters from the pre-trained model on the ImageNet dataset [26].

As a supervised learning approach, the fine-tuned model is adopted to learn a function $f : \chi \rightarrow y$, from a collection of affective training examples $\{(x_n, y_n, s_n)\}_{n=1}^{N}$, where N is the size of the training set, x_n is the affective image, y_n and s_n is the associated ANP label and text. During each training epoch, each image is resized into 256×256 pixels and CNN crops $227 \times 227 \times 3$ patches from each image as input. We use the output from the penultimate layer as the feature $\mathbf{d}_n$ of each patch, the fine-tuning of the last layer is done by maximizing the following log likelihood function:

$$L_{cls} = \sum_{n=1}^{N} \sum_{c \epsilon y} \prod (y_n = c) \log p(y_n = c | \mathbf{d}_n, \mathbf{w}_c) \tag{1}$$

where $\mathbf{W} = \{\mathbf{w}_c\}_{c \epsilon y}$ is the set of model parameters, and $\prod(x) = 1$ if x is true. The sentiment probability $p(y_n = c|\mathbf{d}_n, \mathbf{w}_c)$ can be defined by

$$p(y_n = c|\mathbf{d}_n, \mathbf{w}_c) = \frac{exp(\mathbf{w}_c^T \mathbf{d}_n)}{\sum_{c' \epsilon y} exp(\mathbf{w}_{c'}^T \mathbf{d}_n)} \tag{2}$$

Since the category number of affective dataset is not equal to that of ImageNet, the c-way fc8 classification layer is changed to the class number required by the sentiment dataset, which can produce a probability distribution over the emotional class labels.

3.2 Textual Branch

In order to learn the joint embedding space for image and text modalities, we propose to add the textual branch in the deep model. Given the ANP consisting of adjective and noun word, we first employ the word2vec model to generate textual embedding, which is trained on a large news dataset [2]. The input is denoted as the word sequence $s = <s^{(1)}, s^{(2)}>$, where each word can be represented by a k-dimensional real-valued vector using the word2vec model. For words that do not found in the dictionary of word2vec, we initialize the word vectors randomly. In this work, we use the fixed-length word2vec word embeddings and set $k = 300$ empirically. Thus, the textual embedding of the input ANP can be denoted by

$$\mathbf{v} = [\text{word2vec}(s^{(1)}), \text{word2vec}(s^{(2)})], \tag{3}$$

where $[\cdot]$ denotes the concatenation operation and $\mathbf{v} \in \mathbb{R}^{2k}$.

To facilitate the textual learning process, we insert a fully connected layer T_fc into the middle between fc7 layer and fc8 layer. Similar to the fully-connected layer in CNN, ReLU is utilized as the nonlinear activation function for such fully-connected layer. We add a semantic regularization constraints into the network providing supervision to the semantic space, which is denoted by:

$$L_{sem} = \sum_{n=1}^{N} (||f_{W_T}^{(T)} x_n - \mathbf{v_n}||^2), \tag{4}$$

where $f_{W_T}^{(T)}$ is the output from the T_fc layer of the proposed network.

3.3 Coordinated Feature Learning

To address the problem of cross-modal embedding space learning, we consider the multi-task loss to optimize the correspondence between semantic representation of image and text with the corresponding label vector. Thus, our loss function to be minimized is integrated with two types of losses, denoted as:

$$L = L_{cls}(x, y) + L_{sem}(x, s), \tag{5}$$

where L_{cls} and L_{sem} denote the classification loss and semantic regularization constraint, respectively. The proposed network can thus be optimized with stochastic gradient descent (SGD) in an end-to-end manner. By this way, the semantic representation of pairs of image and text are well consistent, which benefits the alignment between the sentiment in the different modalities.

Based on the trained model, the cross-modality retrieval of a given image can be summarized as follows. For each test image, we first generate the $2k$-dimensional T_fc features as the textual embedding of the sample. To search for an ANP which has a similar emotion as a given query image, we determine the nearest neighbors in terms of the feature representation according to the Euclidean distance, and vice versa.

Table 1. The ANPs with six most popular nouns from the VSO dataset. We use these ANPs to collect images to form our dataset for sentiment-related retrieval task.

Adjective	Noun
Bad, broken, clean, crazy, damaged, dirty, expensive, fancy, hot, lonely, safe, tiny, ugly	Car
Adorable, aggressive, cute, dirty, faithful, fluffy, friendly, funny, happy, lonely, muddy, playful, sad, scared, shy, silly, sleepy, smiling, tiny, tired, wet	Dog
Adorable, cute, fancy, gorgeous, pretty, sexy, traditional	Dress
Adorable, angry, attractive, chubby, clean, crazy, crying, cute, dirty, dumb, excited, funny, grumpy, handsome, hilarious, innocent, mad, pretty, scared, silly, sleepy, sweet, tired, ugly, weird, worried	Face
Beautiful, dry, dying, favorite, fragile, golden, little, prickly, smelly, strange, stunning, sunny, tiny, wild	Flower
Colorful, delicious, dry, excellent, fancy, favorite, greasy, great, hot, natural, super, tasty, yummy	Food

4 Experiment

4.1 Dataset

For the sentiment-oriented cross-modal retrieval, we collect a subset of the large-scale VSO dataset that contains millions of images by querying Flickr with 1553 adjective and noun pairs (ANPs) following [28]. We only select the ANPs related to six most popular nouns, *i.e.*, car, dog, face, dress, flower and food, which are frequently occurred and more detectable in the social media. Such nouns are also associated with different adjectives to form 94 ANPs that show various sentiments, as shown in Table 1. If our method can effectively address the retrieval task on this subset, we can easily extend the work to cover more multimedia data. A total of 36,175 pairs of images and texts are finally collected as our dataset in this paper. Some example ANPs are shown in Fig. 3.

4.2 Experiment Settings

For the text modality, we use a publicly available word embeddings [2] to represent these ANPs, which has been pre-trained on large amount of words from Google News using the continuous bag-of-words model. We randomly split the dataset into 80% and 20% for training and testing set, respectively. The learning rates of the convolutional layers and the last fully-connected layer are initialized as 0.001 and 0.001, respectively. We employ stochastic gradient descent (SGD) to train the deep network using batches of 256, which ensures one batch to cover the whole 94 categories. We implement our framework based on AlexNet [12]. All our experiments are carried out on one NVIDIA GTX 1080 GPU with 32 GB CPU memory.

Fig. 3. Some examples of image-ANP pairs from the collected dataset.

4.3 Baseline

We evaluate our method against the other methods, including *correlation matching* (CM), *semantic matching* (SM), *semnatic correlation matching* (SCM) [1]. CM depends on subspace learning to model correlation between image and text subspaces. SM first abstracts the multi-modal data to a high-level representation, which aims to learn the correlation between the text and image by subspace learning. SCM combines subspace and semantic modeling and utilize logistic regression to maximize correlated subspaces. We also compare the performance of different combination of text feature and visual feature for these methods on this dataset. For visual features, we attempt to compare the hand-craft feature, such as GIST feature and MPEG-7 feature, and deep feature. In this paper, we utilize AlexNet to extract deep visual feature. And we use the public available code with default parameters to extract the GIST visual feature, which results in a 512-dimensional feature vector.

4.4 Evaluation Metrics

To evaluate the effectiveness of our proposed approach, two cross-modal retrieval tasks are conducted: Image as the query to retrieval text and Text as the query to retrieval images. And we also report the average results in two tasks. Following [19], we report the average recall of the gold item at position 10 of the ranked list (R@10), and precision at position 10 of the ranked list (P@10), for cross-modal retrieval tasks. In addition, we compute the F-score by the formulation:

$$F\text{-}score = \frac{2 \times PR}{P + R} \tag{6}$$

Table 2. Cross-modal retrieval performance on the subset of VSO dataset. We evaluate several baselines for cross-modal retrieval, including SM, CM, SCM with different features. Note the GIST, MPEG_7 represents the hand-craft visual feature, and deep feature represent feature extracted from the CNN pre-trained on ImageNet.

Method	Image → Text				Text → Image				Average			
	P@10	R@10	F-score	MAP	P@10	R@10	F-score	MAP	P@10	R@10	F-score	MAP
CM-GIST	0.0428	0.0052	0.0093	0.1308	0.0840	0.0070	0.0129	0.0569	0.0634	0.0061	0.0111	0.0939
CM-MPEG_7	0.0524	0.0054	0.0098	0.1476	0.0989	0.0076	0.0141	0.0689	0.0757	0.0065	0.0120	0.1082
CM-Deep	0.0783	0.0086	0.0155	0.2010	0.1629	0.0106	0.0199	0.1066	0.1206	0.0096	0.0178	0.1538
SM-GIST	0.1049	0.0057	0.0110	0.1673	0.1494	0.0110	0.0204	0.0615	0.1272	0.0084	0.0157	0.1144
SM-MPEG_7	0.1329	0.0065	0.0124	0.1952	0.1525	0.0138	0.0253	0.0769	0.1427	0.0101	0.0189	0.1361
SM-Deep	0.2062	**0.0170**	**0.0314**	0.2811	0.2692	0.0157	0.0297	0.1370	0.2377	**0.0164**	**0.0306**	0.2091
SCM-GIST	0.1024	0.0052	0.0099	0.1646	0.1283	0.0087	0.0163	0.0623	0.1154	0.0139	0.0248	0.1134
SCM-MPEG_7	0.1216	0.0061	0.0116	0.1878	0.1229	0.0090	0.0168	0.0775	0.1223	0.0076	0.0143	0.1327
SCM-Deep	0.1767	0.0089	0.0169	0.2562	0.2036	0.0130	0.0244	0.1285	0.1902	0.0110	0.0208	0.1924
Ours	**0.2250**	0.0118	0.0224	**0.3064**	**0.2781**	**0.0177**	**0.0333**	**0.1897**	**0.2516**	0.0144	0.0272	**0.2481**

The mean average precision (MAP) is used to evaluate the overall performance of approaches for cross-media retrieval task. To compute MAP, we first evaluate the average precision (AP) of a set of R retrieved sets by

$$AP = \frac{1}{R}\sum_{k=1}^{N}\frac{R_k}{k} \times rel_k, \tag{7}$$

where R is the number of relevant items in the retrieved set N, R_k denotes the precision of the top k retrieved results, and $rel_k = 1$ when the k-th retrieved item is relevant and otherwise 0. The MAP is the average value of the AP values over all queries in the query set. The larger MAP indicate better performance.

4.5 Results Analysis

In order to illustrate the effectiveness of our method, we report comparative results with other methods on the dataset in Table 2. As is shown, when we utilize the same method, such as SCM, deep feature improve mAP by 9% compared with hand-craft features, *i.e.*, MPEG_7 feature. When using SM, deep feature outperforms the GIST feature by about 12% in image retrieval text task on mAP. We can conclude that deep visual feature combined with the traditional methods improve the retrieval performance against the hand-craft features, *i.e.*, GIST feature and MPEG_7 feature. It also can be seen that SM outperform CM and SCM, for example, SM improve mAP about 8% compared with CM and 3% against SCM when using deep visual feature and text word2vec feature. Our framework performs favorably against all the baseline methods on all evaluate metrics. For example, our method improves mAP about 2% compared with SM-Deep, 10% compared with CM-Deep in image retrieval text task. Because our framework can optimize both subnetworks simultaneously, *i.e.*, textual network and visual network, to further learn two representation in an end-to-end manner. The two networks can be boosted in the training process of the framework.

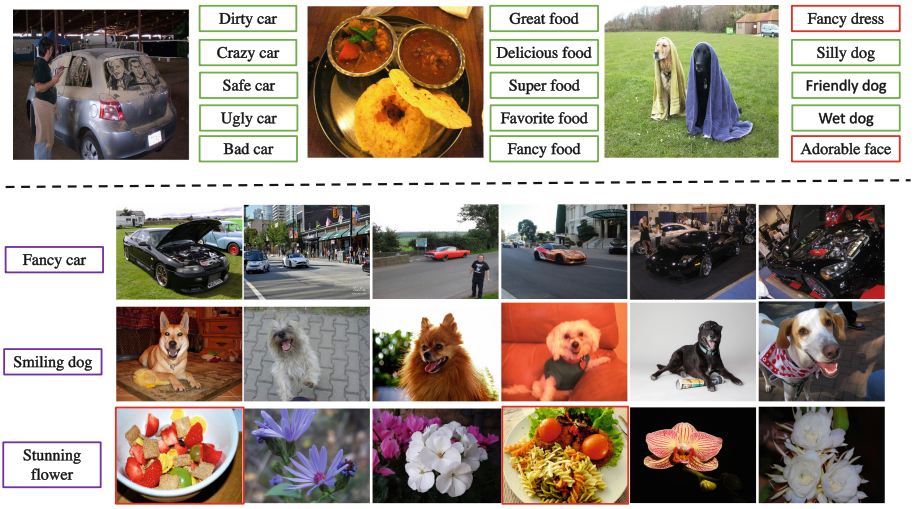

Fig. 4. Several correct and failure retrieved results of our method on the dataset. In the top line, *i.e.*, cases of image query text, the correct related and false ANP is marked with green and red. In the cases of text query images, the first two line show our top-6 retrieval results. While some image retrieved by 'stunning flower' is false, with red border. (Color figure online)

Thus, we draw the following conclusions: First, the deep visual feature can represent that image feature more effectively compared with the traditional feature. Second, the fine-tuned feature on the affective image dataset can represent the visual modality related to sentiment to further improve performance.

4.6 Visualization

To qualitatively evaluate our proposed approach, we visualize some example results, *i.e.*, image retrieved by text query and text retrieved by image query. We show our cross-modal retrieval example results in Fig. 4. In the cases of image query texts, the retrieved ANP is related to object and sentiment in image. For example, for 'dirty car' queries, the first retrieved image is completely corresponding to the image. While other ANP is related to objects 'car'. The right is failure result, which demonstrates that the color of dog is similar to the cloth and confuse the model to return the false results. While the correct result occurs in the fourth position. As is shown, in the cases of text query images, *i.e.*, the first line, our approach retrieval the correct images in top-6. And the false retrieved image with red border. Although the top-1 retrieved example does not have the same ANP label, such image has the related sentiment, *e.g.*, stunning flower. But our approach still achieves good performance on cross-modal retrieval task and can effectively retrieval sentiment-related and object-alignment results.

5 Conclusion

In this work, we introduce a novel problem in the field of cross-modal retrieval, *i.e.*, sentiment oriented cross-media retrieval. Different from traditional task, we focus on the problem of retrieving visual sentiment concepts in affective image and retrieving image according to the emotional text. To address this problem, we propose a deep coordinated neural network combines two branches, i.e. visual and textual branches, which is optimized with a multi-task loss function for the cross-modal representation learning. The experiments on the subset of VSO dataset demonstrate that our method can efficiently establish the correspondence between different modalities, i.e., image and text, and further improve the retrieval performance.

Acknowledgments. This work was partially supported by grants from the NSFC (No. U1533104), the Natural Science Foundation of Tianjin, China (No. 18JCYBJC15400), and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

References

1. Rasiwasia, N., et al.: A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 251–260. ACM, New York (2010)
2. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
3. Pereira, J.C., 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)
4. Joshi, D., et al.: Aesthetics and emotions in images. IEEE Sig. Process. Mag. **28**(5), 94–115 (2011)
5. Truong, Q.T., Lauw, H.W.: Visual sentiment analysis for review images with item-oriented and user-oriented CNN. In: Proceedings of the 2017 ACM on Multimedia Conference, pp. 1274–1282. ACM (2017)
6. Jia, J., Wu, S., Wang, X., Hu, P., Cai, L., Tang, J.: Can we understand van Gogh's mood?: learning to infer affects from images in social networks. In: Proceedings of the 20th ACM International Conference on Multimedia, pp. 857–860. ACM (2012)
7. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 806–813 (2014)
8. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. (2014)
9. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
10. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International Conference on Machine Learning, pp. 1247–1255 (2013)
11. Wang, W., Yang, X., Ooi, B.C., Zhang, D., Zhuang, Y.: Effective deep learning-based multi-modal retrieval. VLDB J. **25**(1), 79–101 (2016)

12. 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)
13. Gong, Y., Ke, Q., Isard, M., Lazebnik, S.: A multi-view embedding space for modeling internet images, tags, and their semantics. Int. J. Comput. Vis. **106**(2), 210–233 (2014)
14. Klein, B., Lev, G., Sadeh, G., Wolf, L.: Associating neural word embeddings with deep image representations using fisher vectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4437–4446 (2015)
15. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: 28th International Conference on Machine Learning, pp. 689–696 (2011)
16. Li, D., Dimitrova, N., Li, M., Sethi, I.K.: Multimedia content processing through cross-modal association. In: Proceedings of the Eleventh ACM International Conference on Multimedia, pp. 604–611. ACM (2003)
17. Zhai, X., Peng, Y., Xiao, J.: Learning cross-media joint representation with sparse and semisupervised regularization. IEEE Trans. Circuits Syst. Video Technol. **24**(6), 965–978 (2014)
18. Wang, K., He, R., Wang, L., Wang, W., Tan, T.: Joint feature selection and subspace learning for cross-modal retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **38**(10), 2010–2023 (2016)
19. Zhai, X., Peng, Y., Xiao, J.: Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In: Association for the Advancement of Artificial Intelligence (2013)
20. Ranjan, V., Rasiwasia, N., Jawahar, C.V.: Multi-label cross-modal retrieval. In: IEEE International Conference on Computer Vision, pp. 4094–4102 (2015)
21. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: 22nd ACM International Conference on Multimedia, pp. 7–16 (2014)
22. Peng, Y., Huang, X., Qi, J.: Cross-media shared representation by hierarchical learning with multiple deep networks. In: 25th International Joint Conference on Artificial Intelligence, pp. 3846–3853 (2016)
23. Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3441–3450 (2015)
24. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: 18th ACM International Conference on Multimedia, pp. 83–92. ACM (2010)
25. You, Q., Luo, J., Jin, H., Yang, J.: Building a large scale dataset for image emotion recognition: the fine print and the benchmark. In: Association for the Advancement of Artificial Intelligence, pp. 308–314 (2016)
26. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
27. Liu, T., Zhao, Y., Wei, S., Wei, Y., Liao, L.: Enhanced isomorphic semantic representation for cross-media retrieval. In: IEEE International Conference on Multimedia and Expo, pp. 967–972 (2017)
28. Chen, T., Yu, F.X., Chen, J., Cui, Y., Chen, Y.Y., Chang, S.F.: Object-based visual sentiment concept analysis and application. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 367–376. ACM (2014)
29. Yang, J., Sun, M., Sun, X.: Learning visual sentiment distributions via augmented conditional probability neural network. In: The Association for the Advancement of Artificial Intelligence, pp. 224–230 (2017)

30. Yang, J., She, D., Sun, M.: Joint image emotion classification and distribution learning via deep convolutional neural network. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 3266–3272 (2017)
31. Yang, J., She, D., Lai, Y., Yang, M.H.: Retrieving and classifying affective images via deep metric learning. In: The Association for the Advancement of Artificial Intelligence (2018)
32. Yang, J., She, D., Sun, M., Cheng, M.M., Rosin, P., Wang, L.: Visual sentiment prediction based on automatic discovery of affective regions. IEEE Trans. Multimed. **PP**(99), 1 (2018)
33. Yang, J., She, D., Lai, Y.K., Rosin, P.L., Yang, M.H.: Weakly supervised coupled networks for visual sentiment analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
34. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. In: Advances in Neural Information Processing Systems, pp. 2222–2230 (2012)