Cross-Modal Event Retrieval: A Dataset and a Baseline Using Deep Semantic Learning

Runwei Situ[1], Zhenguo Yang[1](✉), Jianming Lv[2], Qing Li[3],
and Wenyin Liu[1](✉)

[1] School of Computer Science and Technology,
Guangdong University of Technology, Guangzhou, China
siturunwei@163.com, zhengyang5-c@my.cityu.edu.hk,
liuwy@gdut.edu.cn

[2] School of Computer Science and Engineering,
South China University of Technology, Guangzhou, China
jmlv@scut.edu.cn

[3] Department of Computer Science, City University of Hong Kong,
Hong Kong, China
itqli@cityu.edu.hk

Abstract. In this paper, we propose to learn Deep Semantic Space (DSS) for cross-modal event retrieval, which is achieved by exploiting deep learning models to extract semantic features from images and textual articles jointly. More specifically, a VGG network is used to transfer deep semantic knowledge from a large-scale image dataset to the target image dataset. Simultaneously, a fully-connected network is designed to model semantic representation from textual features (e.g., TF-IDF, LDA). Furthermore, the obtained deep semantic representations for image and text can be mapped into a high-level semantic space, in which the distance between data samples can be measured straightforwardly for cross-model event retrieval. In particular, we collect a dataset called Wiki-Flickr event dataset for cross-modal event retrieval, where the data are weakly aligned unlike image-text pairs in the existing cross-modal retrieval datasets. Extensive experiments conducted on both the Pascal Sentence dataset and our Wiki-Flickr event dataset show that our DSS outperforms the state-of-the-art approaches.

Keywords: Cross-modal event retrieval · Deep learning · Common space

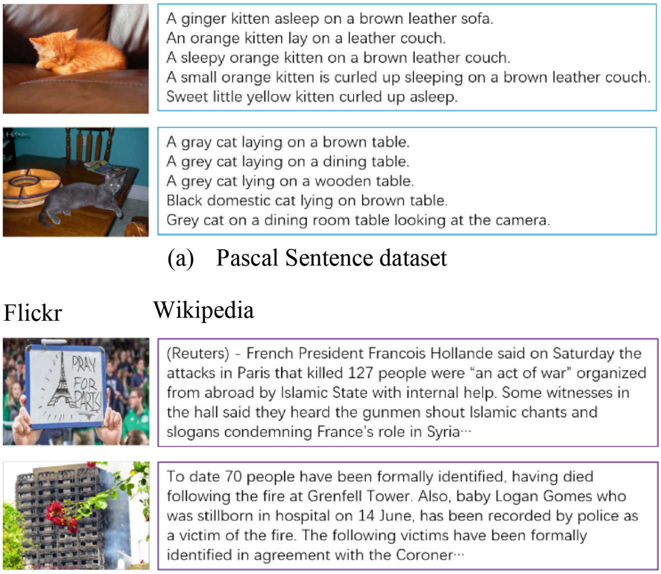
1 Introduction

The development of the Internet and the emerging social media are changing the way people interact with each other. Meantime, multi-modal online data (e.g. images, texts, audios and videos) is growing rapidly. In reality, data in different modalities can be used for describing the same real-world events [1, 2] (e.g., protests, elections, festivals, natural disasters). For instance, a news website may contain textual descriptions, images and audios to report an event. Cross-modal retrieval aims to retrieve the data in different modalities that are relevant to the same event. Multi-modal data can span

R. Hong et al. (Eds.): PCM 2018, LNCS 11165, pp. 147–157, 2018.
https://doi.org/10.1007/978-3-030-00767-6_14

different feature spaces, known as the problem of "semantic gap", making the measurement on the content similarity among the data more challenging.

Researchers have built up quite a few multimodal datasets for cross-modal retrieval, such as Wikipedia image-text pairs [3], Pascal sentence [4], Pascal VOC [5]. However, these datasets focus on strongly aligned data pairs, e.g., an image of cat and its exact textual descriptions as shown in Fig. 1a. In reality, there may exist more complicated cases which cannot be expressed by one-to-one data pairs. For instance, given a photo of news event (e.g., a protest), users may expect to acquire the relevant textual materials, which are usually not the exact description of the photo but they share the same label of event. Therefore, we call such data weakly aligned data as shown in Fig. 1b.



A ginger kitten asleep on a brown leather sofa.
An orange kitten lay on a leather couch.
A sleepy orange kitten on a brown leather couch.
A small orange kitten is curled up sleeping on a brown leather couch.
Sweet little yellow kitten curled up asleep.

A gray cat laying on a brown table.
A grey cat laying on a dining table.
A grey cat lying on a wooden table.
Black domestic cat lying on brown table.
Grey cat on a dining room table looking at the camera.

(a) Pascal Sentence dataset

(Reuters) - French President Francois Hollande said on Saturday the attacks in Paris that killed 127 people were "an act of war" organized from abroad by Islamic State with internal help. Some witnesses in the hall said they heard the gunmen shout Islamic chants and slogans condemning France's role in Syria···

To date 70 people have been formally identified, having died following the fire at Grenfell Tower. Also, baby Logan Gomes who was stillborn in hospital on 14 June, has been recorded by police as a victim of the fire. The following victims have been formally identified in agreement with the Coroner···

(b) Wiki-Flickr event dataset

Fig. 1. Examples of strongly aligned image-text pairs from Pascal sentence dataset, and weakly aligned examples from our Wiki-Flickr event dataset. In particular, the corresponding text is the exact description of an image in the strongly aligned data pairs. In contrast, the weakly aligned textual content does not describe an image exactly, but they share the same event label.

In the context of cross-modal event retrieval, using the user-generated content is very challenging to obtain a joint representation for multimodal data. The performance of the traditional techniques, such as CCA [6], CFA [7], is still far from satisfactory. Recently, due to the development of deep learning, significant progress has been made in the fields of speech recognition, image recognition, sentiment classification, and image caption generation. Inspired by these works, we employ deep learning models for cross-modal event retrieval, especially for the weakly aligned data.

In this paper, we utilize deep learning models to learn a common semantic space in order to measure the content similarity between data in different modalities. More specially, we employ a VGG [8] network to transfer semantic knowledge from ImageNet dataset to our Wiki-Flickr event dataset. At the same time, we devise a fully-connected network to extract deep features from raw textual features, e.g., TF-IDF, LDA. Furthermore, we map the images and texts to a common semantic space with high-level semantics, in which the cross-modal data samples can be matched directly by using similarity measurement. The main contributions of this work are the following:

(1) We propose a deep semantic space (DSS) framework based on the VGG network and fully-connected network for cross-modal retrieval. DSS has the advantage of high discriminative power, and hence can be used to deal with weakly aligned data.
(2) We collect a Wiki-Flickr event dataset, where the data are weakly aligned unlike the usual image-text pairs in the existing datasets. We plan to release the dataset for public use later on.
(3) Extensive experiments conducted on the Pascal Sentence dataset and our Wiki-Flickr event dataset show that the proposed DSS outperforms the state-of-the-art approaches.

The rest of this paper is organized as follows. Section 2 reviews the related work on cross-modal retrieval. Section 3 shows the details of our proposed method. Extensive experiments and conclusions are given in Sects. 4 and 5, respectively.

2 Related Work

Various approaches have been proposed to deal with cross-modal retrieval, which can be roughly divided into four categories: subspace learning, hashing-based methods, rank-based methods, and DNN-based methods. We introduce the basic ideas and a few representative approaches in these categories below.

(1) **Subspace learning**. Canonical correlation analysis (CCA) and Kernel-CCA are representative subspace learning approaches, which aim to learn a common subspace shared among different modalities of data by maximizing their correlations.
(2) **Hashing-based methods.** Considering high-dimensional cross-modal data, Bronstein et al. [9] proposed a cross modal similarity sensitive hashing (CMSSH) for efficient cross-modal tasks. Song et al. [10] proposed a novel inter-media hashing (IMH) model to transform multimodal data into a common Hamming space.
(3) **Rank-based methods.** Rank-based methods usually use the strategy of learning to rank. Bai et al. [11] presented supervised semantic indexing (SSI) for cross-lingual retrieval. Grangier et al. [12] proposed a discriminative kernel-based method to solve the problem of cross-modal ranking by adapting the passive-aggressive algorithm.

(4) **DNN-based methods.** The recent DNN-based methods can utilize the advantage of large-scale data, achieving better performance than the traditional approaches. Srivastava et al. [13] proposed to learn a shared representation between different modalities based on restricted Boltzmann machine. Wang et al. [14] proposed a regularized deep neural network (RE-DNN), which is a 5-layer neural network for mapping visual and textual features into a common semantic space. Furthermore, similarity between different modalities can be measured seamlessly. Wei et al. [15] presented a semantic matching method to address the cross-modal retrieval problem. However, shallow networks usually perform well on the small-scale dataset, which may suffer from underfitting when dealing with large-scale datasets.

3 The Proposed Method

This section elaborates our proposed method for cross-modal event retrieval, which uses the VGG network and a fully-connected network to learn the common semantic space. Figure 2 illustrates the overview of our proposed framework.

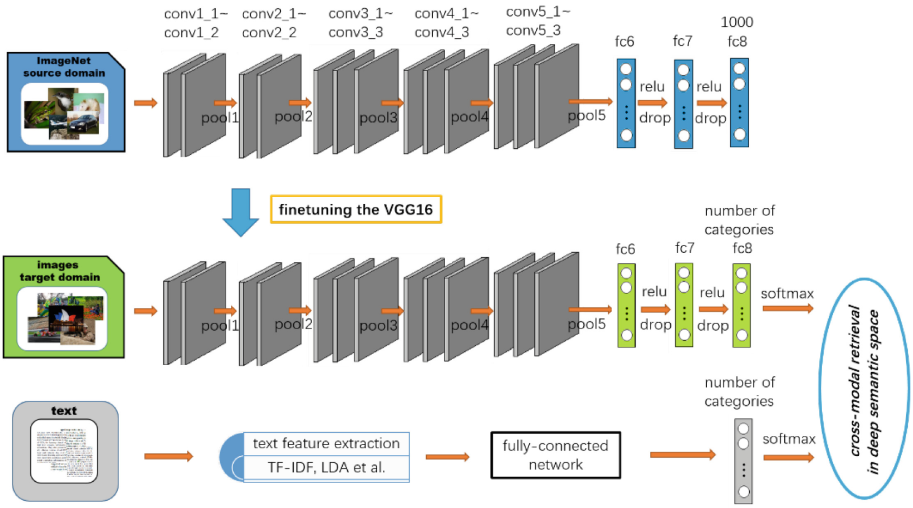


Fig. 2. An overview of our proposed DSS. For images, we use a VGG network to transfer semantic knowledge from ImageNet (in Sect. 3.1). For text, we design a fully-connected network to obtain text semantics (in Sect. 3.2). Finally, the multimodal data is embedded into deep semantic space for cross-modal retrieval (in Sect. 3.3).

3.1 Learning Image Semantics with Knowledge Transfer

Considering the degradation problem faced by the deep learning models on dealing with large-scale dataset, we propose to learn image semantics by VGG network.

More specifically, we fine-tune the VGG network by initializing the parameters with a network pre-trained on the ImageNet dataset. Furthermore, we feed the output of the last fully-connected layer o_I into a softmax, which generates image semantic embedding $S_I \epsilon R^K$ over a number of K event categories. Intuitively, the softmax function maps a K-dimensional vector z to a K-dimensional vector $\sigma(z)$ of real values in the range (0, 1) that add up to 1. The image semantic embedding S_I is defined below:

$$\sigma : R^K \rightarrow \left\{ Z \in R^K | z_i \geq 0, \sum\nolimits_{i=1}^{K} z_i = 1 \right\} \tag{1}$$

$$z_j = (o_I)_j \quad \text{for} \quad j = 1, \ldots, K. \tag{2}$$

$$(S_I)_j = P(y = j|I) = \sigma(z)_j = \frac{e^{z_j}}{\sum\nolimits_{i=1}^{K} e^{z_k}} \quad \text{for} \quad j = 1, \ldots, K. \tag{3}$$

where $P(y = j|I)$ represents the predicted probability for the j-th class given a data sample I. $S_I \in R^K$ is the image semantic embedding vector. $(S_I)_j$ represents the j-th element in the vector.

3.2 Learning Text Semantics by Fully-Connected Network

We design a 4-layer fully-connected network based on raw textual features to obtain the text semantics as show in Fig. 3. More specifically, we take term frequency–inverse document frequency (TF-IDF) as an example to illustrate the semantic learning process for the textual content. Stop words have been removed before obtaining vectors consisting of TF-IDF values for the textual documents. The dimension of the vectors is equal to the number of tokens in the corpus. Furthermore, we design a 4-layer fully-connected network to learn the hidden semantics underlying the documents, which is defined below:

$$f(x) = \max(0, x) \tag{4}$$

$$h_t^{(2)} = f^{(2)} \left(W_t^{(1)} \cdot T + b_t^{(1)} \right) \tag{5}$$

$$h_t^{(3)} = f^{(3)} \left(W_t^{(2)} \cdot h_t^{(2)} + b_t^{(2)} \right) \tag{6}$$

$$o_T = f^{(4)} \left(W_t^{(3)} \cdot h_t^{(3)} + b_t^{(3)} \right) \tag{7}$$

where T represents the input TF-IDF features for each document, $f(x)$ represents the rectified linear unit (ReLU) function, i.e., the activation function, and o_T represents the output of the last fully-connected layer.

Finally, o_T is fed into a K-way softmax, which generates text semantic embedding $S_T \epsilon R^K$ over a number of K categories. The text semantic embedding S_T is defined below:

$$\sigma : R^K \rightarrow \{z \in R^K | z_i \geq 0, \quad \sum\nolimits_{i=1}^{K} z_i = 1\} \tag{8}$$

$$z_j = (o_T)_j \quad \text{for} \quad j = 1, \ldots, K. \tag{9}$$

$$(S_T)_j = P(y = j\,|\,T) = \sigma(z)_j = \frac{e^{z_j}}{\sum\nolimits_{i=1}^{K} e^{z_k}} \quad \text{for} \quad j = 1, \ldots, K. \tag{10}$$

where $P(y = j|\,T)$ represents the predicted probability for the j-th class given a data sample T. $S_T \in R^K$ is the text semantic embedding. $(S_T)_j$ represents the j-th element in the vector.

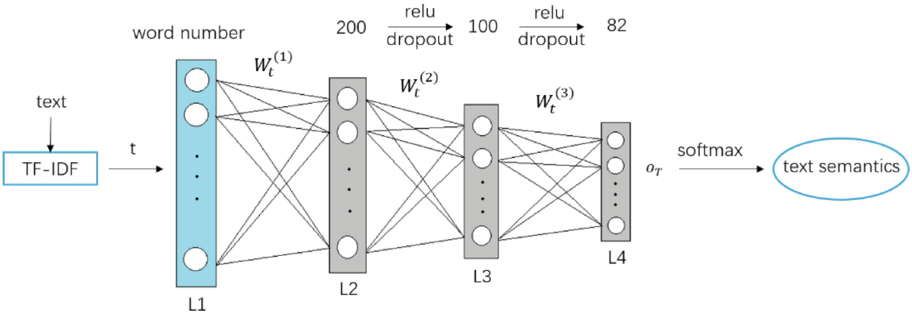


Fig. 3. Learning text semantics by the fully-connected network.

3.3 Semantic Matching in the Deep Semantic Space

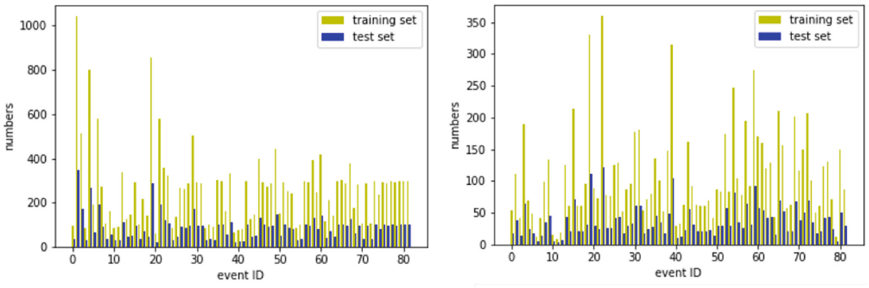
As mentioned previously, we obtain unified vectors for both image and text, which is called the deep semantic space (DSS). Therefore, cross-model data samples can be measured directly in DSS by using distance metrics, e.g., Euclidean distance, cosine distance, Kullback-Leibler (KL) divergence, Normalized Correlation (NC). In the experiments, we will investigate the influence of various distance metrics.

4 Experiments

4.1 Dataset and Data Partitions

(1) **Pascal Sentence dataset**: It is a subset of Pascal VOC, which contains 1000 pairs of image and text descriptions (several sentences) from 20 categories. We randomly select 30% pairs from each category for training and the rest for testing. The text-image pairs are strongly aligned data, i.e., text is the exact description of an image, as shown in Fig. 1.

(2) **Wiki-Flickr Event dataset**: We collect 28,825 images and 11,960 text articles, belonging to 82 categories of events. The images are shared by users on Flickr social media, while the text articles are collected from different news media sites, e.g., BBC News, The New York Times, Yahoo News, Google News. In particular, the data is weakly aligned, as a text article is not the exact description of a certain photo, but they share the same event label. Some examples are shown in Fig. 1, and the statistics on the dataset is shown in Fig. 4. For data partitions, 75% of the data samples are used for training, and the rest are used for testing.



(a) Flickr image

(b) Wikipedia textual articles

Fig. 4. Data partitions of our Wiki-Flickr event dataset.

4.2 Experimental Results

(1) **Implementation Details.** In terms of the images, we crop the images and horizontally flip the images randomly with a given probability of 0.5 for data augmentation. The images are resized to 224 × 224. Then, we normalize the images with mean and standard deviation. The loss function of the network adopts Cross Entropy, which is optimized by using stochastic gradient descent with the momentum of 0.9. The learning rate is 0.001, and the batch size is 16. In terms of the 4-layer fully-connected network for text, we use the same loss function and optimization strategy, the learning rate is 0.01, and the dropout probability is 0.5.

(2) **Evaluation Metrics.** In the experiments, two retrieval tasks are conducted: retrieving text by image queries (denoted as Image → Text) and retrieving images by text queries (denoted as Text → Image). We evaluate the ranking list by mean average precision (MAP). MAP is computed as the mean of average precision (AP) for all the queries.

(3) **Compared with Baselines.** In terms of the performance of our DSS on strongly aligned data pairs, we compare with eight popular approaches on the Pascal Sentence dataset, including CCA [6], CFA [7], JRL [16], LGCFL [17], Multimodal DBN [18], Corr-AE [19], Bimodal-AE [20], and Deep-SM [15]. The performance of the baselines on the Pascal Sentence dataset is reported in their papers, respectively. As shown in Table 1, DNN-based methods [15, 18–20] tend to achieve better performance than the traditional ones [6, 7, 16, 17]. Overall, our

Table 1. MAP performance of the approaches on Pascal Sentence dataset.

Method	Image → Text	Text → Image	Average
CCA	0.110	0.116	0.113
CFA	0.341	0.308	0.325
JRL	0.416	0.377	0.397
LGCFL	0.381	0.435	0.408
Bimodal-AE	0.404	0.447	0.426
Multimodal DBN	0.438	0.363	0.401
Corr-AE	0.411	0.475	0.443
Deep-SM	0.440	0.414	0.427
Our DSS	**0.472**	**0.495**	**0.484**

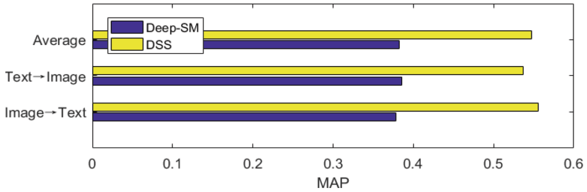


Fig. 5. Comparing DSS with Deep-SM on Wiki-Flickr event dataset.

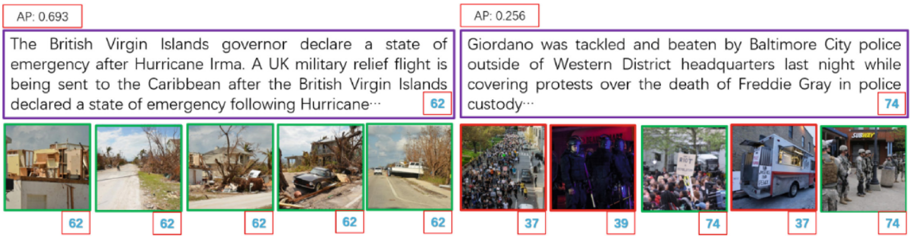
DSS achieves the best performance, giving significant improvement. Furthermore, we compare the proposed DSS with Deep-SM on the Wiki-Flickr event dataset as shown in Fig. 5, which demonstrates the effectiveness of our DSS for weakly align data.

(4) **Evaluation of Network Structures and Distance Metrics.** We evaluate the influence of adopting different network architectures for pre-training, and using different distance metrics in DSS. The evaluations are conducted on the Wiki-Flickr event dataset as shown in Table 2, from which we make two observations: (1) In terms of the network architectures adopted for pre-training, we can observe that deep models, such as VGG [8], and ResNet [21], achieve better performance than the shallow models, such as AlexNet [22], and SqueezeNet [23]. (2) In terms of the distance metrics, normalized correlation and cosine distance perform better than Euclidean distance and KL-divergence in the context of cross-modal event retrieval.

(5) **Examples of the Retrieval Results.** Intuitively, we take text retrieving images as an example to show the performance of DSS on Wiki-Flickr event dataset in Fig. 6. The top-five images are given in the figure, where the event labels are marked at the lower right corner. Red boxes indicate the mismatched retrieval results, while green boxes indicate the correct results. Our DSS returns three mismatched images in the right example. It is probably due to the fact that the categories of ‘Baltimore protests’, ‘Shooting of Michael Brown’ and ‘Death of Freddie Gray’ share quite a few similar words and images, making them hard to be distinguished.

Table 2. MAP performance of adopting different network architectures and distance metrics in our DSS on Wiki-Flickr event dataset

Architecture	Distance metric	Image → Text	Text → Image	Average
AlexNet	KL-divergence	0.451	0.380	0.416
SqueezeNet		0.447	0.409	0.428
ResNet		0.438	0.456	0.447
VGG		**0.494**	**0.447**	**0.471**
AlexNet	Euclidean distance	0.462	0.410	0.436
SqueezeNet		0.466	0.427	0.447
ResNet		0.461	0.464	0.463
VGG		**0.503**	**0.474**	**0.489**
AlexNet	Cosine distance	0.530	0.495	0.513
SqueezeNet		0.539	0.510	0.525
ResNet		0.556	0.537	0.547
VGG		**0.576**	**0.566**	**0.571**
AlexNet	Normalized correlation	0.532	0.498	0.515
SqueezeNet		0.541	0.510	0.526
ResNet		0.560	0.538	0.549
VGG		**0.578**	**0.570**	**0.574**



The British Virgin Islands governor declare a state of emergency after Hurricane Irma. A UK military relief flight is being sent to the Caribbean after the British Virgin Islands declared a state of emergency following Hurricane···

Giordano was tackled and beaten by Baltimore City police outside of Western District headquarters last night while covering protests over the death of Freddie Gray in police custody···

Fig. 6. Two examples of cross-modal retrieval results obtained by DSS on Wiki-Flickr event dataset. Note the numbers refer to the event labels (i.e., 62: Hurricane Irma, 37: Death of Freddie Gray, 39: Shooting of Michael Brown, 74: Baltimore protests, etc.).

5 Conclusion

In this paper, we have proposed a deep semantic space (DSS) learning framework for cross-modal retrieval. DSS embeds multimodal data into a common semantic space with high-level semantics in an end-to-end manner, which casts cross-modal retrieval problem as a homogeneous retrieval task. In particular, we collect a Wiki-Flickr event dataset to advocate the problem of cross-modal retrieval for weakly aligned data. Extensive experiments conducted on a public dataset and Wiki-Flickr event dataset show that our DSS outperforms the state-of-the-art approaches.

Acknowledgments. The authors would like to thank Zehang Lin and Feitao Huang for data collection. This work is supported by the National Natural Science Foundation of China (No. 61703109, No. 91748107, No. U1611461), the Guangdong Innovative Research Team Program (No. 2014ZT05G157), Science and Technology Program of Guangdong Province, China (No. 2016A010101012), and CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, 100190, Beijing, China (No. CASNDST201703), and an internal grant from City University of Hong Kong (Project No. 9610367).

References

1. Yang, Z., Li, Q., Lu, Z., Ma, Y., Gong, Z., Liu, W.: Dual structure constrained multimodal feature coding for social event detection from Flickr data. ACM Trans. Internet Technol. **17** (2), 19 (2017)
2. Yang, Z., Li, Q., Liu, W., Ma, Y., Cheng, M.: Dual graph regularized NMF model for social event detection from Flickr data. World Wide Web **20**(5), 995–1015 (2017)
3. 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: 18th ACM International Conference on Multimedia, pp. 251–260. ACM (2010)
4. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using Amazon's Mechanical Turk. In: NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 139–147. Association for Computational Linguistics (2010)
5. Hwang, S.J., Grauman, K.: Reading between the lines: object localization using implicit cues from image tags. IEEE Trans. Pattern Anal. Mach. Intell. **34**(6), 1145–1158 (2012)
6. Thompson, B: Canonical correlation analysis. In: Encyclopedia of Statistics in Behavioral Science (2000)
7. Li, D., Dimitrova, N., Li, M., Sethi, I. K.: Multimedia content processing through cross-modal association. In: 11th ACM International Conference on Multimedia, pp. 604–611. ACM (2003)
8. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition (2014). arXiv preprint arXiv:1409.1556
9. Bronstein, M. M., Bronstein, A. M., Michel, F., Paragios, N.: Data fusion through cross-modality metric learning using similarity-sensitive hashing. In: Computer Vision and Pattern Recognition, pp. 3594–3601 (2010)
10. Song, J., Yang, Y., Yang, Y., Huang, Z., Shen, H.T.: Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: 2013 ACM SIGMOD International Conference on Management of Data, pp. 785–796. ACM (2013)
11. Bai, B., Weston, J., Grangier, D., Collobert, R., Sadamasa, K., Qi, Y., Weinberger, K.: Learning to rank with (a lot of) word features. Inf. Retr **13**(3), 291–314 (2010)
12. Grangier, D., Bengio, S.: A discriminative kernel-based approach to rank images from text queries. IEEE Trans. Pattern Anal. Mach. Intell. **30**(8), 1371–1384 (2008)
13. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. Adv. Neural Inf. Process. Syst. **5**, 2222–2230 (2012)
14. Wang, C., Yang, H., Meinel, C.: Deep semantic mapping for cross-modal retrieval. In: Tools with Artificial Intelligence, pp. 234–241. IEEE (2015)
15. Wei, Y., Zhao, Y., Lu, C., Wei, S., Liu, L., Zhu, Z., Yan, S.: Cross-modal retrieval with cnn visual features: A new baseline. IEEE Trans. Cybern. **47**(2), 449–460 (2017)

16. 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)
17. Kang, C., Xiang, S., Liao, S., Xu, C., Pan, C.: Learning consistent feature representation for cross-modal multimedia retrieval. IEEE Trans. Multimedia **17**(3), 370–381 (2015)
18. Srivastava, N., Salakhutdinov, R.: Learning representations for multimodal data with deep belief nets. In: International Conference on Machine Learning Workshop, vol. 79 (2012)
19. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: 22nd ACM International Conference on Multimedia, pp. 7–16. ACM (2014)
20. 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)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
22. Krizhevsky, A.: One Weird Trick for Parallelizing Convolutional Neural Networks (2014). arXiv preprint arXiv:1404.5997
23. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size (2016). arXiv preprint arXiv:1602.07360