CROSS-MODALITY MATCHING BASED ON FISHER VECTOR WITH NEURAL WORD EMBEDDINGS AND DEEP IMAGE FEATURES

Liang Han, Wenmin Wang, Mengdi Fan, Ronggang Wang*

School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University
Lishui Road 2199, Nanshan District, Shenzhen, China 518055
bualua61@163.com, wangwm@ece.pku.edu.cn, fanmengdi@sz.pku.edu.cn, rgwang@ece.pku.edu.cn

ABSTRACT

Cross-modal retrieval, which aims to solve the problem that the query and the retrieved results are from different modality, becomes more and more essential with the development of the Internet. In this paper, we mainly focus on the exploration of high-level semantic representation of image and text for cross-modal matching. Deep convolutional image features and Fisher Vector with neural word embeddings are utilized as visual and textual features respectively. To further investigate the correlation among heterogeneous multimodal characteristics, we use multiclass logistic classifier for semantic matching across modalities. Experiments on Wikipedia and Pascal Sentence dataset demonstrate the robustness and effectiveness for both *Img2Text* and *Text2Img* retrieval tasks.

Index Terms— Cross-modal retrieval, Fisher Vector, deep CNN image features, cross-modality matching

1. INTRODUCTION

Over the last decade, with the rapid development of Web 2.0, social media and other information technologies, there has been an explosive growth of data with various modalities. Different types of data are frequently used to describe the same objects or topics, and it's necessary to obtain multimodal data to meet the need of an overall understanding of things. Therefore, cross-modal retrieval is becoming imperative for current Internet environment, such as using pictures to search relevant background music or get pertinent videos via a piece of news. In this paper, we focus on the image-text cross-modal retrieval problem. The challenge of cross-modal retrieval lies in two aspects. The first one is how to represent texts and images effectively. And the second one is the difficulty to match the related multimodal data.

Recently, various approaches have been proposed to address cross-modal retrieval problem. One of the most popular methods is to learn an optimal common representation space of multimodal data. They project representations of multimodal data into a common subspace, in which cross-modal relevance can be computed. Dong et al. [1] categorized existing works into three groups depending on the choice of the common space, namely textual space, visual space, and joint space. In the first group, the image is represented by a bag-of-word vector [2]. For example, in ConSE model [3], an image is embedded into the Word2Vec space, achieved by a convex combination of the word embedding vectors of the visual labels predicted to be most relevant to the image. The second group performs cross-modal retrieval in a visual space only. Dong et al. [4] proposed a deep neural network architecture named *Word2VisualVec* that learns to predict a deep visual encoding of textual input. Lastly, joint space based methods projects both images and texts into a learned space. Rasiwasia et al. [5] proposed a two stage method for cross-modal retrieval. CCA is firstly used to learn a common subspace by maximizing the correlation between two modalities. Then, a semantic space is learned to measure the similarity of different modal features. Sharma et al. [6] extended CCA to generalized multiview analysis (GMA) to map data representations in different modality spaces into a common linear subspace. Habibian et al. [7] contributed to semantic alignment by automatically learning its underlying semantic vocabulary.

Although these methods have made some contributions to the solution of cross-modal retrieval, their performance are mostly far from satisfactory. The reason may be that most of the existing work paid much attention on learning mapping functions but neglected the exploration of high-level semantic representation of multimodal data. So both the text and image features extracted by traditional feature techniques can't effectively express their semantics.

Most of the current works extracted text features by means of Latent Dirichlet Allocation(LDA) [8]. LDA is a high-level semantic representation method which uses a generative probabilistic model for collections of discrete data such as text corpora, where the content is summarized as a mixture of topics. But it has the weakness of ignoring semantics of words. Word2vec [9] is a recently developed technique for building a neural network that maps words to real-number vectors, with the desideratum that words with similar meanings will map to similar vectors. In our work, we employ word2vec and map

This project was supported by Shenzhen Peacock Plan.

every word in the sentence to a vector. All of the vectors that belong to a sentence are then pooled into a single vector by using Fisher Vector with Gaussian Mixture Model (GMM).

As for images, the traditional hand-crafted features such as Scale-Invariant Feature Transform (SIFT) [10] and Histogram of Oriented Gradients (HOG) [11] are frequently used, which have limited the performance of cross-modal retrieval seriously as these traditional visual features can't depict the deep semantics of images sufficiently. Recently, Convolutional Neural Networks (CNN) [12][13] have achieved a great success in large-scale image and video recognition. Especially, a big breakthrough was made by Kriszhevsky et al. [12] who demonstrated for the first time the supremacy of deep learning representation over shallow representations. And deep CNN has now been recognized as an effective image feature extractor. In particular, we employ a pre-trained model, VGG-net [12], to explore the high-level semantic representation of images.

The main contributions of this paper are as follows: 1) Inspired by Fisher Kernels on GMMs applied to image categorization [14], we extend Fisher Vector to represent text features. The proposed method has the ability to excavate the semantics of words and make a high-level representation of texts. 2) We investigate using off-the-shelf VGG-net visual features to represent images. Compared with some hand-crafted visual features, deep CNN visual features can achieve excellent improvement. 3) We apply semantic matching method through multiclass logistic regression to build a natural correspondence between the text and image spaces.

2. PROPOSED METHOD

In this section, we demonstrate using Fisher Vector with neural word embeddings to extract text features, employing VGG-net to extract deep image features and semantic matching to transfer both of the features to a common semantic space. The proposed framework is shown in Fig.1.

2.1. Fisher Vector with Neural Word Embeddings

We employ the Skip-gram architecture of Word2Vec [9] to match every word in a sentence with a specific vector [15]. $X = \{x_i \mid i = 1...T\}$ denotes the set of word vectors extracted from a sentence, where T represents the number of words in the sentence. Let $\lambda = \{\omega_i, \mu_i, \Sigma_i \mid i = 1...G\}$ denote the set of parameters of GMM, where ω_i, μ_i and Σ_i represent the weight, mean vector and covariance matrix of the i^{th} Gaussian function respectively and G denotes the number of Gaussians. We denote $L(X|\lambda)$ as the likelihood that X is generated by GMM model under an independence assumption:

$$L(X|\lambda) = \sum_{i=1}^{T} \log p(x_i|\lambda) \tag{1}$$

where $p(x_i|\lambda) = \sum_{j=1}^{G} \omega_j p_j(x_i|\lambda)$ represents the probability of word vector $x_i (i \in [1, T])$ produced by GMM. The weight ω is constrained by: $\sum_{j=1}^{G} \omega_j = 1$ and $p_i(x|\lambda)$ is given by :

$$p_i(x|\lambda) = \frac{\exp\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\}}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \tag{2}$$

where D represents the dimension of the word vector, and $|\Sigma_i|$ represents the determinant of covariance matrix Σ_i.

By the above definition, we can derive the following functions:

$$\frac{\partial L(X|\lambda)}{\partial \omega_i} = \sum_{t=1}^{T} \left[\frac{\gamma_t(i)}{\omega_i} - \frac{\gamma_t(1)}{\omega_1} \right] \quad for\ i \geq 2 \tag{3}$$

$$\frac{\partial L(X|\lambda)}{\partial \mu_i^d} = \sum_{t=1}^{T} \gamma_t(i) \left[\frac{x_t^d - \mu_i^d}{(\sigma_i^d)^2} \right] \tag{4}$$

$$\frac{\partial L(X|\lambda)}{\partial \sigma_i^d} = \sum_{t=1}^{T} \gamma_t(i) \left[\frac{(x_t^d - \mu_i^d)^2}{(\sigma_i^d)^3} - \frac{1}{\sigma_i^d} \right] \tag{5}$$

where the superscript d denotes the d^{th} dimension of a vector, and $\gamma_t(i)$ represents the likelihood that vector x_t is generated from the i^{th} Gaussian function, which is given by:

$$\gamma_t(i) = \frac{\omega_i p_i(x_t|\lambda)}{\sum_{j=1}^{G} \omega_j p_j(x_t|\lambda)} \tag{6}$$

We need to compute the diagonal of Fisher Information Matrix F to normalize the dynamic range of the different dimensions of the gradient vectors. We denote f_{ω_i}, $f_{\mu_i^d}$ and $f_{\sigma_i^d}$ to be the terms of the diagonal of F that correspond respectively to $\frac{\partial L(X|\lambda)}{\partial \omega_i}$, $\frac{\partial L(X|\lambda)}{\partial \mu_i^d}$ and $\frac{\partial L(X|\lambda)}{\partial \sigma_i^d}$. Then, the normalized partial derivatives are $f_{\omega_i}^{-1/2} \cdot \frac{\partial L(X|\lambda)}{\partial \omega_i}$, $f_{\mu_i^d}^{-1/2} \cdot \frac{\partial L(X|\lambda)}{\partial \mu_i^d}$ and $f_{\sigma_i^d}^{-1/2} \cdot \frac{\partial L(X|\lambda)}{\partial \sigma_i^d}$. We calculate the approximated values of f_{ω_i}, $f_{\mu_i^d}$ and $f_{\sigma_i^d}$ by the following functions:

$$f_{\omega_i} = T(\frac{1}{\omega_i} + \frac{1}{\omega_1}), \quad for\ i \geq 2 \tag{7}$$

$$f_{\mu_i^d} = \frac{T\omega_i}{(\sigma_i^d)^2} \tag{8}$$

$$f_{\sigma_i^d} = \frac{2T\omega_i}{(\sigma_i^d)^2} \tag{9}$$

Due to the constraint of the weight ω: $\sum_{j=1}^{G} \omega_j = 1$, Eq.3 and Eq.7 are defined for i ≥ 2. Fisher Vector is constructed by a concatenation of the normalized partial derivative of all the parameters of the GMM. The dimension of the Fisher Vector is $(2 \times D + 1) \times G - 1$.

2.2. Deep Image Features using VGG-net

In our framework, we use VGG-net [12] to extract image feature vectors. This network was one of the winners of the ImageNet challenge in 2014. The VGG-nets are originally developed for object recognition and detection, which have very

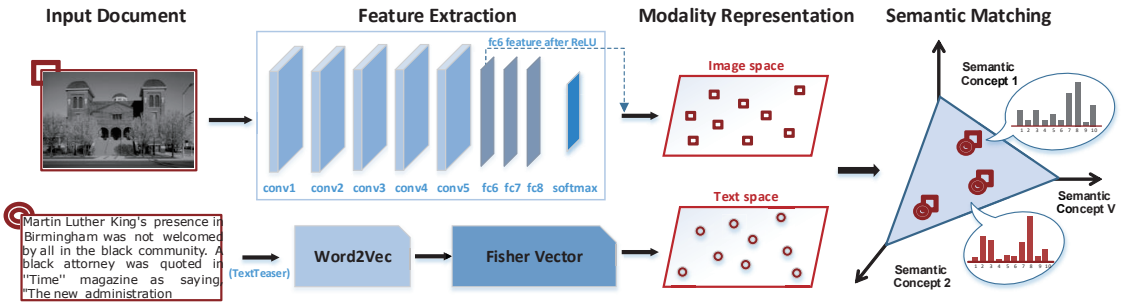

Fig. 1. The framework of the proposed method. The fine-tuned CNN image features are extracted from VGG-net, and Fisher Vector with neural word embeddings using are employed as textual features. Both of them are inputs for semantic matching.

deep convolutional architectures with smaller sizes of convolutional kernel (3×3), stride (1×1), and pooling window (2×2). There are four different network structures, ranging from 11 layers to 19 layers. The model capability increases when the network goes deeper and the computational cost becomes heavier. So we choose the deepest structure with 19 layers. The model is trained on a large image dataset 'Imagenet' and we utilize the pretrained model to directly extract the 4096 dimensional features of the first fully-connected layer after the Rectified Linear Units (fc6-4096) as image feature vectors, which is considered preferable.

2.3. Semantic Matching

After extracting text and image features, we apply Semantic Matching (SM) approach [5] as our matching method to transfer features of different modalities to the common semantic space. Semantic matching uses the multiclass logistic classifier according to a same class label $V = \{j \mid j = 1...M\}$, where M represents the number of classes. For the j^{th} class, logistic regression computes the posterior probability as:

$$P_{V|X}(j|x;\omega) = \frac{1}{Z(x,\omega)} \exp(w_j^{\mathrm{T}} x) \tag{10}$$

where $Z(x,\omega) = \sum_j exp(w_j^{\mathrm{T}} x)$ is a normalization constant, X represents the vector of features in the input space, and w_j stands for the parameter vector of class j. Then we'll get the likelihood that texts and images belong to their specific classes and we employ those probability values as the semantic concept features of the texts and images. Thus, we convert text and image features with different morphology to a higher level of isomorphic abstraction. We adopt several popular similarity measurement methods including Euclidean distance, Kullback-Leibler divergence (KL), Normalized Correlation (NC) and Centered Correlation (CC) to calculate distance between query vectors and retrieved vectors. And Centered Correlation (CC) is found to get better retrieval performance than the others.

3. EXPERIMENTAL RESULTS

3.1. Datasets and Metrics

1) Wikipedia[1] [5]: This dataset consists of 2866 text-image pairs, each annotated with a label from 10 most popular semantic classes. A random split is used to produce a training set of 2173 pairs and a testing set of 693 pairs. For images, we employ 128 dimensional hand-crafted SIFT BoVW features and DeCAF CNN features [16] to compare with our VGG-net image features. As the texts in Wikipedia are presented in paragraph form, we adopt an automatic summarization tool, *TextTeaser*[2], to draw topic sentences. Then we use 10 dimensional LDA [5] and 100 dimensional LDA [17] to compare with our Fisher Vector text features.

2) Pascal Sentence[3] [18]: This dataset contains 1000 pairs of images and text descriptions (about 4-5 sentences) from 20 categories (50 for each category). We randomly select 30 pairs from each category for training and the rest for testing. And we adopt the same features as Wikipedia dataset for fair comparison purpose.

The mean average precision (MAP) score is used to evaluate the retrieval performance. The average precision (AP) of N retrieved target objects is defined as

$$AP = \frac{1}{T} \sum_{k=1}^{N} P(k) rel(k) \tag{11}$$

where T represents the number of the relevant objects in the retrieved set, $P(k)$ is the precision of the top k retrieved objects and $rel(k)$ is an indicator function. $rel(k) = 1$, if the k^{th} retrieved object is relevant, $rel(k) = 0$ otherwise. The MAP score is calculated by averaging the AP values from all the queries in the query set.

[1] http://www.svcl.ucsd.edu/projects/crossmodal/
[2] https://github.com/DataTeaser/textteaser
[3] http://vision.cs.uiuc.edu/pascal-sentences/

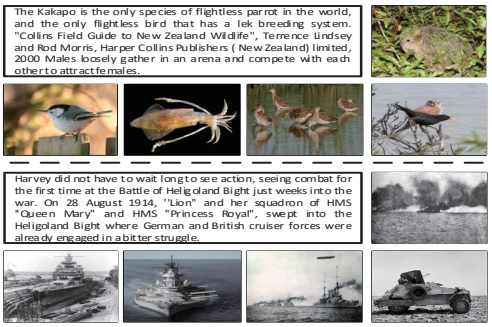

Fig. 2. Two examples of *Text2Img* on Wikipedia dataset. The text query and its associated ground truth image are shown on the top, the retrieved images are shown at the bottom.

3.2. Results on Wikipedia dataset

Wikipedia dataset is frequently used for cross-modal retrieval evaluation, and many articles have performed experiments on this dataset to verify their proposed methods. To validate the effectiveness of our proposed framework with Fisher Vector and Deep image features, we compare some of the state-of-the-art methods which used the same train/test division as ours and the results are show in Table 1. We can observe that our framework outperforms these methods with a large margin of more than 8%. Fig.2 shows two examples of *Text2Img* by our method.

Table 1. Performance comparison with state-of-the-art method on Wikipedia dataset

Methods	*Img2Text*	*Text2Img*	Average
GMLDA [6]	27.2	23.2	25.3
LCFS [19]	28.0	21.4	24.7
CMCP [20]	32.6	25.1	28.9
CFMCM [21]	32.5	23.7	28.1
ClusterKCCA [22]	36.5	28.8	32.7
SCM [23]	36.2	27.3	31.8
Ours	**44.9**	**38.7**	**41.8**

To further explore the effectiveness of Deep CNN image features and Fisher Vector text features, based on our framework, we compare them with other feature vectors and Table 2 presents the results. When fixing the text features of LDA and changing image features (the first 3 lines of the results), it shows that CNN visual features including DeCAF and VGG obtain significant improvements compared with hand-crafted SIFT features, but our VGG features perform even better. We also substitute the Fisher Vector for LDA to compare the performance of different text features (line 1 and 5, 3 and 6). The results show that Fisher Vector can get similar or slightly weaker performance to LDA. We think that extracting topic sentence from text data leads to some information loss. The main reason for the great performance improvement is the robustness and effectiveness of deep image features.

Table 2. MAP scores of different features based on our proposed framework on Wikipedia dataset

#	Text features	Image features	*Img2Text*	*Text2Img*	Average
1	LDA-100	SIFT	30.9	22.3	26.6
2	LDA-100	DeCAF	43.0	37.0	40.0
3	LDA-100	VGG	46.5	39.8	**43.1**
4	LDA-10	VGG	45.1	39.8	42.5
5	FV	SIFT	30.0	21.8	25.9
6	FV	VGG	44.9	38.7	41.8

3.3. Results on Pascal Sentence dataset

We also perform comparison on Pascal Sentence dataset shown in Table 3. Similar to the results of Wikipedia, VGG visual features still perform best among three image features, but Fisher Vector is slightly more effective than LDA. We believe the reason is that each text data is composed of 4-5 sentences from which we can directly extract Fisher Vector features. The optimal average MAP score (56.7%) is achieved by using Fisher Vector text features and VGG visual features. And this is the best result we currently know for this dataset in the field of cross-modal retrieval.

Table 3. MAP scores of different features based on our proposed framework on Pascal Sentence dataset

#	Text features	Image features	*Img2Text*	*Text2Img*	Average
1	LDA-100	SIFT	20.3	11.5	15.9
2	LDA-100	DeCAF	49.6	46.0	47.8
3	LDA-100	VGG	57.3	55.1	56.2
4	FV	SIFT	21.7	11.8	16.8
5	FV	VGG	57.2	56.2	**56.7**

4. CONCLUSION

In this paper, we propose an efficient framework in the processing of cross-modal retrieval, which employs semantic matching to mine the correlation between CNN visual features and Fisher Vector text representations. Experimental results show that VGG-net deep features make better performance than the hand-crafted SIFT features and other CNN features. Fisher Vector text features perform sometimes better than LDA. By comparing with several state-of-the-art methods, ours also demonstrates great superiority. Our future work will focus on exploring some more appropriate neural networks such as RNN to build high-level semantic text features.

5. REFERENCES

[1] J. Dong, X. Li, and S. Liao, "Image retrieval by cross-media relevance fusion," in *ACMMM*, 2015.

[2] Y. Bai, W. Yu, T. Xiao, et al., "Bag-of-words based deep neural network for image retrieval," in *ACMMM*, 2014.

[3] M. Norouzi, T. Mikolov, S. Bengio, et al., "Zero-shot learning by convex combination of semantic embeddings," *E-print arXiv*, 2013.

[4] J. Dong, X. Li, and C. G. M. Snoek, "Word2visualvec: cross-media retrieval by visual feature prediction," in *CVPR*, 2016.

[5] N. Rasiwasia, J. C. Pereira, E. Coviello, et al., "A new approach to cross-modal multimedia retrieval," in *ACM-MM*, 2010.

[6] J. Sharma, H. Kumar, A. Daume III, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *CVPR*, 2012.

[7] A. Habibian, D. Mensink, and C.G.M Snoek, "Discovering semantic vocabularies for cross-media retrieval," in *ICMR*, 2015.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, 2003.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *Computer Science*, 2013.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 2004.

[11] Dalal N. and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2013.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[14] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*, 2007.

[15] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," in *CVPR*, 2015.

[16] J. Donahue, Y. Jia, O. Vinyals, et al., "Decaf: A deep convolutional activation feature for generic visual recognition," *Computer Science*, 2013.

[17] Y. Wei, Z. Yao, C. Lu, and S. Wei, "Cross-modal retrieval with cnn visual features: a new baseline," *IEEE Transactions on Cybernetics*, 2016.

[18] Rashtchian. C., P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Workshop on Creating Speech and Lang.Data with Amazon's Mech.Turk*, 2010.

[19] K. Wang, R. He, W. Wang, et al., "Learning coupled feature spaces for cross-modal matching," in *ICCV*, 2013.

[20] X. Zhai, Y. Peng, and J. Xiao, "Cross-modality correlation propagation for cross-media retrieval," in *ICASSP*, 2012.

[21] M. Fan, W. Wang, and R. Wang, "Coupled feature mapping and correlational mining for cross-media retrieval," in *ICMEW*, 2016.

[22] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," *Aistats*, 2014.

[23] J. C. Pereira, E. Coviello, G. Doyle, et al., "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Transactions on Software Engineering*, 2013.