Image Tag Recommendation via Deep Cross-modal Correlation Mining

Xingmeng Zhang[1], Cheng Jin[1], Yuejie Zhang[1], Tao Zhang[2]

[1]School of Computer Science
Shanghai Key Laboratory of Intelligent Information Processing,
Fudan University, Shanghai 200433, P. R. China
[2]School of Information Management & Engineering,
Shanghai University of Finance & Economics, Shanghai 200433, P. R. China
{15210240104, jc, yjzhang}@fudan.edu.cn, taozhang@mail.shfeu.edu.cn

ABSTRACT . In this paper, a novel image tag recommendation framework is developed by fusing the deep multimodal feature representation and cross-modal correlation mining, which enables the most appropriate and relevant tags to be presented on the image and facilitates more accurate image retrieval. Such an image tag recommendation pattern can be modeled as an inter-related correlation distribution over deep multimodal visual and semantic representations of images and tags, in which the most important is to create more effective cross-modal correlation and measure what degree they are related. Our experiments on a large number of public data have obtained very positive results.

Keywords : Image Tag Recommendation, Deep Multimodal Feature Representation, Cross-modal Correlation Mining, Deep Canonical Correlation Analysis.

1 INTRODUCTION

With the explosive growth of images available both online and offline, especially on some popular websites such as *Flickr*, how to explore the involved contents in images to achieve more effective image retrieval has become an important research focus. Usually, common users are allowed to annotate each image with a series of tags in accordance with their own tendencies. However, existing investigations reveal that only around 50% tags provided by *Flickr* users are indeed related to the images. The quality of tags is far from satisfactory due to the ambiguity, incompleteness and over-subjectivity. The main reason is that because of the semantic gap, there may exist huge uncertainty on the correspondence relationships among visual contents and semantic tags. Thus how to integrating multimodal information sources to enable the objective image tag recommendation has become a critical issue for supporting image retrieval.

The general image recommendation approach attempts to describe an image as a set of tags based solely on image contents [4]. However, tailoring the general image recommendation to image retrieval is challenging in two aspects: a) the semantic gap still largely exists, and annotation tags are often very unreliable; b) due to having little or no semantic information, fail to combine the enough semantics of images, as well as capture the multimodal association of different modalities. In the general recommendation framework [8], the missing of such important

semantic information for image description may result in the huge uncertainty on the correspondence relationships between visual contents and semantic tags, and the performance of retrieval accuracy, efficiency and effectiveness is not very ideal. Therefore, more efforts are put into abstracting images in the deep level and integrating the semantic information from tags to form a novel expression with strong descriptive ability, so as to fully exploit the multimodal attributes in images and tags to support more precise tag recommendation and further improve the retrieval performance [1]. To achieve effective image tag recommendation, two inter-related issues should be addressed simultaneously: 1) the in-depth discovery of valuable multimodal features to characterize images and tags more reasonably; and 2) the cross-modal correlation mining to identify better multimodal correlations between the visual features for images and semantic features for tags. To address the first issue, it is very important to leverage large-scale annotated images for robust visual and semantic mining to achieve more comprehensive multimodal feature representation. To address the second issue, it is very interesting to develop new algorithms for exploiting multimodal features and efficiently explore the cross-modal correlations between different modality attributes in images and tags [13].

Based on these observations above, a novel scheme is developed in this paper for facilitating automatic image tag recommendation to enable more effective image retrieval. Our scheme significantly differs from other earlier work in: a) The meticulous deeper representation for multimodal feature is constructed to learn more representative attributes for visual images and semantic tags; b) The deep cross-modal correlation mining has a strong ability to characterize the deep multimodal correlations between the visual features in images and semantic features in tags, which can alleviate the problem of semantic gap to a great degree; c) A novel image tag recommendation framework is built by fusing the deep multimodal feature representation and cross-modal correlation mining strategies, which enables the most appropriate and relevant tags to be presented on the image. Such an image tag recommendation pattern can be modeled as an inter-related correlation distribution over multimodal visual and semantic representations of images and tags, in which the most important is to create more effective cross-modal correlation and measure what degree they are related. Our experiments on a large number of public data have obtained very positive results.

2 Deep Multimodal Feature Representation

The Convolution Neural Network (CNN) is a feed-forward artificial neural network in machine learning [16], which is biologically-inspired [7]. Recently deep convolutional neural networks have demonstrated promising results in computer vison tasks, such as single-label image classification, image recognition, etc. We consider utilizing a popular deep neural network, Alex-Net, to extract CNN deep visual features [14]. Alex-Net is a typical convolutional neural network, which consists of 5 convolutional layers and 2 fully connected layers. It's really a simple but high-efficient network. Given an image, we extract a 4,096-dimensional feature vector using the pre-trained CNN on the ILSVRC-2012 dataset. We explore Alex layered architecture and use the Alex-Net features for all our experiments. In order to making faster training, we use the Auto-Encoder (AE) strategy to reduce the dimension from 4,096 to 300. Hence, the dimension for the feature expression reduces 93% and the amount of information decreases 20%, thus some redundant information in image contents can be eliminated and the information representation ability can be also further improved.

The Skip-gram model is an efficient method to learn the distributed representations of textual words in a vector space from large amounts of unstructured text data [15]. It learns deep word vector representations that are good at predicting the nearby textual words, which can capture more precise syntactic and semantic relationships among textual words and group similar textual

words together. Thus we leverage the Skip-gram model to construct the deep textual word for better representing the deep semantic properties in annotation tags.

Let D^T be the textual annotation part of the whole multimodal dataset, W denotes all the raw textual words in D^T, and V is the textual word vocabulary. For each textual word w in W, I_w and O_w are the input and output vector representations for w, *Context* (w) represents the nearby textual words of w, here the context window size is set as 5. We define the set of all the input and output vectors for each textual word as a long vector $\omega \in R^{2*|V|*dim}$ and *dim* is the dimension number of the input or output vector, thus the objective function of Skip-gram can be described as:

$$\begin{aligned} BSG(\omega) &= \underset{\omega}{\text{argmax}} \frac{1}{|W|} \sum_{i=1}^{|W|} \sum_{j=1}^{|Context(w_i)|} logP(w_j|w_i) \\ &= \underset{\omega}{\text{argmax}} \frac{1}{|W|} \sum_{i=1}^{|W|} \sum_{i=1}^{|Context(w_i)|} \frac{exp(O_{w_j} \cdot I_{w_i})}{\sum_{k=1}^{|V|} exp(O_{w_j} \cdot I_{w_i})} \end{aligned} \tag{1}$$

Since the computing cost is extremely high for the standard softmax formulation of Skip-gram, the Negative Sampling is utilized to compute $\log P(w_j|w_i)$ approximatively.

$$log(w_j|w_i) = \log\sigma(O_{w_j} \cdot I_{w_i}) + \sum_{k=1}^{m} E_{w_k(w) \sim P(w)} log\sigma(O_{w_j} \cdot I_{w_i}) \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function; and m is the number of negative samples, each sample is drawn from the noise distribution $P(w)$ based on the textual word frequency. The entire textual word vectors are clustered to acquire the new deep textual word vocabulary, and then each word/tag in the annotation is projected to this vocabulary. Thus each annotation can be represented in the form of bag-of-deep-textual-words. Compared to the raw textual word, the main advantage of deep textual word is the consideration of the semantic relationships among raw words, which makes the deep words more representative to describe textual annotations, i.e., semantic tags.

3 DEEP CROSS-MODAL CORRELATION MINING WITH DCCA

To achieve more effective cross-modal image tag recommendation, we have developed a multimodal feature embedding scheme based on Deep Canonical Correlation Analysis (DCCA) [3][9][11] to exploit multiple features of annotated images and explore the multimodal associations between deep visual property features and semantic expression features. To make a clear presentation, we first introduce the general CCA [16], and then develop our DCCA-based multimodal feature embedding pattern.

The standard CCA algorithm is a classic statistical method to multi-view and multi-scale analysis for multiple data sources, which has received much attention in the field of cross-media/cross-modal processing [5]. It aims at finding linear projections for different types of data with the maximum correlation. Due to the limitation of standard CCA, the kernel representation is always integrated into CCA to increase the computation power of linear learning machines, that is, Kernel CCA (KCCA) [6]. As an extension of standard CCA, KCCA can provide a nonlinear function learning method by projecting different types of data into a high dimension feature space. However, it's still difficult for KCCA to find a better correlation between different types of data, because there are not simple linear or nonlinear relationships among real data.

To overcome the drawback of KCCA that the representation for real data is limited by the fixed kernel, deep networks can be integrated to learn flexible nonlinear representations, that is DCCA [2][10]. Deep network is commonly used in the field of feature learning and classification, which plays a great role in learning a complex and nonlinear form to fit real data. DCCA with deep networks means a model with multiple hidden layers, which can simultaneously learn two deep nonlinear mappings of two views that are maximally correlated. The key difference between DCCA and other closely related approaches is that DCCA learns two separate deep encodings with the objective that the learned encodings are as correlated as possible, and such different objectives may have advantages in different settings. Thus DCCA, which could obtain more accurate highly abstract expression of real data via the complex linear and nonlinear transformation among layers, is introduced to make a better solution for image tag recommendation via cross-modal correlation mining.

For projecting multi-modal features to cross-modal features with DCCA, it aims at projecting the multimodal features with different modalities on multiple views into a common subspace and making sure that the correlation between visual features and semantic features could be maximized. We illustrate our architecture in Fig. 1.

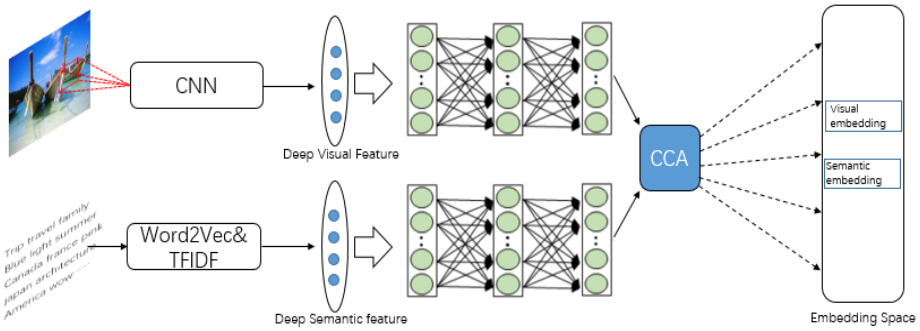


Fig. 1. Our proposed DCCA&DMFR model.

Let *IMG* be the set of annotated images that consists of *N* samples; $X_V \in R^{D_V * N}$ is the visual feature vector for *IMG* and $X_S \in R^{D_S * N}$ represents the semantic feature vector, D_V and D_S are the corresponding dimensionality values for these two vectors, generally $D_V \neq D_S$. Taking these multimodal feature vectors as the input data of the corresponding deep networks including the visual and semantic networks, that is, $a_{V1}=X_V$ and $a_{S1}=X_S$. The visual and semantic networks have D_V and D_S units in the input layers, V_h and S_h units in the hidden layers, and hold V_o and S_o layers in the final output layers respectively. Thus in each network, the feed-forward from the i^{th} layer to the $(i+1)^{th}$ layer can be defined as:

$$a_{Vi+1} = f_V\left(W_{Vi} a_{Vi} + b_{Vi} I^T\right), \quad a_{Si+1} = f_S\left(W_{Si} a_{Si} + b_{Si} I^T\right) \tag{3}$$

where W_{Vi} and W_{Si} represent the weight matrices between the i^{th} layer to the $(i+1)^{th}$ layer in two networks respectively, $W_{Vi} \in R^{D_{Vi+1} * D_{Vi}}$ and $W_{Si} \in R^{D_{Si+1} * D_{Si}}$, D_{*i} and D_{*i+1} are the unit numbers in the i^{th} and $(i+1)^{th}$ layers of two networks; b_{Vi} and b_{Si} denotes the bias values in two networks, $b_{Vi} \in R^{D_{Vi+1}}$, $b_{Si} \in R^{D_{Si+1}}$; I denotes an identity matrix; a_{Vi} and a_{Si} denotes the output data in the i^{th} layers of two networks; and f is equivalent to a nonlinear transformation function for the data in the corresponding layer. Hence, the final matrices from the output layer in two networks can be

achieved as $H_{Vo}, H_{So} \in R^{D_o * N}$. Based on the learning with deep networks, the projection can be implemented as Formula (4).

$$DCCA_{VV} = \frac{1}{n-1}\tilde{H}_{Vo}\tilde{H}_{Vo}^{T} + r_V I,\ DCCA_{SS} = \frac{1}{n-1}\tilde{H}_{So}\tilde{H}_{So}^{T} + r_S I$$
$$DCCA_{VS} = \frac{1}{n-1}\tilde{H}_{Vo}\tilde{H}_{So}^{T} \tag{4}$$

where $\tilde{H}_{Vo}$ and $\tilde{H}_{So}$ represent the central data in the output matrices H_{Vo} and H_{So} from two networks respectively; and r_V and r_S are two regularization factors, which are usually set as small values to avoid the numerically ill-conditioned problem.

To find a projection relationship between the visual feature space and semantic feature space that could maximize the correlation between different feature views, the following Formula (5) is adopted to achieve such a goal.

$$(\theta_V^*, \theta_S^*) = \arg\max_{(\theta_V, \theta_S)} Corr(f_V(X_V;\theta_V), f_S(X_S;\theta_S))$$
$$Corr(f_V(X_V;\theta_V), f_S(X_S;\theta_S)) = Corr(\tilde{H}_{Vo}, \tilde{H}_{So}) = \|T\|_{tr} = tr(T^T T)^{\frac{1}{2}} \tag{5}$$
$$T = DCCA_{VV}^{-\frac{1}{2}} DCCA_{VS} DCCA_{SS}^{-\frac{1}{2}}$$

where θ_V and θ_S are two parameter sets of (W_{Vi}, b_{Vi}) and (W_{Si}, b_{Si}) form the input layer to the output layer in the visual and semantic network respectively. Because the maximal correlation value is 1, to convert the above maximization problem into a usual minimization problem, Formula (5) can be further slightly transformed into Formula (6).

$$(\theta_V^*, \theta_S^*) = \arg\min_{(\theta_V, \theta_S)} \left(\frac{1}{2}\sum_{j=1}^{D_o} \left(1 - Corr(f_V(X_V;\theta_V), f_S(X_S;\theta_S))^2\right)\right) \tag{6}$$

To optimize the correlation measure, the gradient-based optimization is introduced to achieve better training for the parameters of W and b. Suppose the singular value decomposition of T is $T = UDV^T$, U and V are singular vectors, the gradient of correlation function can be computed with respect to $\tilde{H}_{Vo}$ and $\tilde{H}_{So}$ and then the backpropagation is utilized, as shown in Formula (7).

$$\frac{\partial Corr(\tilde{H}_{Vo}, \tilde{H}_{So})}{\partial \tilde{H}_{Vo}} = \frac{1}{n-1}\left(2\nabla_{VoVo}\tilde{H}_{Vo} + \nabla_{VoSo}\tilde{H}_{So}\right)$$
$$\nabla_{VoVo} = -\frac{1}{2} DCCA_{VoVo}^{-\frac{1}{2}} UDU^T DCCA_{VoVo}^{-\frac{1}{2}} \tag{7}$$
$$\nabla_{VoSo} = DCCA_{VoVo}^{-\frac{1}{2}} UDU^T DCCA_{SoSo}^{-\frac{1}{2}}$$

Similarly, $\frac{\partial Corr(\tilde{H}_{Vo}, \tilde{H}_{So})}{\partial \tilde{H}_{So}}$ has a symmetric expression. Hence, the set of the projection matrices of $P=\{p_1, p_2, \ldots, p_R\}$ and $Q=\{q_1, q_2, \ldots, q_R\}$ can be obtained by considering them as a symmetric eigenvalue problem. Thus based on the matrices of $P, Q \in R^{D_o * D_o}$, we can embed the multimodal features (i.e., visual feature H_{Vo} and semantic feature H_{So}) into a common subspace that can generate the cross-modal correlation, as shown in Formula (8).

$$P = DCCA_{V_oV_o}^{-\frac{1}{2}}U, \qquad Q = DCCA_{S_oS_o}^{-\frac{1}{2}}V \tag{8}$$

In fact, for visual features and semantic features involved in each annotated image, they belong to different feature spaces with different dimensions. Our DCCA-based multimodal feature embedding for deep cross-modal correlation mining can provide a relatively perfect feature representation with associations between various features in different modalities, which can mitigate the problem of semantic gap to a certain extent and achieve better multimodal feature associations.

4 Experiment and analysis

4.1 Dataset and Evaluation Metrics

The evaluation for image tag recommendation via deep cross-modal correlation mining requires an image collection with paired images and annotation texts. Thus our dataset is established based on three benchmark datasets of *Corel*5*k*, *Corel*30*k* and *Nus-Wide*. The first and second datasets are two subsets of the *Corel* database, which contain 5,000 images, 260 words or labels and 87 categories, and 31,695 images, 5,587 words or labels and 320 categories respectively. The latter is a web image dataset created by *NUS*'s Lab for Media Search, which includes 269,648 images, the associated tags from *Flickr* with a total of 5,018 unique tags, and 704 categories. Because our tag recommendation method is a supervised task, we follow [8][12] to split the dataset to the training set and testing set.

Our algorithm evaluation focuses on three criteria: 1) how well our deep multimodal feature representation can identify the valuable multimodal features from the images and annotation texts of large-scale annotated image collections; 2) how well our deep cross-modal correlation mining can measure the deep inter-related correlations between visual and semantic features and make an effective integration to construct cross-modal associations; 3) how well our model can support automatic image tag recommendation for large-scale images. To evaluate the effectiveness of these criteria for our image tag recommendation, we compare the annotation results with the available ground-truth labels and employ the benchmark metrics of *Average Precision* (*AP*), *Average Recall* (*AR*) and *Average F-measure* (*AF*), which are defined as:

$$AP = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Pr}ecison(Img_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{|RecTag(Img_i) \cap TagSet(Img_i)|}{|RecTag(Img_i)|} \tag{9}$$

$$AR = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Re}call(Img_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{|RecTag(Img_i) \cap TagSet(Img_i)|}{|TagSet(Img_i)|} \tag{10}$$

$$AF = \frac{|2 * AP * AR|}{|AP + AR|} \tag{11}$$

where $TagSet(Img_i)$ denotes the semantic annotation set for the image Img_i; $RecTag(Img_i)$ denotes semantic annotation set recommended for Img_i; $|RecTag(Img_i) \cap TagSet(Img_i)|$ represents the number of correct semantic annotation tags recommended for Img_i; and n is the number of images in the whole dataset. Meanwhile, a specific *Cross-modal Correlation Score* (*CCS*) is introduced to better exhibit the effect of cross-modal correlation measure, which belongs to the range of [0, 1].

4.2 Experiment on Cross-modal Correlation

The image-tag association mainly focuses on mining the valid and reasonable cross-modal correlation between deep visual features and semantic features. As the definition for the above evaluation metric of *CCA*, cross-modal correlation considers the positive correlation in the common space, that is, the maximum correlation for each multimodal visual-semantic representation pair in the same common space. Thus to verify the effect of our DCCA-based cross-modal correlation mining for acquiring the deep association between different modalities, we make a comparison analysis between different correlation models on three datasets, as shown in Table 1.

Metric	***Cross-modal Correlation Score* (*CCS*)**								
Dataset	***Corel*5*k***			***Corel*30*k***			***Nus-Wide***		
Top-k	**CCA**	**KCCA**	**DCCA**	**CCA**	**KCCA**	**DCCA**	**CCA**	**KCCA**	**DCCA**
1	0.929448	0.93294	0.99481	0.353207	0.820499	0.999997	0.724740	0.55652	0.995945
2	1.84468	1.86209	1.98828	0.676485	1.63814	1.998032	1.429714	1.02952	1.991429
3	2.74386	2.77899	2.98118	0.996206	2.45429	2.995678	2.095074	1.48892	2.986730
4	3.63837	3.67983	3.97299	1.309763	3.26965	3.99262	2.666284	1.92911	3.981379
5	4.52701	4.57421	4.96399	1.610355	4.08462	4.989262	3.226887	2.35637	4.975107
6	5.4081	5.45657	5.95395	1.901771	4.89915	5.985365	3.764116	2.77975	5.968340
7	6.27834	6.33457	6.94298	2.188280	5.71323	6.980934	4.273572	3.1918	6.960630
8	9.58159	7.02072	7.93152	2.466425	6.52701	7.976069	4.761882	3.59934	7.952770
9	10.3853	8.05096	8.91915	2.733498	7.3404	8.970905	5.231056	4.0051	8.944524
10	11.1852	8.88718	9.90616	2.993295	8.15309	9.965175	5.683072	4.40765	9.935961
11	11.9751	9.7185	10.8925	3.245362	8.96542	10.95874	6.125599	4.80745	10.92652
12	12.7614	10.5444	11.8785	3.494521	9.77719	11.95198	6.558548	5.20064	11.91612
13	13.5379	11.3587	12.8635	3.740007	10.5888	12.94488	6.981282	5.59122	12.90495
14	14.31.3	12.1617	13.8479	3.984139	4535408	13.93722	7.399391	5.97982	13.89345
15	15.061	12.9619	14.8311	4.220558	13.0214	14.92937	7.809606	6.36719	14.88162
16	15.8035	13.7567	15.8133	4.452925	13.9317	15.92118	8.213106	6.75276	15.86897
17	16.5491	14.5485	16.7953	4.68259	14.6419	16.91252	8.612662	7.13786	16.85460
18	17.2753	15.3372	17.7760	4.904568	15.1354	17.90378	9.000176	7.52246	17.83928
19	17.9891	16.9006	18.7553	5.125131	15.70336	18.89465	9.377608	7.90547	18.82347
20	18.6975	16.6693	19.7339	5.343721	16.2628	19.88460	9.747613	8.287272	19.80716

Table 1. The comparison results between our and other correlation models.

It can be seen from Table 1 that for cross-modal correlation measure on three datasets, we can obtain the best *CCS* value of 0.9999 for *Top*-1 and the best *CCS* sum of 19.8846 for *Top*-20 on *Corel*30*k* in the evaluation pattern of our DCCA-based cross-modal correlation mining. In comparison with the general CCS, the correlation mining performance could be promoted to a great degree based on KCCA and DCCA, and DCCA exhibits more significant positive impact to cross-modal correlation, which further confirms the obvious advantage of our deep cross-modal

correlation mining mechanism with the deep multimodal feature information. Compared the results on *Corel*5*k*, *Corel*30*k* and *Nus-Wide*, the results on *Nus-Wide* appear less performant due to the obvious characteristic differences between these three datasets. *Corel*5*k* and *Corel*30*k* have more normative and consistent annotations, and the images in the same cluster have the higher visual similarity. *Nus-Wide* is established based on the social annotated images from *Flickr*, in which the annotation information is very abundant but with many noisy or error tags, and the images have much higher visual diversity. Although when evaluating on *Nus-Wide* the cross-modal correlation mining performance maybe slightly influenced by such extra phenomena, the *CCS* value for *Top*-1 and the *CCS* sum for *Top*-20 can still reach a relatively high value. We can still observe the promising and positive performance exhibition under the complicated tag recommendation environment of *Nus-Wide*. The best *CCS* value of 0.9959 for *Top*-1 and the *CCS* sum of 19.8071 for *Top*-20 on *Nus-Wide* approach to those on *Corel*30*k* acquired under a relatively more pure recommendation environment. It's also evidenced once more that our cross-modal correlation mining mechanism can be effectively applicable for both the unsophisticated case with less interference information and the complex case with more misinformation effect.

4.3 Experiment on Tag Recommendation

To further explore the applicability and usefulness of our cross-modal correlation mining for supporting image tag recommendation, we particularly carry out four runs to make the evaluation for the recommendation performance on *Corel*5*k* [*C*5] and *Corel*30*k* [*C*30], that is, *DMFR*&*CCA*[*C*5], *DMFR*&*CCA*[*C*30], *DMFR*&*DCCA*[*C*5] and *DMFR*&*DCCA*[*C*30] (with Deep Multimodal Feature Representation [*DMFR*] and CCA/DCCA-based Cross-modal Correlation Mining). The *AF* curves with different feature dimension settings on two datasets are exhibited in Figure 1. The *AP-AR* curves and the *AF* values for such six runs on three datasets are listed in Fig. 2.

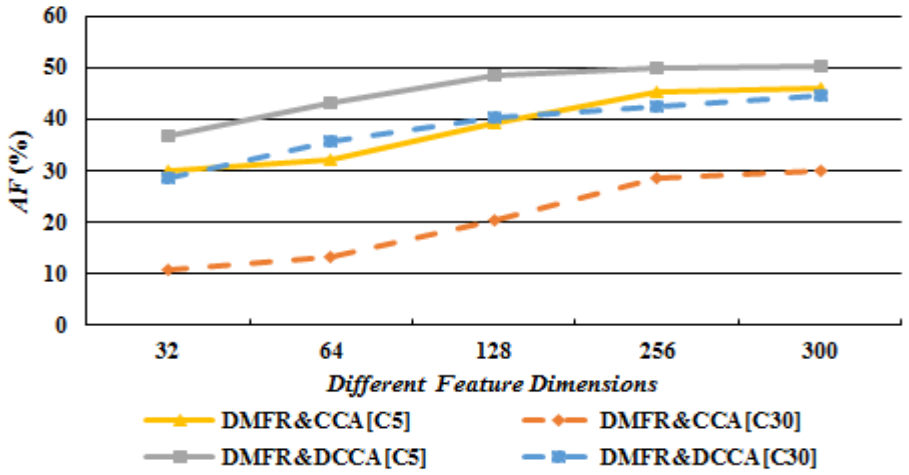
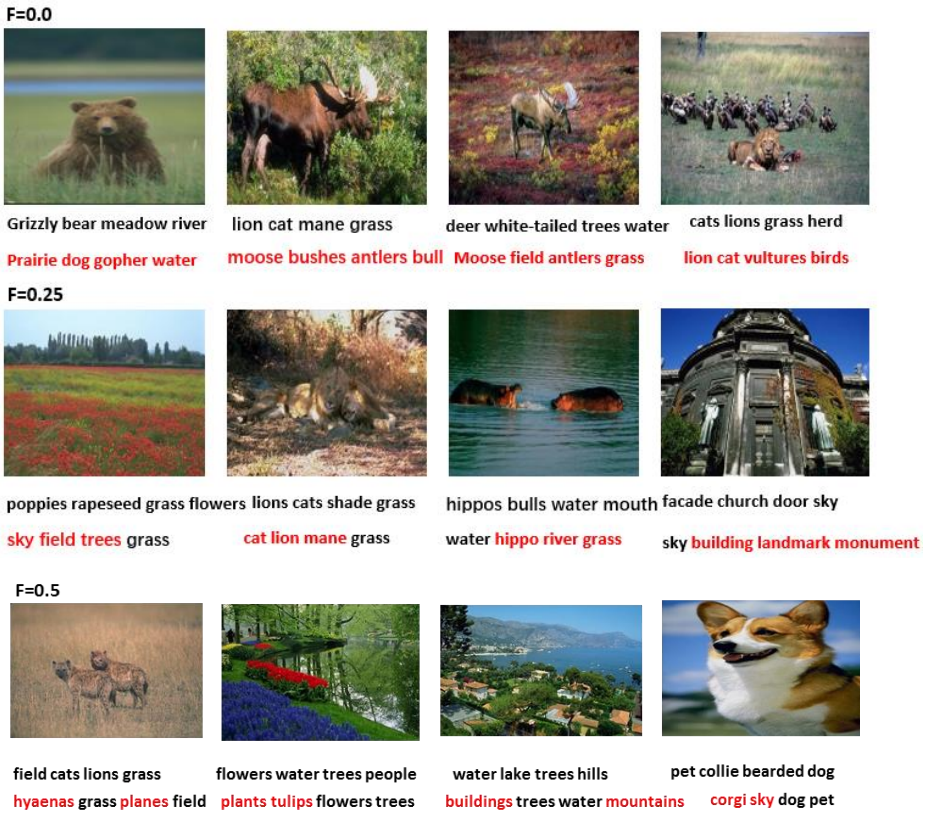


Fig. 2. The *AF* curves with different feature dimension settings.

It can be observed from Fig. 2 that for the tag recommendation on *Corel*5*k* and *Corel*30*k*, we can obtain the best recommendation performance in the evaluation pattern of fusing *DMFR* and *DCCA*. In comparison with the patterns using CCA-based cross-modal correlation measure, the performance could be greatly promoted by integrating the deep multimodal feature representation

and the deep correlation mining, which confirms the obvious advantage of our deep representation and mining for tag recommendation. Comparing the results on two datasets, the results on *Corel*30*k* appear less performant on the whole *AF* curves, while the performance difference for *Corel*5*k* and *Corel*30*k* is not obvious, which reflects the performance advantage to some degree. Meanwhile, compared the results for different feature dimensions, we can observe that with the number of feature dimensions increasing, the recommendation performance gradually becomes pretty good, especially for our approach with *DMFR&DCCA*. These results are consistent with what we expect considering more valuable deep multimodal feature information or given more detailed deep multimodal feature description. It's worth noting that the in-depth multimodal analysis is available and presents more impactful ability for discovering the meaningful deep multimodal features and cross-modal correlations. An instantiation of some recommendation results is shown in Fig. 3.



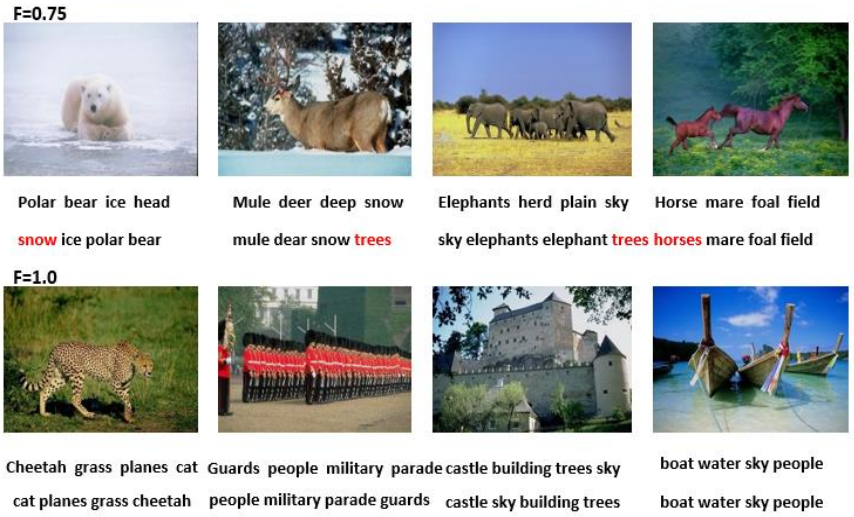


Fig. 3. An instantiation of some recommendation results with our model, in which the upper line represents the ground-truth tag sequence, the lower line denotes the recommended results, and the red words denote the mismatched words.

To give full exhibition to the superiority of our tag recommendation model with deep cross-modal correlation mining, we have also performed a comparison between our approach and the other existing classical methods in recent years. Two methods developed by Makadia *et al*. (2008) [8] and Murthy *et al*. (2014) [3] respectively are analogous with ours to some extent, and then we accomplished them on the same datasets. The experimental results are presented in Table 2, which reflect the difference of power between these four patterns.

Table 2. The comparison results among different approaches.

Approach	**Dataset**								
	Corel*5*k			***Corel*30*k***			***Nus-Wide***		
	AP	***AR***	***AF***	***AP***	***AR***	***AF***	***AP***	***AR***	***AF***
***JEC* [8]**	0.2700	0.3200	0.2900	-	-	-	-	-	-
General CCA	0.3500	0.4600	0.4000	0.2270	0.2250	0.2260	0.1080	0.1296	0.1178
***CCA*-based [3]**	0.4200	0.5200	0.4600	0.2990	0.2980	0.2990	0.1426	0.1716	0.1557
***DMFR*&*DCCA* [Ours]**	0.4991	0.5034	**0.5012**	0.4454	**0.4440**	**0.4446**	**0.2293**	**0.2456**	**0.2371**

It can be found from Table 2 and Fig. 3 that the best performance can be acquired on *Corel*5*k*, *Corel*30*k* and *Nus-Wide* by our approach. When integrating *DMFR* and *DCCA*, we can acquire the obviously better performance than the other patterns. Although based on our approach the values of *AP*, *AR* and *AF* on *Corel*5*k* and *Corel*30*k* may exhibit the obvious increase over those by the other patterns, we can observe that these values on *Nus-Wide* have been also dramatically higher than those by the other patterns. Under such a complicated learning environment of *Nus-Wide*, our approach reveals more significant advantage. This further confirms the prominent roles

of deep multimodal feature representation and cross-modal correlation mining in tag recommendation, which implies that our model is exactly a better way for determining deep multimodal associations between images and annotation tags.

5 Analysis and Discussion

Through the analysis for the tag recommendation results, it can be found that our deep correlation with deep feature for tag recommendation is effective., but it's quality is highly related to the training set. Because correlation method is sensitive to data noise. It's easier to introduce error or noisy detections for visual and semantic feature information, which will seriously affect the whole clustering performance. (2) There is abundant information connotation involved in visual image. It's empirically realized that only using visual features is not sufficient for well formulating the distinguishability among image classes. The intensive visual feature expression can be utilized to further improve the multimodal correlation effectiveness and stableness. (3) There are different multimodal attributes among different annotated images. Although more obvious performance superiority has been exhibited via our CCA-based multimodal feature fusion, it's very beneficial to exploit an adaptive fusion between visual and semantic feature for each annotated image. (4) Some annotated images present an extreme vision with wrong or even without any valid annotations. With very limited useful annotation information and too much noises, it's hard for such images to successfully to recommend correct tags. This may be the stubbornest problem.

6 CONCLUSIONS

A new framework is implemented to exploit deep cross-modal correlations among deep visual and semantic features to enable more effective image tag recommendation. The in-depth multimodal feature analysis is established for characterizing the deep multimodal attributes for images and annotations. The DCCA-based cross-modal correlation mining is introduced to acquire the specified multimodal association expressions. Our future work will focus on making our system available online, so that more Internet users can benefit from our research.

Acknowledgments. This work is supported by the National Key Research and Development Plan (Grant No. 2016YFC0801003). Yuejie Zhang is the corresponding author.

7 REFERENCES

1. Venkatesh N. Murthy. 2015. Automatic image annotation using deep learning representations. University of Massachusetts, Amherst, MA, USA.
2. Wang W., Arora R., Livescu K., et al. 2015. Unsupervised learning of acoustic features via deep canonical correlation analysis[C]//Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 4590-4594.
3. Murthy V.N., Can E.F., Manmatha R. 2014. A hybrid model for automatic image annotation[C]//Proceedings of International Conference on Multimedia Retrieval. ACM.
4. Guillaumin M., Mensink T., Verbeek J., et al. 2009. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation[C]//Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 309-316.

5. Hardoon D.R., Szedmak S., Shawe-Taylor J. 2004. Canonical correlation analysis: An overview with application to learning methods[J]. Neural computation, 16(12): 2639-2664.
6. Jin C., Mao W., Zhang R., et al. 2015. Cross-Modal Image Clustering via Canonical Correlation Analysis[C]//Twenty-Ninth AAAI Conference on Artificial Intelligence.
7. Gong Y., Jia Y., Leung T., et al. 2013. Deep convolutional ranking for multilabel image annotation[J]. arXiv preprint arXiv:1312.4894.
8. Makadia A., Pavlovic V., Kumar S. 2008. A new baseline for image annotation[M]//Computer Vision–ECCV 2008. Springer Berlin Heidelberg, 316-329.
9. Andrew G., Arora R., Bilmes J., et al. 2013. Deep canonical correlation analysis[C]//Proceedings of the 30th International Conference on Machine Learning, 1247-1255.
10. Wang W., Arora R., Livescu K., et al. 2014. On deep multi-view representation learning[C]//Proceedings of the 32nd International Conference on Machine Learning (ICML-15), 1083-1092.
11. Andrew G., Arora R., Bilmes J., et al. 2013. Deep canonical correlation analysis[C]//Proceedings of the 30th International Conference on Machine Learning, 1247-1255.
12. Sigurbjörnsson B., Van Zwol R. 2008. Flickr tag recommendation based on collective knowledge[C]//Proceedings of the 17th international conference on World Wide Web. ACM, 327-336.
13. Murthy V.N., Can E.F., Manmatha R.A. 2014. A hybrid model for automatic image annotation[C]//Proceedings of International Conference on Multimedia Retrieval. ACM.
14. Krizhevsky A., Sutskever I., Hinton G.E. 2012. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems, 1097-1105.
15. Pennington J., Socher R., Manning C. 2014. Glove: Global vectors for word representation[C]// Conference on Empirical Methods in Natural Language Processing.
16. Thompson B. 2005. Canonical correlation analysis[J]. Encyclopedia of statistics in behavioral science.