Unsupervised Concept Learning in Text Subspace for Cross-media Retrieval

Mengdi Fan, Wenmin Wang*, Peilei Dong,
Ronggang Wang, and Ge Li

School of Electronic and Computer Engineering, Peking University
Lishui Road 2199, Nanshan District, Shenzhen, China 518055
fanmengdi@sz.pku.edu.cn, wangwm@ece.pku.edu.cn, pldong@pku.edu.cn,
rgwang@ece.pku.edu.cn, gli@pkusz.edu.cn

Abstract. Subspace (i.e. image, text or latent subspace) learning is one of the essential parts in cross-media retrieval. And most of the existing methods deal with mapping different modalities to the latent subspace pre-defined by category labels. However, the labels need a lot of manual annotation, and the label concerned subspace may not be exact enough to represent the semantic information. In this paper, we propose a novel unsupervised concept learning approach in text subspace for cross-media retrieval, which can map images and texts to a conceptual text subspace via the neural networks trained by self-learned concept labels, therefore the well-established text subspace is more reasonable and practicable than pre-defined latent subspace. Experiments demonstrate that our proposed method not only outperforms the state-of-the-art unsupervised methods but achieves better performance than several supervised methods on two benchmark datasets.

Keywords: unsupervised, concept learning, text subspace, cross-media retrieval, neural networks

1 Introduction

With the expansion of multimedia data on the Internet, cross-media retrieval, which focuses on mining the correlations between different modalities, is becoming increasingly important. It aims to align the feature spaces to bridge the heterogeneity gap across modalities. As a solution, a majority of the existing predominant methods utilize the pre-defined category information to explore the semantic correlation particularly on image-text mixed multimedia data. However, this practice may be problematic in two aspects: Firstly, it strongly bases a possibly invalid hypothesis that a query will fall into a pre-defined category in the process of retrieval; Secondly, it might well be the case that labeled data is not available due to the expensive annotation process. Thus the primary challenge of cross-media retrieval in practice consists in finding a common semantic

[*] Corresponding Author

subspace to eliminate the heterogeneity as well as getting rid of the dependency on pre-defined category information. From the above point of view, the recent progresses of cross-media retrieval can be categorized by two perspectives: 1) the works mapping into one of three subspaces, i.e. image subspace, text subspace and latent subspace; 2) the works using class labels or not, i.e. the supervised methods[1] which provide class labels for each modality, and the unsupervised methods[2] where only image-text pairs are used without class labels.

Many supervised methods for cross-media retrieval have been proposed in recent years. Wang *et al.* [22] unified coupled linear regression, L_{21} norm and trace norm regularization terms into a generic framework to jointly perform common subspace learning and coupled feature selection. Several years later, they have proposed a joint learning framework [21] which consists of subspace learning for different modalities, the L_{21} norm for coupled feature selection, and the multimodal graph regularization for preserving the inter-modality and intra-modality similarity. On the basis of [13], Wei *et al.* [23] proposed a deep semantic matching method to address the cross-modal retrieval problem with respect to samples annotated with one or multiple labels. Wang *et al.* [20] proposed a regularized deep neural network (RE-DNN) for semantic mapping across modalities and learned a joint model which can capture both intra-modal and inter-modal relationships. Peng *et al.* [11] proposed a cross-media multiple deep network (CMDN) to exploit the complex cross-media correlation by hierarchical learning. However, it is usually time-consuming and requires a lot of human annotation effort to manually label large-scale datasets.

Besides, a few unsupervised methods without using class labels are proposed. Canonical correlation analysis (CCA) [13] can learn a common latent subspace for different media types, which is able to maximize the correlation of them. Yan *et al.* [24] addressed the problem of matching images and captions in a joint latent space learnt with deep canonical correlation analysis (DCCA). The bilinear model (BLM) [19] can separate style and content by using singular value decomposition (SVD). It learns a shared subspace in which data of the same content and different modality is projected to the same coordinates. Partial least square (PLS) [14] uses the least square method to correlate the subspaces of CCA in order to avoid information dissipation in the process of different modal correlations. Corr-AE [5] is constructed by correlating hidden representations of two uni-modal autoencoders. Multimodal DBN [16] learns a joint representation of multimodal data, which models each media type with a separate two-layer DBN, and combines two networks by learning a joint RBM on top of them. Sun *et al.* [17] proposed an automatic visual concept discovery algorithm for image-sentence bi-directional retrieval. Liang *et al.* [8] proposed a self-paced cross-modal subspace matching method, where a joint feature learning and data grouping formulation has been developed.

In general, from the perspective of three subspaces, most of the cross-media retrieval methods are based on latent subspace. Some of them aim to align the

[1] https://en.wikipedia.org/wiki/Supervised_learning

[2] https://en.wikipedia.org/wiki/Unsupervised_learning

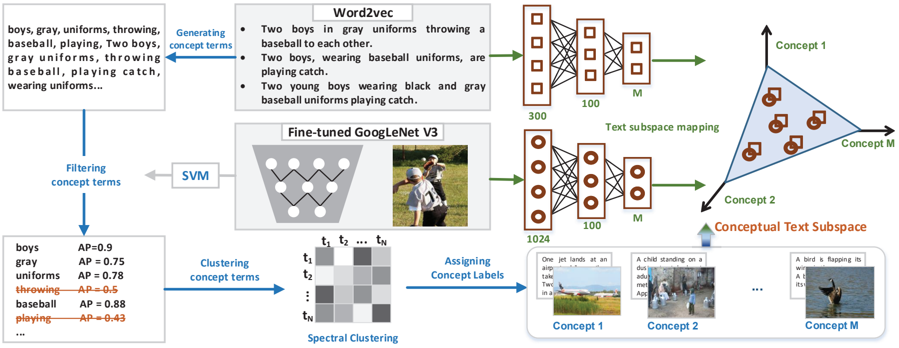

Fig. 1. Overview of our approach. The framework consists of four procedures: firstly generating concept terms, then filtering the terms without visual discrimination, thirdly clustering the terms into several compact concepts and lastly constructing two neural networks to map all texts and images into the self-learned conceptual text subspace.

images and texts directly from the low-level features in the latent subspace, such as CCA [13], KCCA [2], BLM [19] and PLS [14], but it's not the most effective place to find correlations, as the semantic gap is maximized [6]. Some of them first extract low-level features from each modality then the alignment is learned jointly by Corr-AE [5], Multimodal DBN [16] or CMDN [11]. However, when there is not enough labeled training data available, the deep networks might be overfitted due to their large number of parameters. Others explore the latent subspace with semantic associations heavily depending on pre-defined class labels, such as LCFS [22], JFSSL [21], CFMCM [4] and so on. More rarely, Dong *et al.* [3] proposed to embed the text into a visual feature space (image subspace) to capture both semantic and visual similarities.

Different from the above methods, we introduce a novel unsupervised concept learning approach in text subspace for cross-media retrieval, shown in Figure 1, which can pretrain the text set to get the semantic concepts of texts without using class labels, then cast all modalities into a conceptual text subspace trained by neural networks based on the aggregated concept labels. The contributions of this paper are as follows:

- We introduce the idea of unsupervised concept learning for cross-media retrieval, which can automatically generate the concept labels to cope with the problem of missing class labels in practical applications.
- According to the self-learned concept labels, we propose two 3-layer neural networks to map all images and texts respectively into an isomorphic text subspace, which is more reasonable and interpretable than other subspaces.
- The retrieval results on Pascal Sentence dataset outperform all of the unsupervised and supervised methods we compare, and better results are also obtained for Wikipedia dataset.

2 Our Approach

In this section, we describe the details of the proposed unsupervised concept learning in text subspace for cross-media retrieval.

2.1 Concept Terms Generating

Given a paralleled corpus of images and their descriptive texts, we firstly extract the semantic information from the text data. Considering a sentence "Two boys in gray uniforms throwing a baseball to each other", we can extract some key words or several dependencies between individual words. For example, "boys", "uniforms" and "baseball" belong to nouns, "throwing" belongs to verb, and "grey" belongs to adjective. "Grey uniforms" is an adjective modifier dependency, and "throwing baseball" is the direct objective dependency.

To collect the above candidate concepts, we use Stanford CoreNLP Parser [9] to extract the unigrams and dependencies of interest, shown in Table 1. After parsing the whole corpus, we count the number of occurrences of the entire terms, including unigrams and dependencies. In order to eliminate the influence of very rare terms, we remove the terms which occur fewer than k times.

Table 1. Unigrams and dependencies of interest

Unigrams	Dependencies
Noun: NN, NNS, NNP, NNPS	acomp, agent, amod
Verb: VB, VBD, VBG, VBN, VBP, VBZ	dboj, iobj, nsubj
Adjective: JJ, JJR, JJS, ADJP	nsubjpass, nummod
Adverb: RB, RBR, RBS	prt, vmod

2.2 Concept Terms Filtering

In order to guarantee the semantic consistency of images and texts, the concept terms have to be associated with visual features as much as possible. But the terms generated in Section 2.1 may not have visual discrimination or may not be distinguished by visual features easily, we decide to filter the terms with the help of corresponding image features pretrained on deep convolutional neural networks (CNN).

For the images associated with a certain term, we do a 2-fold cross validation with LIBSVM [1], and the negative samples of which are randomly selected from training set. We can get the average precision (AP) for each term, and remove the terms with AP lower than a threshold. The filtered and preserved terms are partially listed in Table 2.

From Table 2, we can see that a term may be filtered if it's not visually discriminative (the first line), not detailed enough (the second line) or too complicated (the third line). And some terms beginning with capital letters are also filtered out (the last line).

Table 2. The filtered and preserved terms from Pascal Sentence dataset

Filtered terms	Preserved terms
facing, group, several	glasses, flowers
yellow, brown	yellow bus, brown horse
jet, airport	plane, air
People, Woman	people, woman

2.3 Concept Clustering

Many of the remaining concept terms are synonyms or belong to various transformations. If they serve as different concepts, the classifier in concept space will be confused in the training process. Thus, it's important to cluster the terms with the same semantics. Word2vec [10] is a recently developed technique for building a neural network (NN) that maps words to real-number vectors, with the desideratum that words with similar meanings will map to similar vectors. The pre-trained 300-dimensional Word2vec vectors used to represent texts are obtained from `code.google.com/p/word2vec/`. We take the average of the word vectors from each word of the concept term, and $L2$-normalize the average vector. For each two concept terms t_i and t_j, the similarity can be computed by the cosine distance of their word embeddings:

$$S(i,j) = cosine(t_i, t_j) \tag{1}$$

For N concept terms $T = \{t_1, t_2, \cdots, t_N\}$, we use Spectral Clustering [25] to cluster the similarity matrix $S \in \mathbb{R}^{N \times N}$ into M concept groups $C = \{c_1, c_2, \cdots, c_M\}$. Each group is represented as a set of terms. Table 3 shows some of the clustered concepts discovered by our framework. We can find that the objects or scenes which often co-occur have been gathered together. Different transformations or patterns (the third line) are also merged. Besides, it can automatically generate the concepts related to positions, activities, colors and synonyms. We observe that there are some mixed type concepts (the last line) of activities (*plant*) and objects (*trees*). It's possibly because they often appear in one sentence.

Table 3. The clustered concepts from Pascal Sentence dataset

Type	Concept
Object	{bed, chair, coach, sofa}
Scene	{city, sidewalk, street}
Transform	{black cat, white cat, cats}
Position	{behind, back, up, down}
Activity	{laying, looking, running}
Color	{blue, green, orange, red}
Synonyms	{bicycle, bike}, {motorbike, motorcycle}
Mixed	{plant, trees, flowers, grass, wood}

2.4 Text Subspace Mapping

Each clustered concept is served as one dimension in text subspace. And what we need to solve is how to map all the images and texts into this subspace under the circumstances of preserving the isomorphic semantic information as much as possible.

The previous work of [7] shows the efficiency of "Semantic Matching", which tried to learn a shallow linear classifier with a probabilistic interpretation to produce a probability distribution over classes. Differently, we learns the 3-layer neural networks of several non-linear transformations to produce a probability distribution over our pre-defined concept labels as the semantic embedding.

For the i^{th} pair of image and text, the most relevant concept $l_i \in \mathbb{R}^M$ is defined as:

$$l_i = \underset{j}{\operatorname{argmax}} \underset{c_j \in C}{cosine}(w_i, c_j) \tag{2}$$

where w_i is the average Word2vec vectors of the words in the i^{th} text.

According to the self-learned concept label l, we can train a 3-layer fully-connected neural networks for images and texts separately. We take the image input as an example. $x_i^I \in \mathbb{R}^{D_I}$ represents the deep CNN features of the i^{th} image, which is the input to the network. The output y_i, also called semantic embedding, is calculated as:

$$y_i = Softmax(W_2 \cdot \sigma(W_1 x_i^I + b_1) + b_2) \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function, W_1 and W_2 are the weights to be learned. Actually, the output of $Softmax$ produces a probability distribution over M concepts, the meaning of which is essentially the same as Semantic Matching but with more non-linear deep correlation.

We define the cost function as:

$$loss = \frac{1}{2n} \sum_{i=1}^{n} \|y_i - l_i\|^2 + \frac{\lambda}{2}(\|W_1\|^2 + \|W_2\|^2) \tag{4}$$

where n is the training number, and λ is the regularizer penalty parameter. Equation 4 is minimized by using stochastic gradient descent (SGD).

We can train a neural network similarly for texts given a fixed training set $\{(x_1^T, l_1), \cdots, (x_n^T, l_n)\}$, where $x_i^T \in \mathbb{R}^{D_T}$ equals to w_i of the i^{th} text features. After casting all of the heterogeneous image and text features into this conceptual text subspace, we can easily perform *Img2Text* and *Text2Img* evaluated by the traditional similarity measurement, such as Centered Correlation (CC).

3 Experiments

In this section, we evaluate the performance of our method by conducting extensive experiments and compare with the existing unsupervised and supervised approaches on two benchmark datasets.

3.1 Datasets

1) *Pascal Sentence dataset*[3] [12]: Pascal Sentence dataset is a subset of Pascal VOC, which contains 1,000 pairs of an image along with several sentences from 20 categories. The whole dataset is randomly split into a training set and a testing set with 600 and 400 pairs.

2) *Wikipedia dataset*[4] [13]: Wikipedia dataset, generated from Wikipedia featured articles, is composed of 2,866 image-text pairs from 10 semantic classes. A random split was used to produce a training set of 2,173 documents, and a testing set of 693 documents.

3.2 Evaluation Scheme & Baseline Methods

We use mean average precision (mAP) as the performance measurement. For a query q and a set of R retrieved target documents, the average precision (AP) is defined as:

$$AP(q) = \frac{1}{L_q} \sum_{j=1}^{R} P_q(j)\delta_q(j) \tag{5}$$

where L_q is the number of ground truth neighbors in the retrieved set, $P_q(j)$ is the precision of the top j retrieved results and $\delta_q(j)$ is an indicator function equaling 1 if the item at rank j is relevant, 0 otherwise.

With regards to baseline methods, CCA [13], Multimodal DBN [16], Corr-AE [5], SCSM [8] are four unsupervised methods, which merely use the pair-wised multimedia information to learn a common latent subspace. On the contrary, supervised methods such as LCFS [22], Deep-SM [23], JFSSL [21], CMDN [11] all take the category information into account, and they are also compared here.

3.3 Experimental Results

The parameter k in Section 2.1 is set to 15 for Pascal Sentence dataset and 20 for Wikipedia dataset. The threshold of AP in Section 2.2 is set to 0.7 for Pascal Sentence dataset and 0.6 for Wikipedia dataset. Figure 2 shows the curves of mAP when selecting different numbers of concept groups. We fix M to 100 for Pascal Sentence dataset and 20 for Wikipedia dataset to achieve the best resluts in our experiments. In view of [23], the visual features obtained from the pretrained or fine-tuned CNN model should be the primary candidate for cross-media retrieval. So our deep CNN features extracted from the 1024×1 final layer of Inception-v3 model [18] are fined-tuned by the images from two target datasets in our experiments. For the 3-layer fully-connected image and text neural networks in Section 2.4, the node number of hidden layer is set to 100, the learning rate is set as 0.01, the momentum equals to 0.9 and the epoch number is fixed at 2,000. We have attempted to change the layer number or node number but perform not as good as these two 3-layer networks.

[3] http://vision.cs.uiuc.edu/pascal-sentences/

[4] http://www.svcl.ucsd.edu/projects/crossmodal/

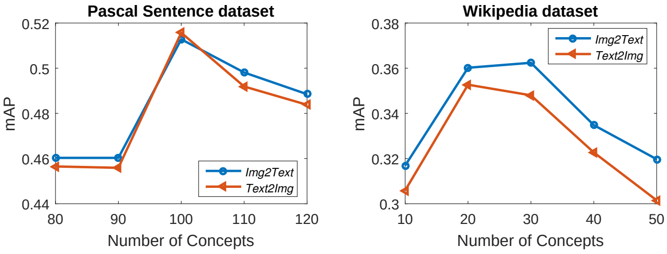

Fig. 2. Curves of the mAP values under different clustering numbers of semantic concept.

Table 4. The mAP scores on Pascal Sentence dataset and Wikipedia dataset

Methods	Pascal Sentence dataset			Wikipedia dataset		
	Img2Text	*Text2Img*	*Average*	*Img2Text*	*Text2Img*	*Average*
CCA [13][23]	0.364	0.387	0.376	0.251	0.199	0.225
Corr-AE [5]	0.290	0.279	0.285	0.335	0.368	0.352
Multimodal DBN [16][11]	0.149	0.150	0.150	0.197	0.183	0.190
SCSM [8]	-	-	-	0.274	0.217	0.245
LCFS [22]	0.497	0.488	0.493	0.280	0.214	0.247
Deep-SM [23]	0.446	0.478	0.462	0.398	0.354	**0.376**
JFSSL [21]	-	-	-	0.306	0.228	0.267
CMDN [11]	0.334	0.333	0.334	0.393	0.325	0.359
Ours	**0.513**	**0.516**	**0.515**	**0.362**	**0.353**	**0.359**

Table 4 shows the mAP scores of different approaches on Pascal Sentence and Wikipedia dataset. On the whole, the supervised methods achieve higher scores than unsupervised methods. This is because supervised methods rely on semantic labels to reduce the semantic gap of different modalities, but unsupervised methods only use pair-wised information. For Pascal Sentence dataset, we can observe that the proposed method achieves the best performance with average mAP 0.515 among unsupervised methods even supervised methods. For Wikipedia dataset, our method obtains the best scores among unsupervised methods. Even though our method performs a little lower than Deep-SM [23], it still can beat other supervised methods, such as LCFS [22] and JFSSL [21].

Figure 3 shows two examples of *Text2Img* by our proposed approach. Row 1 and 2 correspond to the samples from Pascal Sentence dataset and Wikipedia dataset respectively. The first column contains the text queries and the second represents the paired images of the text queries. Column 3∼7 are the top 5 retrieved images.

The excellent performance of Pascal Sentence dataset may derive from two aspects: Firstly, the concepts of the sentences are relatively simple, and from which we can exploit rich semantic information. Secondly, the 20 classes are all included in ImageNet and the images of Pascal Sentence dataset are very similar as those from ILSVRC 2012 [15]. In contrast, the texts in Wikipedia dataset are

Airplane on runway in front of buildings. An aircraft sitting on the runway. A white airplane is on a runway. Front image of a parked silver airplane. Plane on runway with buildings in the background.

In 1943, the Spanish Army and government decided to replace much of its armor. Tanks such as the Panzer I and T-26 were obsolete, as compared to newer tanks such as the Soviet T-34 and German Panther tank.

Fig. 3. Two examples of *Text2Img* by our method for Pascal Sentence dataset and Wikipedia dataset respectively.

hard to parse and there are huge gaps between the latent meaning of the texts and the visual perception of the images. Despite the improved performance of some supervised methods, semantic labels are usually too expensive and time-consuming to obtain. Thus, it's of great significance to improve the effectiveness of unsupervised approaches for cross-media retrieval.

4 Conclusion

In this paper, we innovatively conduct cross-media retrieval through unsupervised concept learning in text subspace. The procedures of concept terms generating, filtering and clustering aim to automatically obtain concept labels, which lay a good foundation for text subspace mapping. And then, the isomorphic text subspace is trained by two well-designed 3-layer fully-connected neural networks individually for images and texts. Our method can successfully cope with the problem of missing class labels in practical applications. The retrieval results show that the proposed approach not only surpasses the unsupervised methods, but outperforms several supervised methods, achieving the state-of-the-art performance. In the future work, we will further explore the semantic correlation between the images and texts by unsupervised learning.

Acknowledgement. This project was supported by Shenzhen Peacock Plan (20130408-183003656), Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (ZDSYS201703031405467), and Guangdong Science and Technology Project (2014B010117007).

References

1. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (2007)
2. Costa, P.J., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G.R., Levy, R., Vasconcelos, N.: On the role of correlation and abstraction in cross-modal multimedia retrieval. TPAMI (2014)
3. Dong, J., Li, X., Snoek, C.G.M.: Word2visualvec: Cross-media retrieval by visual feature prediction. In: arXiv

4. Fan, M., Wang, W., Wang, R.: Coupled feature mapping and correlation mining for cross-media retrieval. In: ICME Workshop (2016)
5. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: ACM MM (2014)
6. Habibian, A., Mensink, T., Snoek, C.G.M.: Discovering semantic vocabularies for cross-media retrieval. In: ACM ICMR (2015)
7. Han, L., Wang, W., Fan, M., Wang, R.: Cross-modality matching based on fisher vector with neural word embeddings and deep image features. In: ICASSP (2017)
8. Liang, J., Li, Z., Cao, D., He, R.: Self-paced cross-modal subspace matching. In: ACM SIGIR (2016)
9. Marneffe, M.C.D., Maccartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. LREC (2006)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR (2013)
11. Peng, Y., Huang, X., Qi, J.: Cross-media shared representation by hierarchical learning with multiple deep networks. In: IJCAI (2016)
12. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using amazon's mechanical turk. In: NAACL Workshop (2010)
13. Rasiwasia, N., Costa Pereira, J., Coviello, E., et al.: A new approach to cross-modal multimedia retrieval. In: ACM MM (2010)
14. Rosipal, R., Kramer, N.: Overview and recent advances in partial least squares. In: International Conference on Subspace, Latent Structure and Feature Selection (2006)
15. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV (2015)
16. Srivastava, N., Salakhutdinov, R.: Learning representations for multimodal data with deep belief nets. In: ICML Workshop (2012)
17. Sun, C., Gan, C., Nevatia, R.: Automatic concept discovery from parallel text and visual corpora. In: ICCV (2015)
18. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR (2015)
19. Tenenbaum, J.B., Freeman, W.T.: Separating style and content with bilinear models. Neural Computation (2000)
20. Wang, C., Yang, H., Meinel, C.: Deep semantic mapping for cross-modal retrieval. In: ICTAI (2015)
21. Wang, K., He, R., Wang, L., Wang, W., Tan, T.: Joint feature selection and subspace learning for cross-modal retrieval. TPAMI (2016)
22. Wang, K., He, R., Wang, W., Wang, L., Tan, T.: Learning coupled feature spaces for cross-modal matching. In: ICCV (2013)
23. Wei, Y., Zhao, Y., Lu, C., Wei, S.: Cross-modal retrieval with cnn visual features: A new baseline. IEEE Transactions on Cybernetics (2016)
24. Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: CVPR (2015)
25. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. In: NIPS (2004)