Large Scale Image Annotation via Deep Representation Learning and Tag Embedding Learning

Yonghao He, Jian Wang, Cuicui Kang, Shiming Xiang, Chunhong Pan
{yhhe, jian.wang, cckang, smxiang, chpan}@nlpr.ia.ac.cn

ABSTRACT

In this paper, we focus on the issue of large scale image annotation, whereas most existing methods are devised for small datasets. A novel model based on deep representation learning and tag embedding learning is proposed. Specifically, the proposed model learns an unified latent space for image visual features and tag embeddings simultaneously. Furthermore, a metric matrix is introduced to estimate the relevance scores between images and tags. Finally, an objective function modeling triplet relationships (irrelevant tag, image, relevant tag) is proposed with maximum margin pursuit. The proposed model is easy to tackle new images and tags via online learning and has a relatively low test computation complexity. Experimental results on NUS-WIDE dataset demonstrate the effectiveness of the proposed model.

Keywords

Large Scale Image Annotation; Deep Representation Learning; Tag Embedding Learning

1. INTRODUCTION

In the era of big data, there are many image related applications, such as keyword based image search and picture recommendation, in need of image annotation. Plentiful tags, as high level semantic features, are assigned to images to definitely improve the quality of these applications. In this paper, we address the issue of large scale image annotation, namely a large number of images with tags engaged. Many existing methods [2, 19, 14, 10, 3, 15, 1] for image annotation are established on small datasets, such as Corel5K [6], IAPRTC-12 [9] and ESP-game [18]. These datasets have only around 5000 to 20000 images, and these methods are difficult to be applied to large datasets. By contrast, we focus on designing an algorithm that can handle a large number of images.

There exist some methods [20, 8] for large scale image annotation (we omit the review of those methods for small scale datasets). Weston *et al.* [20] proposed a method (called WSABIE) to optimize top k tag lists for images. In this method, tag embeddings are initialized with random variables in a low-dimensional space. Then a projection matrix is used to transform the image features into the tag embedding space. In such space, the relevance scores between images and tags are measured by inner product. Finally, the objective is to minimize weighted approximate-rank pairwise (WARP) loss. The model parameters are tag embeddings and projection matrix that can be learned recursively via online updating. Recently, in [8], Gong *et al.* used convolutional neural network (CNN) [12] to map raw image data to tag indicator space. Three categories of loss functions for CNN are studied, namely cross-entropy loss with softmax, pairwise hinge loss and WARP. The results show that ranking based CNN (pairwise hinge loss and WARP) is better than classification based CNN (cross-entropy loss with softmax). One of the advantages of ranking based CNN is that the test computation complexity is a constant. However, when new tags appear, the number of output layer nodes has to be increased and the network needs to be retrained. Besides these two methods, kNN and probability SVM [8] are feasible to solve this problem. In kNN based method, tag votes are collected through top k nearest neighbors. This straightforward method can achieve state-of-the-art performance, which will be presented in experiment section. Whereas, it is time-consuming to find nearest neighbors in large datasets. The probability SVM based method is to train many one-vs-all classifiers for each tag. The drawbacks of this method is that tags have no mutual interactions and the classifiers may not be sufficiently trained due to sample imbalance.

In this paper, we propose a novel model based on **D**eep **R**epresentation learning and tag **E**mbedding **L**earning (**DR EL**). First, images and tags are represented as visual features and embedding vectors, respectively. Tag embedding vectors are created for each tag and randomly initialized. They will be updated in the learning process. The merit of tag embeddings is to make the direct interactions between tags and images possible in the unified latent space. Then, two deep neural networks (DNN) are used to do feature learning while mapping image features and tag embeddings from their original spaces to an unified latent space. In the latent space, the relevance scores between images and tags are estimated by a metric matrix. Finally, a pairwise hinge loss for triplets (irrelevant tag, image, relevant tag) are adopted with maximum margin pursuit. The model parameters, including weights of two DNNs, the metric matrix and tag embeddings, are learned through stochastic gradient descent

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

ICMR '15, June 23 - 26, 2015, Shanghai, China
Copyright 2015 ACM 978-1-4503-3274-3/15/06 ...$15.00
http://dx.doi.org/10.1145/2671188.2749330.

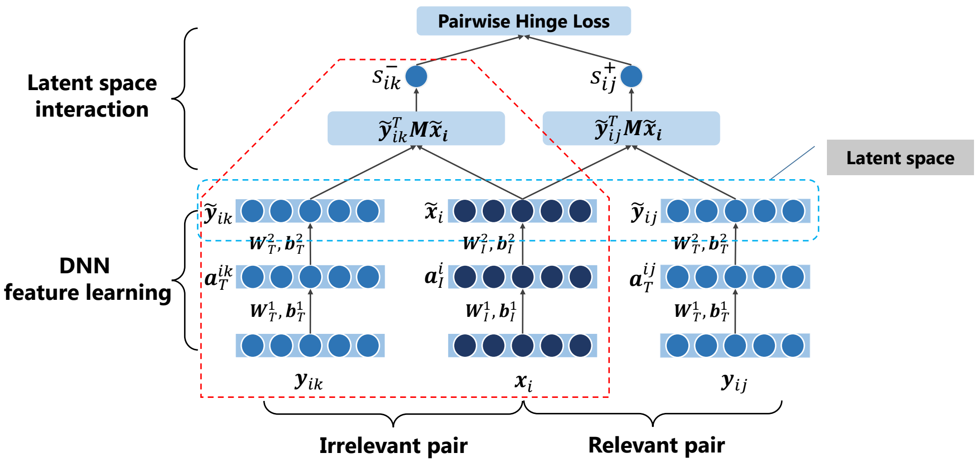

Figure 1: Architecture of the proposed model. The input is a triplet (irrelevant tag, image, relevant tag). Through two specific DNNs, the image features and tag embeddings are nonlinearly mapped to the latent space. In the latent space, we use a metric matrix to calculate the relevance scores between images and tags. The irrelevant and relevant image-tag pairs are put to the pairwise hinge loss based objective function to optimize the model parameter. The process of test phase is highlighted in the red dashed box.

and backpropagation. The proposed model is easy to handle new images and tags via online learning. Actually, the model in [20] can be viewed as a "shallow" type of ours. Both methods have a relatively low test computation complexity of $O(T)$ (T is the number of tags). The experimental results demonstrate the effectiveness of the proposed model compared to state-of-the-art methods.

2. THE PROPOSED MODEL

We first introduce some notations: the training set contains N images $\{\mathbf{x}_i\}_{i=1}^{N}$ ($\mathbf{x}_i$ are visual feature vectors) and T tags $\{\mathbf{y}_i\}_{i=1}^{T}$ ($\mathbf{y}_i$ are randomly initialized embedding vectors). For each image $\mathbf{x}_i$, its associated relevant tag set is $\mathcal{T}_{i+} = \{\mathbf{y}_{ij}\}_{j=1}^{j=n_i}$ and irrelevant tag set is $\mathcal{T}_{i-} = \{\mathbf{y}_{ik}\}_{k=1}^{k=T-n_i}$, n_i is the number of relevant tags.

The architecture of the proposed model is illustrated in Fig. 1. Generally, the model consists of two parts: the DNN feature learning and latent space interaction. In the DNN feature learning, image features and tag embeddings are mapped from the original spaces to the latent space. For image features, the transformation is formulated as:

$$\mathbf{a}_I^i = f(\mathbf{W}_I^1 \mathbf{x}_i + \mathbf{b}_I^1), \tag{1}$$

$$\tilde{\mathbf{x}}_i = f(\mathbf{W}_I^2 \mathbf{a}_I^i + \mathbf{b}_I^2), \tag{2}$$

where $\mathbf{W}_I^1, \mathbf{W}_I^2$ and $\mathbf{b}_I^1, \mathbf{b}_I^2$ are weights and bias respectively, $f(\cdot)$ is sigmoid function and $\tilde{\mathbf{x}}_i$ are representations of images in the latent space (the dimensions of these variables are presented in Section 3.2&3.5). Similarly, $\tilde{\mathbf{y}}_i$ are representations of tags in the latent space. There are two hidden layers for both DNNs, and the second hidden layer is for the latent space (highlighted in blue dashed box in Fig. 1). The reasons of this nonlinear transformation process is two-fold: 1) the resulting latent space makes images and tags interact easily for relevance score estimation; 2) this process is the feature learning directly driven by the model objective.

In Fig. 1, images and tags are organized into irrelevant pairs (the left part of Fig. 1) and relevant pairs (the right part of Fig. 1). A metric matrix $\mathbf{M}$ is employed to compute the relevance scores of these pairs:

$$s_{ik}^{-} = \tilde{\mathbf{y}}_{ik}^{T} \mathbf{M} \tilde{\mathbf{x}}_i, \tag{3}$$

$$s_{ij}^{+} = \tilde{\mathbf{y}}_{ij}^{T} \mathbf{M} \tilde{\mathbf{x}}_i, \tag{4}$$

where s_{ik}^{-} and s_{ij}^{+} are relevance scores for irrelevant and relevant pairs (the larger, the more relevant). The matrix $\mathbf{M}$ can capture more complex relationships than just inner product used in [20], that is, the interactions of each dimension between $\tilde{\mathbf{y}}_{ij}$ ($\tilde{\mathbf{y}}_{ik}$) and $\tilde{\mathbf{x}}_i$ can be adjusted by matrix $\mathbf{M}$.

The objective of the our model is to increase relevance scores of relevant pairs, while decreasing relevance scores of irrelevant pairs. Here, we adopt pairwise hinge loss:

$$\max(0, m + s_{ik}^{-} - s_{ij}^{+}), \tag{5}$$

where m is used to control margin size (m is set to 1 in experiments). Hinge loss is used in a ranking objective with the notion of large margin. In this case, it makes relevant pairs rank ahead of irrelevant pairs. The final objective function is formulated as:

$$\min_{\boldsymbol{\theta}} \frac{1}{T_N} \sum_{i=1}^{N} \sum_{\mathbf{y}_{ij} \in \mathcal{T}_{i+}} \sum_{\mathbf{y}_{ik} \in \mathcal{T}_{i-}} \max(0, m + s_{ik}^{-} - s_{ij}^{+}), \tag{6}$$

where $\boldsymbol{\theta}$ is the model parameter and T_N is the number of triplets generated according to N training images with their tags. Since the deep model has many parameters to be learned, a large number of samples need to be trained to prevent overfitting. The form of triplets can expand N training images to much more training triplets[1]. Parameters involved in the proposed model are DNN weights ($\mathbf{W}_I^1, \mathbf{b}_I^1, \mathbf{W}_I^2, \mathbf{b}_I^2, \mathbf{W}_T^1, \mathbf{b}_T^1, \mathbf{W}_T^2, \mathbf{b}_T^2$), the metric matrix $\mathbf{M}$

[1] For example, there are 100 images and 10 tags. Each image has 2 tags on average. The number of triplets is 1600 ($C_{100}^1 \times C_2^1 \times C_8^1$).

and tag embeddings ($\mathbf{y}_i$). Error backpropagation and stochastic gradient descent are used to optimize the proposed model. In the test phase, relevance scores between the test image and all tags are calculated through the forward pass of the structure (highlighted in red dashed box in Fig. 1), and tags with highest scores are taken as the final annotations.

Furthermore, the proposed model can handle new training images with new tags. The new data can be processed by the same way in which the original training data is processed. This online updating only fine-tunes the model parameters without changing the model structure, which is not feasible for method in [8].

Test computation complexity. According to the description of the proposed model, the final annotations are determined by the relevance scores between the test image and all tags. Therefore, test computation complexity of the proposed model is $O(T)$. Similarly, the text complexity of method in [20] and probability SVM [17] based method is also $O(T)$. For kNN based method, the test complexity will increase to $O(N)$ because the distances between the test image and all training images need to be calculated (in practice, $N \gg T$). The method in [8] only requires a single forward pass of CNN for testing, resulting in a computation complexity $O(C)$, C is a constant.

3. EXPERIMENTS

3.1 Dataset

The dataset used in the experiment is NUS-WIDE [4], which is the largest publicly released multilabel dataset. In this dataset, 269,648 images with 81 different concepts[2] are collected from Flickr. Since the compared method [8] uses CNN, it needs raw images as inputs that are not provided instead of source URLs. However, some URLs are no longer valid and some images have no tags, the final set contains 113,290 images. We randomly divide it into the training set (103,290 images) and the test set (10,000 images).

3.2 Features

We conduct two groups of experiments with low level visual features and off-the-shelf CNN features (it is widely accepted to use learned CNN models as feature extractors [16, 7]), respectively. The low level features include 64-dimensional color histogram, 144-dimensional color correlogram, 73-dimensional edge direction histogram, 128-dimensional wavelet texture, 255-dimensional block-wise color moments and 500-dimensional bag of words based on SIFT descriptor [13]. As for CNN features, we use the model trained on Imagenet [5], provided by Caffe[3] [11]. The dimension of CNN feature is 4096. Additionally, the method in [8] with raw images as inputs is compared with other methods using off-the-shelf CNN features

3.3 Compared Methods

The methods listed below are taken for comparison:

- **WSABIE** [20]. This method projects images features to tag embedding space, and then relevance scores are calculated using inner product.
- $\mathbf{CNN}_{softmax}$, $\mathbf{CNN}_{hingeloss}$, $\mathbf{CNN}_{WARP}$ [8]. CNN based methods must take raw images as inputs, and we implement them using toolkit MatConvNet[4]. Furthermore, the parameters of these CNN models are initialized with the same model that is used to extract off-the-shelf CNN features.

Table 1: Annotation results with low level features.

Method	mP	mR	F1	N_+
kNN	0.1896	0.2216	0.2043	74
SVM	0.0878	0.2358	0.1279	78
WASBIE	0.1083	0.3025	0.1595	81
DREL	0.1614	0.3193	0.2144	81

Table 2: Annotation results with CNN features.

Method	mP	mR	F1	N_+
kNN	0.2566	0.3826	0.3072	81
SVM	0.1147	0.4192	0.1801	81
WASBIE	0.1667	0.4677	0.2458	81
$\mathbf{CNN}_{softmax}$	0.1903	0.4896	0.2741	81
$\mathbf{CNN}_{hingeloss}$	0.2063	0.4718	0.2871	81
$\mathbf{CNN}_{WARP}$	0.1996	0.4637	0.2790	78
DREL	0.2295	0.4223	0.2973	81

- **SVM**. For each tag, a probabilistic SVM [17] is trained. The tags with highest probabilities are determined as the final annotations.
- **kNN**. The kNN based method follows the same manner as in [8]. The k is set to 100 for both experiments.

3.4 Evaluation Protocol

We follow the conventional protocol in [14] to evaluate all methods, namely mean recall (mR) over all tags, mean precision (mP) over all tags, F1 ($F1 = \frac{2*mP*mR}{mP+mR}$) score and N_+ (the number of tags that have non-zero recall). The top 5 tags are regarded as the final annotations for each image.

3.5 Setting of the Proposed Structure

Since we take two groups of experiments using low level features and CNN features respectively, two different groups of DNNs are used. Specifically, for low level features, the number of nodes for each layer in image DNN and tag DNN are 1164-1024-1024 and 256-512-1024. For CNN features, the structure is changed to 4096-2048-1024 for image DNN and 256-512-1024 for tag DNN.

3.6 Experimental Results

The results of two groups of experiments are presented in Table 1 and Table 2, respectively.

From Table 1 and Table 2, we can make some observations. In low level feature based results, our method and kNN can achieve state-of-the-art performances with high F1 scores. kNN trends toward a higher mean precision but relatively a lower mean recall and N_+. WASBIE and our method can obtain higher mean recall and full N_+, but mean precision of WASBIE is lower than ours partially due to its "shallow" structure. SVM performs poor because it isolates all tags from each other.

[2] Following [8], we directly regard concepts as tags.

[3] http://caffe.berkeleyvision.org/

[4] http://www.vlfeat.org/matconvnet/

In Table 2, CNN feature based results are shown. It can be observed that the overall performances are much higher than those using low level features. Actually, off-the-shelf CNN features always result in high performance [16]. kNN, three CNN based methods and our method can achieve state-of-the-art performances. All methods obtain full N_+ except for CNN_{WARP}. kNN still shows the same trend of higher mean precision and lower mean recall, but the highest F1 score. Three CNN based methods have higher mean recall, in particular, $CNN_{softmax}$ gain additional 10% mean recall compared to kNN. Our method outperforms WASBIE in mean precision and F1, again demonstrating the superiority of the deep structure.

In summary, first, CNN features are much better than low level features for the task of image annotation. Second, our method and kNN can achieve state-of-the-art performances, whereas the proposed method has much lower test computation complexity. Third, the proposed method can significantly outperform WASBIE, which demonstrates the effectiveness of deep representation learning. Finally, CNN based methods are effective. However, they have to be retrained with new images and tags, which is not necessary in our method.

4. CONCLUSIONS

In this paper, a novel model is proposed for large scale image annotation via deep representation learning and tag embedding learning. Image features and randomly initialized tag embeddings are refined and mapped to the latent space through two deep neural networks. In such latent space, relevance scores between images and tags are calculated through a metric matrix. Finally, a pairwise hinge loss based objective function is adopted to optimize the model parameters with notion of large margin. The proposed model can accept new images and tags for training via online learning without modifying the structure. Furthermore, the proposed method has low test computation complexity of $O(T)$. The comparative results with state-of-the-art methods on NUS-WIDE dataset demonstrate the effectiveness of our method.

5. ACKNOWLEDGMENTS

This work was supported in part by the National Basic Research Program of China under Grant 2012CB316304, the National Natural Science Foundation of China under Grants 61331018, and the Beijing Nature Science Foundation under Grant 4142057.

6. REFERENCES

[1] L. Ballan, T. Uricchio, L. Seidenari, and A. Del Bimbo. A cross-media model for automatic image annotation. In *ACM International Conference on Multimedia Retrieval*, pages 73–80, 2014.

[2] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.

[3] M. Chen, A. Zheng, and K. Weinberger. Fast image tagging. In *International Conference on Machine Learning*, pages 1274–1282, 2013.

[4] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: a real-world web image database from national university of singapore. In *ACM International Conference on Image and Video Retrieval*, page 48, 2009.

[5] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: a large-scale hierarchical image database. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision*, pages 97–112. 2006.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013.

[8] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. 2014.

[9] M. Grubinger, P. Clough, H. Muller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop OntoImage*, pages 13–23, 2006.

[10] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: discriminative metric learning in nearest neighbor models for image auto-annotation. In *IEEE International Conference on Computer Vision*, pages 309–316, 2009.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, pages 675–678, 2014.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[14] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *European Conference on Computer Vision*, pages 316–329. 2008.

[15] V. Murthy, E. Can, and R. Manmatha. A hybrid model for automatic image annotation. In *ACM International Conference on Multimedia Retrieval*, pages 369–376, 2014.

[16] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *arXiv preprint arXiv:1403.6382*, 2014.

[17] V. Vapnik and V. Vapnik. *Statistical learning theory*, volume 2. Wiley New York, 1998.

[18] L. Von Ahn and L. Dabbish. Labeling images with a computer game. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 319–326, 2004.

[19] C. Wang, S. Yan, L. Zhang, and H. Zhang. Multi-label sparse coding for automatic image annotation. In *IEEE International Conference on Computer Vision*

and Pattern Recognition, pages 1643–1650, 2009.

[20] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine Learning*, 81(1):21–35, 2010.