TAG-AWARE IMAGE CLASSIFICATION VIA NESTED DEEP BELIEF NETS

Zhaoquan Yuan[1,2], *Jitao Sang*[1,2], *Changsheng Xu*[1,2]

[1]National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China
[2]China-Singapore Institute of Digital Media, Singapore
{zqyuan, jtsang, csxu}@nlpr.ia.ac.cn

ABSTRACT

With the rising of internet photos-sharing web sites, the rich aware text information surrounding images on the sites are proved helpful to improve the image classification. This paper presents a novel nested deep learning model called Nested Deep Belief Network(NDBN) for tag-aware image classification. A multi-layer structure of Deep Belief Network(DBN) is established to learn a unified representation of visual feature and tag feature for an image, and an additional Gaussian Restricted Boltzmann Machine is built to capture the tag-tag dependency. Compared with conventional methods, the proposed model can not only find correlations across modalities, but mine the importance for different tags, and also bring about low-rank tag feature representation. We conduct experiments over the MIR Flickr dataset and the results show that the proposed NDBN model outperforms the existing image classification techniques.

***Index Terms*—** Deep belief network, image classification, deep learning, singular value decomposition

1. INTRODUCTION

Image classification is the most important part of digital image content analysis. Many efforts have been made about it. e.g., SVM [1], sparse representation [2], Boosting [3]. However, in real world application, image classification is still a challenging task because of the semantic gap.

On the other hand, with the dramatic growth of the internet photo sharing web, we could have millions of internet images which are surrounded with rich text information, such as text, tag, etc. These information are indicative of the image content and provide a more direct gateway to image analysis. Actually, some related research works [4] [5] [6] [7] [8], etc. have proved that these text information consistently improve the performance on image classification problems. These methods could be roughly divided into three categories: classifier-level methods, latent space methods and feature-level methods. The difference among these categories lies in the fusion level.

Classifier-level methods for tag-aware image classification train separate classifiers for text and visual feature respectively, and then correlate the visual information and text information in the classifying phase. Based on the motivation that images with similar surrounding text are similar in visual feature space, [4] builds the text feature for images by finding the K nearest neighbor images which are similar to the target image in visual. Two separate classifiers are trained, and the then combined for the final prediction. In [6], author assumes that the text is not available in test dataset and only a training subset is labeled. A strong Multiple Kernel Classifier is learned using the text and visual feature to aid the final visual classification performance.

Latent space methods assume that although the raw text and image are in different feature spaces, but there could be a latent semantic space, where text and visual features with same semantic meaning have same statistical property. A representative work is [7], in which a latent space bridges the text and image visual feature. Through the bridge, a cross modal transfer learning model is built for image classification.

Feature-level methods focus on learning an unified representation from multimodal feature. For the sake of the goal, deep networks which have been successful applied to feature learning are usually built. [9] uses deep sparse Restricted Boltzmann Machine to learn sharing representation for multimodal feature. A closely related work [10] introduces a multimodal data representation learning model with Deep Belief Network.

However, there are still some difficulties in image classification even with text information. Firstly, image low-level visual features and the associated text features are belong to different modal spaces respectively, and each modal is characterized by different statistical properties. Models in the first two kinds of methods belong to shallow architectures and have limited representation capabilities [11]. Hence, they can not fuse the multimodal data well. Secondly, the text information usually is very noisy and ambiguous. Taking the tags for example, some tags are meaningless and even indicate error information for the associated image content. Thirdly, the dimension of the whole tag/text feature space is high. If the raw tag features are used directly as inputs of a model, the training process will become very difficulty. In addition, for each image, the associated tags have different indicative intensity for the image content, and these intensity can not be reflected by the word counts, etc. simple feature.

Our work builds on three insights. First, deep learning with the strong feature learning ability provides us a powerful framework to learn a more discriminative feature representation. Based on Deep Belief Nets [12], we learn an unified representation from multimodal feature. Second, considering that there are some degree of dependence between tags of an image, and these dependence could constitute a dependence spectrum for the image, an additive Restricted Boltzmann Machine in this paper is built to capture these tag-tag dependence. Third, in order to handle the high dimension problem of tag space and mine the tag indicative intensity, we use SVD algorithm for image-tag relation matrix, which bring out a brief but informative low-rank representation for tag feature.

The remainder of this paper is organized as follows. section 2 introduces a popular deep learning model which is the background of our work. section 3 describes our proposed Nested Deep Belief Network(NDBN) model. Experiments on the MIR Flickr dataset are presented in Section 4. section 5 summarizes the conclusion and mentions future work.

2. BACKGROUND: DEEP BELIEF NETS

Recently, deep learning is successfully applied to multiple areas due to its strong feature learning ability. Among these models, Deep Belief Nets [12] is of special concern. For our work is based on this architecture, we give an brief introduction to it in this section. DBN uses multiple non-linear layers to learn semantic feature. The learning processes include layer-wise pre-training and a following fine-tuning stage.

The greedy layer-wise pre-training is the phase of constructing the deep architecture based on Restricted Boltzmann Machine [13]. RBM is an undirected graphical model with a hidden layer and another visible layer shown as left plot in Fig. 1. The two layers are connected by symmetric weights, but there are no intra-layer connection. The energy of the joint configuration is give by [14]:

$$E\left(\mathbf{v},\mathbf{h}\right) = -\sum_i a_i v_i - \sum_j b_i h_j - \sum_{ij} \nu_i h_j W_{ij} \qquad (1)$$

where v_i and h_j are the binary states of units in $\mathbf{v} \in \{0,1\}^D$ and hidden units in $\mathbf{h} \in \{0,1\}^F$, and a_i and b_j are their biases. W_{ij} denotes the symmetric weights. The probability of visible vector $\mathbf{v}$ can be computed using the energy function:

$$P\left(\mathbf{v}\right) = \sum_h \frac{exp\left(-E\left(v,h\right)\right)}{\sum_{u,g} exp\left(-E\left(u,g\right)\right)} \qquad (2)$$

The conditional distribution of hidden units with value 1 given the states of visible units is:

$$Q\left(h_j = 1|\mathbf{v}\right) = \sigma\left(b_j + \sum_i \nu_i W_{ij}\right) \qquad (3)$$

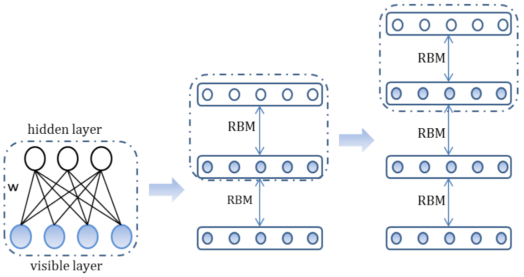

Fig. 1: Layer-wise pre-training process.

where $\sigma\left(x\right)$ denotes the logistic function. Similarly, the probability on visible units given the hidden units:

$$P\left(v_i = 1|\mathbf{h}\right) = \sigma\left(a_i + \sum_j h_j W_{ij}\right) \qquad (4)$$

There is no closed solution for the parameters of RBM, but they can be obtained by alternating Gibbs sampling. For simplicity, a contrastive divergence [15] is carried out, which is an approximate version of Gibbs sampling:

$$\triangle w_{ij} = \epsilon(< v_i h_j >_{data} - < v_i h_j >_{recon}) \qquad (5)$$

where ϵ is the learning rate. $< . >_{data}$ and $< . >_{recon}$ are the expectation with respect to the data distribution and the expectation with respect to the reconstruction distribution after running k steps of Gibbs sampling. The biases parameters could be updated in a similar way.

Multiple layers deep network is built in a bottom-up fashion. Each pair of two adjacent layers can be regarded as a RBM by taking the lower layer as visible layer **v** and the upper layer as hidden layer **h**. Fig. 1 shows the layer-wise pre-training process.

After having greedily pre-trained the deep multiple layers, a up-down algorithm is used to adjust the parameter of all the layer globally in the fine-tuning stage, which is a contrastive form of “wake-sleep” algorithm [16].

3. NESTED DEEP BELIEF NETWORK

In this section, we describe the details of our proposed Nested Deep Belief Network(NDBN) model. We first overview the architecture of the NDBN model, and then show how to apply the singular value decomposition technique to learn tag feature for images. Next, the two training stages of NDBN model will be introduced. Finally, we summary the algorithm.

3.1. Model Overview

Let **x** be the visual feature vector of an image, and matrix $\mathbf{C} \in \mathcal{R}^{N \times M}$ are the image-tag matrix. N and M is the number of images and tags in tag dictionary. Let **y** be the label vector

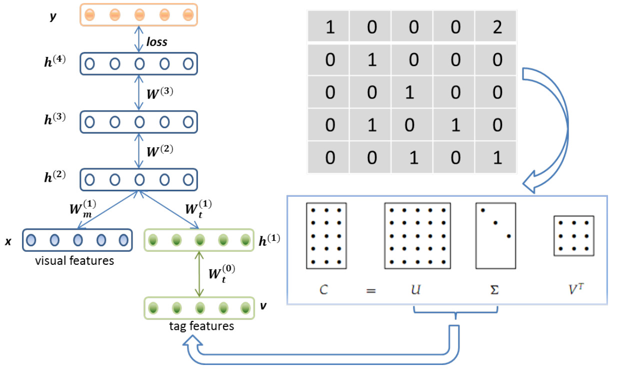

Fig. 2: Nested Deep Belief Network(NDBN) model framework.

corresponding the image with visual feature $\mathbf{x}$, which is the form as follows:

$$\mathbf{y} = (y_1, y_2, \ldots, y_K) \tag{6}$$

where K is the number of total classes, and y_k is binary such that if $\mathbf{x}$ belong to the kth class, then y_k is 1, otherwise 0. Each image may belong to multiple classes.

Based on these data above, what we want to do is to learn a model that can achieve the tag-aware image classification purpose. To address the problem, we propose a novel Nested Deep Belief Network model. Fig. 2 show the architecture of NDBN. Multiple layers is built to fuse the visual feature and tag feature. In order to capture the tag-tag dependence, an additional Gaussian RBM is contained for tag feature, which forms a nested architecture. To handle the high dimension and mine the tag indicative intensity, a SVD [17] algorithm is used to the image-tag matrix for getting tag feature.

3.2. SVD for Image-tag Matrix Factorization

In linear algebra, the singular value decomposition (SVD) is a factorization of a real or complex matrix, with many useful applications in relational data modeling and statistics, which maps the high dimension feature into the low dimension feature space through discovering the principle components of the data. Here, we factorize the image-tag matrix $\mathbf{C}$ to get the singular value matrix Σ.

$$\mathbf{C} = \mathbf{U}\Sigma\mathbf{V}^T \tag{7}$$

where $\mathbf{U} \in \mathcal{R}^{N\times N}$ and $\mathbf{V} \in \mathcal{R}^{M\times M}$ are unitary matrixes. The entries of matrix Σ reflect the indicative degrees of the tags for the image content. A low dimension singular value matrix $\Sigma^{'} \in \mathcal{R}^{D\times D}$ is got by discarding the part of smallest singular values, which is the approximation of Σ. It contains only the largest D singular values. Under the new singular value matrix, low dimensional tag features are computed by:

$$\mathbf{T} = \mathbf{U}\Sigma^{'} \tag{8}$$

where $\mathbf{T} \in \mathcal{R}^{N\times D}$ and ith row vector $\mathbf{v}_i$ of $\mathbf{T}$ can be treated as a latent semi-semantic representation of the ith image, which remove the noise of text information. Through the SVD factorization process, new tag feature set $\mathbf{T}$ reflect the indicative intensity of tags, and also bring out low-rank representaion. It could enhance for our multimodal feature fusion in deep architecture.

3.3. Nested Deep Architecture for Classification

After discovering the low-rank tag feature, a deep multiple layers architecture is constructed to fuse the visual feature and tag feature. However, before that, an additive RBM is established to capture tag-tag dependence explicitly in our model. In this case, the input feature $\mathbf{v}$ is real-value rather than binary, so the Gaussian RBM [18] is used to model them. The energy of the states in visible layer $\mathbf{v}$, and hidden layer $\mathbf{h}^{(1)}$ is defined as follows:

$$E\left(\mathbf{v}, \mathbf{h}^{(1)}\right) = \textstyle\sum_{i=1}^{D} \frac{\left(\nu_i - \mu_i\right)^2}{2\sigma_i^2} \\ - \sum_{i=1}^{D}\sum_{j=1}^{F} \frac{\nu_i}{\sigma_i} W_{ij}^{(0)} h_j^{(1)} - \sum_{j=1}^{F} a_j h_j^{(1)} \tag{9}$$

where, σ_i is standard deviation with predetermined value and μ_i denotes the expectation of the ith unit in the visible layer $\mathbf{v}$. This lead to conditional probabilities on visible units given the hidden layer:

$$P\left(\nu_i|\mathbf{h}^{(1)}\right) = \mathcal{N}\left(b_i + \sigma_i \sum_{j=1}^{F} W_{ij}^{(0)} h_j^{(1)}, \sigma_i^2\right) \tag{10}$$

The corresponding conditional distribution on hidden layer given the visible layer is:

$$Q\left(h_i^{(1)} = 1|\mathbf{v}\right) = \sigma(\mu_j + \sum_i W_{ij}^{(0)} \frac{v_i}{\sigma_i}) \tag{11}$$

Learning of the parameters in this work is carried out using one-step Contrastive Divergence [15].

The additive Gaussian RBM together with the latter deep network forms a nested architecture. we argue that it is necessary for it could make the compact tag-tag relation be obtained and the tag feature more discriminative.

In order to fuse the visual feature $\mathbf{x}$ and tag feature, a similar deep belief net is built with a different RBM which we called Multimodal RBM where visual feature layer $\mathbf{x}$ and layer $\mathbf{h}^{(1)}$ are regarded as visible layer together. Considering visual feature units $\mathbf{x} \in \{0,1\}^G$, units in the layer $\mathbf{h}^{(1)} \in \{0,1\}^D$ and as hidden layer in Multimodal RBM $\mathbf{h}^{(2)} \in \{0,1\}^F$, the joint energy configuration is defined as follows:

$$E\left(\mathbf{x},\mathbf{h}^{(1)},\mathbf{h}^{(2)}\right) = -\sum_{i=1}^{D}\sum_{j=1}^{F} h_i^{(1)} W_{\mathtt{t}ij}^{(1)} h_j^{(2)} - \sum_{g=1}^{G}\sum_{j=1}^{F} x_g \cdot W_{\mathtt{m}ij}^{(1)} h_j^{(2)} - \sum_{i=1}^{D} a_{\mathtt{t}i}^{(1)} v_i - \sum_{j=1}^{F} a_{\mathtt{t}j}^{(2)} h_j^{(2)} - \sum_{g=1}^{G} a_{\mathtt{m}g}^{(1)} x_g \quad (12)$$

Exact inference likelihood learning in this model is intractable. However, an analogous Contrastive Divergence approximation algorithm [15] could be applied.

Based on the Multimodal RBM, we also use the greedy layer-wise pretraining strategy to construct the deep network and initial the model parameters simultaneously by learning a RBM at a time.

In order to make the classification task more precision, we use a discriminative fine-tuning method to adjust the parameter globally so as to find the local optimal solution. This goal is achieved by minimizing the empirical risk on the training data, and the optimization problem can be formulated as:

$$\arg_{\Theta} \min L\left(\mathbf{Y},\widehat{\mathbf{Y}},\Theta\right) \quad (13)$$

where Θ is the parameters of the whole network, and the $L\left(\mathbf{Y},\widehat{\mathbf{Y}},\Theta\right)$ is the loss function for measuring the training error on the training data, Y is the groundtruth labels of the training samples, and $\widehat{Y}$ is the predicted labels from our model. we use the exponential loss function as optimization criteria:

$$L\left(\mathbf{Y},\widehat{\mathbf{Y}},\Theta\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} exp\left(l\left(y_{nk},\widehat{y}_{nk}\right)\right) \quad (14)$$

where, N denotes the number of the training samples and K represents the image classes. If $y_{nk} = \widehat{y}_{nk}$, we set $l\left(y_{nk},\widehat{y}_{nk}\right)$ as 1, otherwise 0.

Combined the SVD algorithm introduced in section 3.2, the proposed Nested Deep Belief Nets model is summarized in Algorithm 1.

4. EXPERIMENTS

4.1. Dataset and Experiment Setup

In order to evaluate the effectiveness of our proposed NDBN model, we conduct a series experiments on a public MIRFLICKR-25000 collection dataset [19]. The data set consist of $25,000$ annotated images which are collected from Flickr along with their tags. It includes 24 labeled categories among which 14 classes were stricter labeled. Therefore, there are 38 classes in total. Each image may belongs to multiple classes. Fig. 3 shows some sample images from the dataset.

```
Procudure 1: NESTED DEEP BELIEF NETS
Input: Image-tag matrix C; visual feature dataset of
    images X; Corresponding labels set Y;
    Number of network layers L; Number of
    training samples N; Random initial bias
    parameters
    $\mathbf{a} = \{a_t^{(0)}, a_t^{(1)}, a_m^{(1)}, a^{(2)}, \ldots, a^{(L)}\}$; Random
    initial weight parameters
    $\mathbf{W} = \{W_t^{(0)}, W_t^{(1)}, W_m^{(1)}, W^{(2)}, \ldots, W^{(L-1)}\}$;
    Dimension of tag feature $D$.
Output: Optimal parameter space $\widehat{\Theta} = \left[\widehat{\mathbf{W}}, \widehat{\mathbf{a}}\right]$
SVD for matrix C: $\mathbf{C} = \mathbf{U}\Sigma\mathbf{V}^T$;
$\Sigma'$ approximates $\Sigma$ ;
$\mathbf{T} = \mathbf{U}\Sigma'$;
Train Gaussian RBM consdiering $\mathbf{v_i}$ as visible layer
  and $\mathbf{h}^{(1)}$ as hidden layer;
Train Multimodal RBM consdiering $\mathbf{x}_i$ and $\mathbf{h}^{(1)}$ as
  visible layer, $\mathbf{h}^{(2)}$ as hidden layer;
for $each\ layer\ from\ \boldsymbol{h}^{(2)}\ to\ \boldsymbol{h}^{(L)}$ do
    Greedy layer-wise pretraining by learning a RBM
    at a time;
Return optimal parameter $\widehat{\Theta} = \arg_{\Theta} \min L\left(\mathbf{Y},\widehat{\mathbf{Y}},\Theta\right)$;
```

In our experiment, we use randomly selected $15,000$ images for pretraining, $5,000$ for fine-tuning and $5,000$ for test. Gray values are extracted as visual feature for each image and they are represented by $1024(32 \times 32)$ dimensional vectors. We set the dimension of the approximation singular value matrix Σ' to 200. In GRBM, the numbers of units in both layers are 200 and in visual feature layer are $1,024$. Mean Average Precision(MAP) is served as performance metric and the results are averaged over 10 trials.

4.2. Experimental Results and Analysis

Table 1 presents the AP scores for the our comparison with Support Vector Machines(SVMs) and Multimodal DBN(MDBN) [10]. For the SVM model, we use the concatenated visual and tag feature as input feature, and use $20,000$ images for training and $5,000$ for testing. Table 1 shows that proposed NDBN model outperforms the SVM and MDBN models.

For the sake of studying the influence of the components SVD and Gaussian RBM(GRBM), we also conduct the experiment with SVD and DBN but no Gaussian RBM(SVD+DBN). Meanwhile, we discard the SVD step and replace the tag feature in our model with the tag counts(GRBM+DBN), where we just consider the 2000 most

Table 1: Experiment Results

LABELS	ANIMALS	BABY	BABY*	BIRD	BIRD*	CAR	CAR*	CLOUDS	CLOUDS*	DOG
SVM	0.477	0.171	0.157	0.249	0.223	0.341	0.311	0.627	0.573	0.261
MDBN	**0.498**	0.129	0.134	0.184	0.255	0.309	0.354	0.759	**0.691**	0.342
SVD+DBN	0.481	0.135	0.153	0.219	0.238	0.285	0.332	**0.762**	0.674	0.438
GRBM+DBN	0.483	0.147	0.166	0.207	0.264	0.293	0.305	0.705	0.662	**0.451**
NDBN	0.474	**0.268**	**0.269**	0.191	0.234	**0.378**	0.317	0.724	0.684	0.447
LABELS	DOG*	FEMALE	FEMALE*	FLOWER	FLOWER*	FOOD	INDOOR	LAKE	MALE	MALE*
SVM	0.219	0.459	0.405	0.524	0.586	0.341	0.485	0.475	0.411	0.385
MDBN	0.376	0.540	0.478	**0.593**	**0.679**	**0.447**	**0.750**	0.262	0.503	0.406
SVD+DBN	0.465	0.582	0.567	0.568	0.663	0.325	0.691	0.326	**0.571**	0.479
GRBM+DBN	0.445	0.627	0.583	0.556	0.647	0.351	0.678	0.314	0.552	0.492
NDBN	0.490	**0.646**	**0.603**	0.531	0.606	0.349	0.632	**0.365**	0.570	**0.509**
LABELS	NIGHT	NIGHT*	PEOPLE	PEOPLE*	PLANT-LIFE	PORTRAIT	PORTRAIT*	RIVER	RIVER*	SEA
SVM	0.558	0.493	0.458	0.436	0.531	0.417	0.408	0.407	0.069	0.538
MDBN	**0.655**	0.483	0.800	0.730	**0.791**	**0.642**	**0.635**	0.263	0.110	**0.586**
SVD+DBN	0.621	0.516	0.819	**0.752**	0.613	0.627	0.622	0.342	0.246	0.576
GRBM+DBN	0.637	0.540	0.764	0.725	0.665	0.641	0.596	0.354	**0.257**	0.574
NDBN	0.638	**0.559**	**0.826**	0.718	0.679	0.632	0.613	**0.272**	0.152	0.537
LABELS	SEA*	SKY	STRUCTURES	SUNSET	TRANSPORT	TREE	TREE*	WATER	**MEAN**	
SVM	0.073	0.609	0.488	0.672	0.418	0.495	0.217	0.532	0.407	
MDBN	0.259	0.873	0.787	**0.648**	0.406	**0.660**	0.483	**0.629**	0.503	
SVD+DBN	0.263	0.801	0.774	0.632	0.437	0.624	0.492	0.593	0.508	
GRBM+DBN	**0.271**	0.816	0.733	0.642	**0.481**	0.651	0.497	0.608	0.510	
NDBN	0.188	**0.829**	**0.791**	0.635	0.456	0.602	**0.503**	0.615	**0.514**	

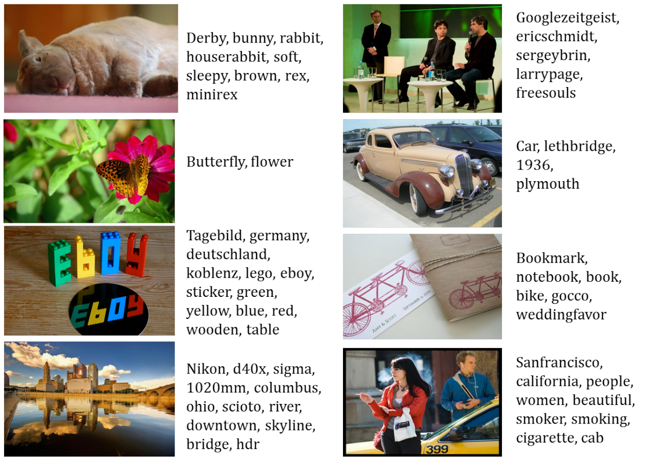
Derby, bunny, rabbit, houserabbit, soft, sleepy, brown, rex, minirex
Googlezeitgeist, ericschmidt, sergeybrin, larrypage, freesouls
Butterfly, flower
Car, lethbridge, 1936, plymouth
Tagebild, germany, deutschland, koblenz, lego, eboy, sticker, green, yellow, blue, red, wooden, table
Bookmark, notebook, book, bike, gocco, weddingfavor
Nikon, d40x, sigma, 1020mm, columbus, ohio, scioto, river, downtown, skyline, bridge, hdr
Sanfrancisco, california, people, women, beautiful, smoker, smoking, cigarette, cab

Fig. 3: Sample images from MIRFLICKR-25000 dataset.

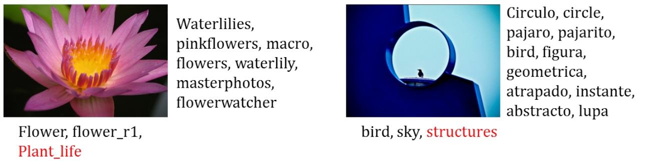
Waterlilies, pinkflowers, macro, flowers, waterlily, masterphotos, flowerwatcher
Flower, flower_r1, Plant_life
Circulo, circle, pajaro, pajarito, bird, figura, geometrica, atrapado, instante, abstracto, lupa
bird, sky, structures

Fig. 4: Example images which are classified incorrectly by SVM, while classified correctly by NDBN model

frequent tags. In the experiments of SVD+DBN and GRBM+DBN, the training and tuning methods are the same as our proposed NDBN model. From the results showed in table 1, both the SVD+DBN and GRBM+DBN outperform SVM and MDBN models slightly.

For illustrating the advantages of the proposed NDBN mode demonstrably, we give a brief analysis about some cases in the experiments. Fig. 4 show two images as well as the associated tags on the right side of the pictures, which are classified incorrectly by SVM, while classified correctly by NDBN model. The groundtruth labels are provided below the pictures. For the left image, the SVM model classifies the label "flower"and "flower_r1" correctly, but misclassifies the label "Plant_life". We think that "Plant_life" is a more abstract concept than "flower", and it needs more training samples to cover the concept, while in the tags set for all samples, the "flower"and "Plant_life"usually appear together. The SVD step in our NDBN model could capture this type of correlation and corresponding singular value of "Plant_life"is bigger than the other tag such as "macro"although it does not appear in the raw tags.

As for the right image case, it is misclassified in the SVD+DBN experiment as well while correctly by NDBN model. We believe that it owes to the Gaussian RBM as well as the deep architecture. Although the tags "lupa", "geometrica", and "figura" are seldomly arise along with the tag "structures", they constitute a specific dependence spectrum captured by the Gaussian RBM. This kind of dependence spectrum implies some semantic concept and the semantic "structures" is learned through the following deep architecture. By the way, some other tags which are hardly indicative of the image content will be filtered by the SVD algorithm, which can be explained the small corresponding singular values.

5. CONCLUSION

In this paper, a Nested Deep Belief Network model is proposed to solve the tag-aware image classification problem. In this framework, we use SVD to educe low-rank tag feature representation which reflect the indicative intensity of different tags. Obtained the brief but informative tag feature, tag dependence spectrum is captured by a Gaussian RBM. we generalize a Multimodal RBM together with deep network to fuse the visual feature and text feature. A series of experiments are conducted to MIR Flickr dataset.

Our future work include: 1) considering the fact that images with similar content exist similar tags, image-image relationship is more mined. 2) modeling topic distributions of tag space which can reflect more semantic content for images.

6. ACKNOWLEDGMENTS

This work is supported in part by National Basic Research Program of China (No. 2012CB316304), National Natural Science Foundation of China (No. 61225009), and Beijing Natural Science Foundation (No. 4131004). This work is also supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office.

7. REFERENCES

[1] Ankita Kumar and Cristian Sminchisescu, “Support Kernel Machines for Object Recognition,” in *International Conference on Computer Vision*, 2007, pp. 1–8.

[2] Bingkun Bao, Guangyu Zhu, Jialie Shen, and Shuicheng Yan, “Robust image analysis with sparse representation on quantized visual features,” 2013.

[3] Andreas Opelt, Michael Fussenegger, Axel Pinz, and Peter Auer, “Weak hypotheses and boosting for generic object detection and recognition,” *European Conference on Computer Vision*, vol. 2, pp. 71–84, 2004.

[4] Gang Wang, Derek Hoiem, and David Forsyth, “Building text features for object image classification,” in *Computer Vision and Pattern Recognition*, 2009, pp. 1367–1374.

[5] Yunpeng Li, David J. Crandall, and Daniel P. Huttenlocher, “Landmark classification in large-scale image collections,” in *International Conference on Computer Vision*, 2009, pp. 1957–1964.

[6] Matthieu Guillaumin, Jakob J. Verbeek, and Cordelia Schmid, “Multimodal semi-supervised learning for image classification,” in *Computer Vision and Pattern Recognition*, 2010, pp. 902–909.

[7] Wolfram Burgard and Dan Roth, Eds., *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*. AAAI Press, 2011.

[8] Shuhui Wang, Shuqiang Jiang, Qingming Huang, and Qi Tian, “Multi-feature metric learning with knowledge transfer among semantics and social tagging,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2240–2247.

[9] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng, “Multimodal deep learning,” in *ICML*, 2011, pp. 689–696.

[10] Nitish Srivastava and Ruslan Salakhutdinov, “Learning Representation for Multimodal Data with Deep Belief Nets,” in *International Conference on Machine Learning Workshop*, 2012.

[11] Yoshua Bengio, “Learning deep architectures for ai,” *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[12] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh, “A Fast Learning Algorithm for Deep Belief Nets,” *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[13] P Smolensky, *Information processing in dynamical systems: Foundations of harmony theory*, vol. 1, pp. 194–281, MIT Press, 1986.

[14] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, pp. 504–507, 2006.

[15] Geoffrey E. Hinton, “Training Products of Experts by Minimizing Contrastive Divergence,” *Neural Computation*, vol. 14, pp. 1771–1800, 2002.

[16] G E Hinton, P Dayan, B J Frey, and R M Neal, “The wake-sleep algorithm for unsupervised neural networks.,” *Science*, vol. 268, no. 5214, pp. 1158–1161, 1995.

[17] Scott C. Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard A. Harshman, “Indexing by Latent Semantic Analysis,” *Journal of The American Society for Information Science and Technology*, vol. 41, pp. 391–407, 1990.

[18] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle, “Greedy layer-wise training of deep networks,” *Processing*, vol. 19, no. d, pp. 153, 2007.

[19] Mark J Huiskes and Michael S Lew, “The mir flickr retrieval evaluation,” *Proceeding of the 1st ACM international conference on Multimedia information retrieval MIR 08*, vol. 4, no. November, pp. 39, 2008.