Learning Features from Large-Scale, Noisy and Social Image-Tag Collection

Hanwang Zhang[†], Xindi Shang[†], Huanbo Luan[†§], Yang Yang[‡], Tat-Seng Chua[†]
[†]National University of Singapore
[‡]University of Electronic Science and Technology of China
[§]Tsinghua University
{hanwangzhang,xindi1992,luanhuanbo,dlyyang}@gmail.com;dcscts@nus.edu.sg

ABSTRACT

Feature representation for multimedia content is the key to the progress of many fundamental multimedia tasks. Although recent advances in deep feature learning offer a promising route towards these tasks, they are limited in application to domains where high-quality and large-scale training data are hard to obtain. In this paper, we propose a novel deep feature learning paradigm based on large, noisy and social image-tag collections, which can be acquired from the inexhaustible social multimedia content on the Web. Instead of learning features from high-quality image-label supervision, we propose to learn from the image-word semantic relations, in a way of seeking a unified image-word embedding space, where the pairwise feature similarities preserve the semantic relations in the original image-word pairs. We offer an easy-to-use implementation for the proposed paradigm, which is fast and compatible for integrating into any state-of-the-art deep architectures. Experiments on NUSWIDE benchmark demonstrate that the features learned by our method significantly outperforms other state-of-the-art ones.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Model

Keywords

feature learning; visual-semantic embedding; multimodal analysis

1. INTRODUCTION

The progress in multimedia applications is largely due to the advances of feature representations for multimedia content. For example, over the past decades, we have witnessed the evolution of visual features from color histogram to SIFT interest points and to the recent deep learning features, that help to move a large varieties of multimedia applications from academic prototypes into industrial products [14]. Today, a general consensus is that learning-based features by deep neural networks can outperform most hand engineered features and therefore free us to focus on designing algorithms and end applications.

In order to learn strong features, we need a large-scale and high-quality dataset. At this point, ImageNet with millions of human-labeled images across thousands of semantic categories has offered us a reliable incubator to develop features. However, this dataset is built by Web images five years ago and hence it lags behind the fast evolving semantic and visual diversities in real-world scenarios. For example, different emerging vertical domains like fashion (*e.g.*, Taobao and Amazon) would need different datasets in order to learn specific features for shoes and clothes domain. Moreover, videos from emerging popular social networks (*e.g.*, Snapchat and Vine) would love features different from what were learned from images. Building such datasets not only requires heavy and tedious labeling efforts but also expert domain knowledge, any of which is expensive. This awkward situation

(a) Samples from ImageNet and SBU

(b) Feature Visualization

Figure 1: t-SNE visualization for the features of a million SBU images from Flickr. Different colors represent different semantic categories. One of the distribution is learned by deep learning a million images across 1,000 categories from high-quality labeled ImageNet and the other is learned by a million images weakly-labeled by over 30,000 unique tags by SBU (cf. Section 3 for details). Both sides show that the features are representative. Can you tell which side is learned by SBU? (see answers below)

[1]Left: ImageNet. Right: SBU.

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806286.

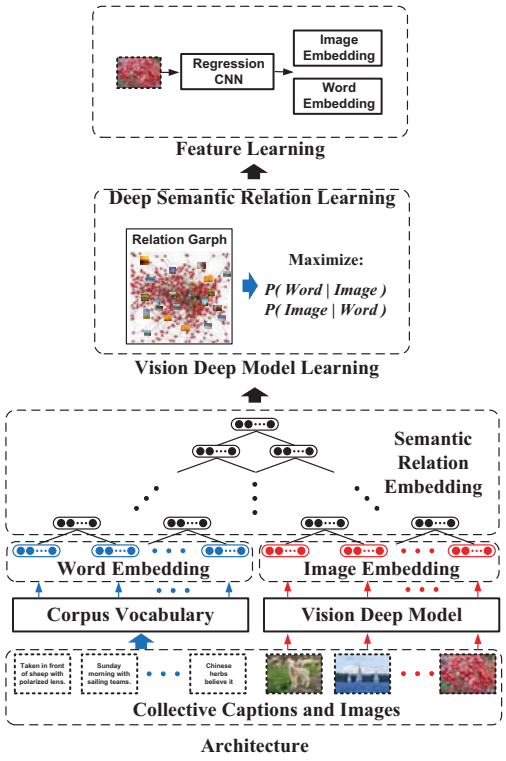

Figure 2: The overview of the proposed feature learning paradigm. Architecture: a hierarchical model that learns a unified vector space for words and images, which embeds the semantic relations. Feature Learning: After learning the embedded space by *Deep Semantic Relation Learning*, we learn a deep feature map from raw image pixels to the desired embedding vector by *Regression CNN*, which composes our *Vision Deep Model Learning*.

highlights the weakness of current feature learning methods: even though it requires no human expertise on feature engineering, it still needs a great labor intensity in building a good quality dataset as payoff.

In this paper, we propose a novel feature learning paradigm based on large, noisy and social image-tag collections which are ubiquitously accessible to us (*e.g.*, SBU in Figure 1(a)). The overview of our proposed feature learning paradigm is shown in Figure 2. The architecture is designed for mapping both images and words into the same embedding space of semantic relations, which are discovered from the large amount of social images and captions. The off-line feature learning includes two core components. The first is a deep relational learning method to learn the semantic relation embedding from the heterogeneous relation graph of images and words, which are noisy and sparse. In particular, we break down the topology of the graph into a tree-structured deep hierarchy, where the leaves are images and words. Each non-leaf node encodes the information about the semantic relations (*i.e.*, word-image, image-image and word-word) in terms of a feature vector. In order to generalize the learned collective intelligence to the content of images, the second component maps visual information into the semantic embedding. Our vision deep learning regresses images to the image embedding (by Regression CNN). So, our on-line feature extraction is simply to apply the learned deep vision model on an input image.

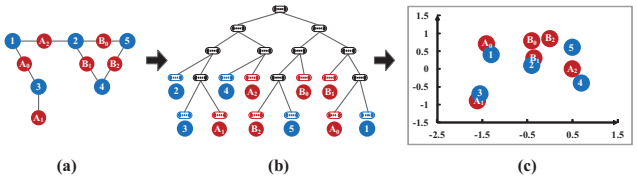

Figure 3: We want to transform the image-word graph into a continuous vector space. We take the graph (left) as the input to the tree-structured deep architecture (middle). It then learns the image and word feature representations (right) as the output. Here is a simple illustration on a toy graph: gray ones are words and red ones are items. We can see that the similarities between the learned features perserve the orginal graph topology.

We use a million-scale social media of Flickr images and captions: SBU dataset, as our source of collective intelligence. On the social image benchmark: NUSWIDE images, we conduct extensive experiments to show the effectiveness of the proposed feature learning paradigm in generating superior features. We believe that our work has a great potential in relieving human labor for feature learning.

2. RELATED WORK

Although deep learning features have been shown to outperform hand-engineered features on many vision benchmarks [3], there is limited research on how to further finetune the features for specific domains, especially for the social multimedia in the wild. One of the most probable reason might be the sparsity and noise in social media, which makes the problem very challenging. Zhang *et al.* [15] developed a multimodal autoencoder to discover the common patters of images and text and the binary classifier responses of the discovered patterns were considered as the semantic featuers. Gong *et al.* [6] used a large weakly annotated photo collection from Flickr (similar to SBU used in our work) to learn a generic image-sentence embeddings. The above methods lack a principled way to tackle the challenge of social media but generally resort to the deep architecture to discover the semantic relations between the noisy images and text. However, we specially propose to explicitly learn the inherent semantic relations embedded in social multimedia. This learning method is inspired by recent advances in social network learning [12] and natural language modeling [10] and has been shown to be robust to the sparse, diverse and noisy nature of data. In this way, our work is related to a latest work by Fang *et al.* [4], where they applied a matrix factorization method to learn the embeddings of images and users from a fashion collection. However, their method requires a reasonable measure for "negative" relations, which is not a generic solution for general social collections.

3. DEEP SEMANTIC RELATION LEARNING

Inspired by the recent advance in network topology learning [12], we first construct an image-word graph for social image-description collection and then introduce the deep architecture to learn the representations. We denote images

as $\mathcal{I}$ and words as $\mathcal{W}$. So, a image-word relation graph is denoted as $\mathcal{G} = \{\mathcal{I}, \mathcal{W}, \mathcal{E}\}$, where $\mathcal{E} \subseteq \mathcal{U} \times \mathcal{I}$ is the set of edges that connect images and words. We define that image i connects to word w if and only if w is a word appeared in the description of i. Our goal is to learn vector representations for the symbolic vertices in the graph. We denote $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{x}_w \in \mathbb{R}^d$ as vectors in a d-dimensional space. In particular, we want the similarities between vectors $s_{iw} = \mathbf{x}_i^T \mathbf{x}_w$ to preserve the topology of the graph. That is, we expect $s_{iw} > s_{iw'}$ if i is connected to w but not w'. Also, we expect the intra-image/word similarities to have similar properties such that $s_{i_j i_k} > s_{i_j i_l}$ (or $s_{w_j w_k} > s_{w_j w_l}$) if i_j and i_k (or w_j and w_k) share more words (or images) than i_j and w_l (or w_j and w_l). Figure 3 illustrates a toy example of such properties. In order to capture the topology information of the graph, we explore a series of short random walks in the graph. Rooted from each vertex v (v can be i or w), we randomly choose K-length list $\{v_1, ..., v_K\}$ ($K = 5$ in our paper), where v_{k+1} is a vertex chosen at random from the neighbors of vertex v_k.

Specifically, they statistically preserve the frequencies of image-word co-occurrences, that is, the number of optimizations for frequent connections will be larger than that of rare ones. Therefore, the learned vectors will reveal the desired semantic relations. As a result, our learning objective is to maximize the similarities between the vertices within a random walk. A widely used objective function is the softmax that models the probability of vertices i and j within a random walk: $p(i|j) = exp\left(\mathbf{x}_i^T \mathbf{x}_j\right) / \sum_{i'} exp\left(\mathbf{x}_{i'}^T \mathbf{x}_j\right)$, where we discard the distinct notations of image and word vertices for simplicity. We deploy the Hierarchical Softmax [9] to implement the probability as in [12] for efficiency and robustness.

4. VISUAL DEEP LEARNING FOR IMAGE-WORD EMBEDDING

Recall that our goal is to learn image features from collective intelligence, here, the discovered image-word embeddings. Although the embeddings have successfully gained semantic relations, visual information of images are not taken into account. In this section, we first use deep vision model to transform images into the embeddings and then exploit the visual information to fine-tune them. As a result, we expect to obtain image features that are close to their semantically related words.

Suppose the core visual model (*i.e.*, usually a convolution neural network pretrained by ILSVRC dataset), with its prediction layer removed, is denoted as $f(i)$, where i is the input image. For example, if we use AlexNet, $f(i)$ will be a 4,096-D sparse nonnegative vector which is the ReLU ouput of the 7-th fully-connected layer. We want to map $f(i)$ into the same embedding space as learned above. In particular, we expect that a linear mapping $\mathbf{W}f(i)$ to be sufficiently as:

$$loss(i; \mathbf{W}) = \|\mathbf{W}f(i) - \mathbf{x}_i\|_1 \quad (1)$$

which is a simple regression task with ℓ_1-norm since ℓ_2-norm is sensitive to outliers. However, one could argue to use a more complex and discriminative loss function as in [5]:

$$loss(i; \mathbf{W}) = \sum_{j \neq i} \max\left(0, margin - \mathbf{x}_i^T \mathbf{W} f(i) + \mathbf{x}_j^T \mathbf{W} f(i)\right) \quad (2)$$

Unfortunately, in practice, we find that training a deep model with the loss in Eq. (2) is very tricky. First, it is difficult to define the negative samples of i. For millions or billions images, even maintaining a small set of margin-violating negative samples can take hours. Second, the negative samples are dependent on the learning of $f(i)$ and hence the training cannot be easily asynchronized. Third, the margin brings in additional hyperparameters which need to be tuned.

So far, we have incorporated visual information into the image-word embedding. After the learning process, at test time, we can extract features for newly arrived images from the fully connected layers before the loss function. Also, we find that the visually fine-tuned image-word embeddings has great potentials in some fundamental multimedia tasks such as key-word based image search and image annotation (see Figure 4).

5. EXPERIMENTS

5.1 Datasets

SBU: the SBU Captioned Photo Dataset, was originally used as a gallery for retrieval based image description generations [11]. It contains a million images with associated descriptive text. In this paper, we further removed stopwords like "is" and "that" and words with frequency less than 5 from the captions. This gives rises to a vocabulary of the size 3,0456. We used SBU as our social multimedia source.

NUSWIDE: it is a popular social image benchmark [1]. It contains 269,648 images across 81 general noun concepts. We followed the official "161,789/107,859" as the "tran/test" split.

5.2 Implementation Details

For the image-word graph deep learning method in Section 3, we used the source codes in [12] with minor modifications. We applied 5-step random walks with 5 repetitions and window-size of 5 for pair-wise optimization. We used 8 computing threads on a 8-core machine. It took about 10 minutes for a good solution.

For deep vision models used in Section 4, we depoyed Caffe framework [7] for CNN implementation on a NVIDIA Titan Z GPU. In particular, we used the well-known AlexNet architecture [8], which consists of 5 convolutional layers with max-pooling and 2 fully connected layers before the loss layer. We used the author provided ImageNet pretrained model (in Caffe format) as initializations for the proposed Regression CNN. The initial learning rate was set to $1e^{-4}$ with dynamic momentum. The size of the batch was 128 and it took 20 epochs to converge. Each epoch took about 90 mins. We used ℓ_2-norm weight decay with $5e^{-5}$ coefficient.

The choice of the embedding dimension is crucial. We tuned the values within $\{300, 400, ..., 2,000\}$ and found that 1,000 was the best choice for accuracy and efficiency.

For classification task, we used LIBLINEAR linear SVM toolbox[3] for classifier implementation. We ℓ_2-normalized all the input features and the trade-off parameter was tuned within $\{0.001, 0.01, 0.1, 1, 10\}$.

5.3 Results

Good features should perform well in classification task. We extracted two kinds of features by our method: 4,096-

[3] www.csie.ntu.edu.tw/ cjlin/liblinear/

Figure 4: Examples of the 5 most similar images to those concepts that are not labeled in NUSWIDE. From collective intelligence, we can generalize visual properties to any word, even for those "words" that do not exist in the target domains. For example, there is no "relax"concept in NUSWIDE, but we can return "vocation" and "beach" concepts, which most relates to "relax".

Table 1: Performance (mAP%) of classification on NUSWIDE.

Dataset/Method	DeCAF	ICMAE	Ours-fc7
NUSWIDE	38.6	32.7	**39.1**

D **Ours-fc7**, which is the 7-th fully-connected layer output of our R-CNN. We compared our features with the following two strong baselines: **DeCAF** [3] , which is a 4,096-D feature from the 7-th fully-connected layer of AlexNet, and **ICMAE** [15], which is a 1,898-D feature learned from a million Flickr images with tags[5].

Table 1 lists the performance of classification using various features. On both datasets, we can see that our two methods considerably outperform DeCAF. This demonstrates the effectiveness of learning from large social collection. We note that ICMAE, which is also learned from social images, performs the worst. The reasons are two folds. First, the dimension is the lowest. Second, it does not explicitly tackle the sparsity and diversity in large but noisy social media. Note that sparse coding image embeddings in terms of word dictionaries performs the best. This is because as compared to the lower-level fc7 feature, the 4,000-D sparse codes endow higher-level semantic information. Moreover, we find that our feature learning framework can also learn useful image-tag embeddings which can be used for querying images out of the concept domain in a data collection (see Figure 4).

6. CONCLUSIONS

We presented a novel feature learning paradigm from large-scale but noisy social multimedia. Our key idea is to acquire collective intelligence from big multimedia data, which helps to better understand the semantic relations between images and words. In order to achieve this, we proposed a deep learning method to learn image-word embeddings from a million image-caption pairs of SBU dataset. Then, we proposed a visual deep learning method to map images into the embeddings. Experiments on NUSWIDE benchmark validated the effectiveness of our learned features. We believe that the proposed paradigm is of great practical use since it is effective, efficient and easy-to-use.

7. ACKNOWLEDGMENTS

This work was supported by NUS-Tsinghua Extreme Search (NExT) project under the grant No.: R-252-300-001-490 and the National Natural Science Foundation of China, No. 61303075.

[5]http://pub.nextcenter.org/deeplearning/modelParams.tar

8. REFERENCES

[1] T.-S. 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 *CIVR*, 2009.

[2] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, et al. Large scale distributed deep networks. In *NIPS*, 2012.

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[4] C. Fang, H. Jin, J. Yang, and Z. Lin. Collaborative feature learning from social media. In *CVPR*, 2015.

[5] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[6] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*. 2014.

[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint*, 2014.

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

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

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[11] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

[12] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *KDD*, 2014.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint*, 2014.

[14] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *arXiv preprint*, 2014.

[15] H. Zhang, Y. Yang, H. Luan, S. Yan, and T.-S. Chua. Start from scratch: Towards automatically identifying, modeling, and naming visual attributes. In *MM*, 2014.