ENCYCLOPEDIA ENHANCED SEMANTIC EMBEDDING FOR ZERO-SHOT LEARNING

Zhen Jia[1,2], *Junge Zhang*[1,2], *Kaiqi Huang*[1,2,3], *Tieniu Tan*[1,2,3]

[1] CRIPAC & NLPR, Institute of Automation, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] CAS Center for Excellence in Brain Science and Intelligence Technology
{zhen.jia, jgzhang, kqhuang, tnt}@nlpr.ia.ac.cn

ABSTRACT

There are tremendous object categories in the real world besides those in image datasets. Zero-shot learning aims to recognize image categories which are unseen in the training set. A large number of previous zero-shot learning models use word vectors of the class labels directly as category prototypes in the semantic embedding space. But word vectors cannot obtain the global knowledge of an image category sufficiently. In this paper, we propose a new encyclopedia enhanced semantic embedding model to promote the discriminative capability of word vector prototypes with the global knowledge of each image category. The proposed model extracts the TF-IDF key words from encyclopedia articles to acquire the global knowledge of each category. The convex combination of the key words' word vectors acts as the prototypes of the object categories. The prototypes of seen and unseen classes build up the embedding space where the nearest neighbour search is implemented to recognize the unseen images. The experiments show that the proposed method achieves the state-of-the-art performance on the challenging ImageNet Fall 2011 1k2hop dataset.

Index Terms— zero-shot learning, image classification

1. INTRODUCTION

Image classification has gained huge progress in recent years, due to the impressive improvement of deep learning methods, such as convolutional neural networks (CNNs) [1, 2, 3, 4], and large scale datasets [5]. Some CNN-based image classification methods [6] even trump human performance on ImageNet classification task. Meanwhile, almost all successful image classification methods mentioned above are supervised models, which take large scale captioned image data to get convergent. Early research on human cognition [7] shows that human have the ability to recognize more than 30,000 object categories and objects with components removed or non-rigid deformation. What's more, human can recognize objects they've never seen before. For instance, human can easily tell apart different cat categories by just reading their text descriptions. A child can also recognize a zebra at the first sight if he has seen a horse before and known that a zebra looks like a horse with white and black stripes. We strongly hope that the machine image classification systems have the similar ability as human beings to transfer knowledge from other modalities to visual area, i.e. to recognize image categories which don't appear in the training set.

Zero-shot learning (ZSL) aims to deal with image classification task in which the test categories have no overlap with training categories. This topic draws an increasing attention of computer vision researchers. Many computer vision and machine learning methods, such as probabilistic models [8, 9, 10], canonical correlation analysis [11, 12], metric learning methods [13, 14] and graphical models [15] are exploited to solve the ZSL problem. In order to classify unseen images, the first step is to build a semantic embedding space where all the image classes are represented as their prototypes. Attribute features, word vectors and image descriptions of the categories are the typical side information to form the embedding space. C. Lampert et al. [8, 9] come up with probabilistic models – direct and indirect attribute prediction models (DAP and IAP) to predict the unseen images using their attribute features as prototypes. The deep visual-semantic embedding model (DeViSE) [16] maps CNN image features to the word vector embedding space. DeViSE model explore the semantic and syntactic properties of word vector as shown in [17]. Recently, Z. Akata et al. [18, 19] utilize image descriptions as side information to build the embedding space. In these three kinds of side information, word vectors demonstrate more advantages than attribute features and image descriptions to materialize prototypes, because they are liberated from human annotations which are quite expensive and time consuming. Thus word vector is an ideal prototype to solve large scale ZSL problem.

Many proposed ZSL methods [11, 12, 16, 20, 21, 22] use the word vectors of the class labels as the classification prototypes directly, which has negative effect on zero-shot classification. The word vectors extracting algorithms, such as skip-gram method [17], usually set the size of training window to a small number that makes word vectors unable to gain the global knowledge of a category in the corpus. The global knowledge is the more comprehensive and scientific repre-

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

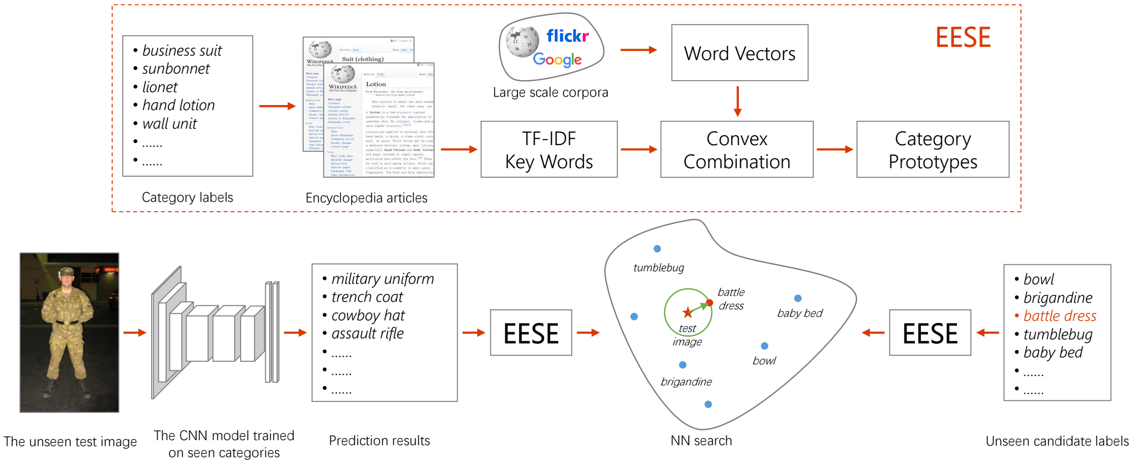


Fig. 1. The framework of encyclopedia enhanced semantic embedding

sentation of a category than the local knowledge obtained in the short context. Thus global knowledge can represent distinctiveness of a category better than local knowledge. For zero-shot learning, especially in large scale tasks, some categories are usually close to each other in the semantic space, which are difficult to tell apart, let alone they are unseen in the training set. It is imperative for us to use the prototypes with global knowledge to improve ZSL models.

In order to enhance the discriminative ability of word vector prototypes with category global knowledge, we present a new semantic embedding model – encyclopedia enhanced semantic embedding (EESE) for ZSL. We explore the online encyclopedias, such as Wikipedia and Britannica, which contains articles defining various object categories. Each article embodies the peculiar and comprehensive knowledge of the object category. The proposed EESE model synthesizes the global knowledge of each object category by extracting the term frequency-inverse document frequency (TF-IDF) key words of corresponding articles. TF-IDF is popular in the research of information retrieval and natural language processing. The TF-IDF key words reflect the most important and unique words of an article in the corpus. Thus TF-IDF key words contain the global knowledge which is significant to a category and distinct from others. This property of TF-IDF key words makes them appropriate to dig the global knowledge of each category. With the global knowledge extracted from the encyclopedia articles, the prototypes' discriminative ability get boosted. The experiments show that the EESE model achieves much better performance than previous models in ZSL task on the challenging ImageNet Fall 2011 dataset.

This paper is organized as follows. Section 2 describes the ZSL problem statement. Section 3 and Section 4 show the proposed method and experimental results in detail. Finally, Section 5 makes the conclusion of this paper.

2. ZERO-SHOT LEARNING STATEMENT

In the zero-shot learning task, the training image categories are referred to as *seen classes* and the test image categories are referred to as *unseen classes*. The visual knowledge of the unseen classes is extracted from other modal data which is called *side information*. With the side information, we build up an *embedding space*, in which the *prototypes* exhibit the essential semantic features of each category.

We define the training set as $\mathcal{S} \equiv \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, where $\mathbf{x}_i$ is a p-dim vector which represents an image in the training set. Meanwhile, $y_i \in Y = \{1, 2, \ldots, n\}$ indicates the category label of image $\mathbf{x}_i$. There are m images from n distinct seen classes in the training set. The test set is defined as $\mathcal{U} \equiv \{(\mathbf{x}'_j, y'_j)\}_{j=1}^{m'}$. We have n' unseen classes in the test set, i.e. $y'_j \in Y' = \{n+1, \ldots, n+n'\}$. The purpose of ZSL is to train a classifier on $\mathcal{S}$. At the same time, we expect the classifier can achieve an favourable performance on $\mathcal{U}$. Considering that the seen and unseen classes are disjoint, i.e. $Y \cap Y' = \emptyset$, side information is needed to transfer the classifier learned on seen classes to unseen classes.

3. METHODOLOGY

The proposed encyclopedia enhanced semantic embedding method is illustrated in Fig. 1. The EESE model is proposed to produce prototypes of the seen and unseen classes in the semantic embedding space. Firstly, the model extracts TF-IDF key words of each image category's encyclopedia article. These key words contain more global knowledge of corresponding categories than the mere category labels. Then the model combines word vectors of the key words convexly based on their TF-IDF weights. The combined vector acts as the prototype of the image category to build up the semantic embedding space.

When an test image comes, a CNN classification model trained on the training set will predict its classification probabilities on the seen categories. Then the prototypes of seen classes with the highest classification probabilities are convexly combined to form the semantic vector for the test image. Finally, we put this semantic vector into the embedding space and search for the nearest neighbour prototype as the final zero-shot prediction result of the unseen image.

3.1. Zero-Shot Learning Framework

In this paper, we apply the convex combination of semantic embeddings (ConSE) [20] as the baseline as well as the ZSL framework to collaborate with the proposed EESE model. During the training stage, we train a CNN classification model using the seen images. We also need to train the word vectors from the large scale corpora, such as Wikipedia, Google News et al. Word vectors act as the bridge to correlate the seen and unseen classes semantically.

In the test stage, the pre-trained CNN model will classify the unseen test image **x** firstly as shown in Equation 1.

$$y(\mathbf{x}, 1) = \arg\max_{y \in Y} p(y|\mathbf{x}) \ . \tag{1}$$

Note that the highest probability classification result, $y(\mathbf{x}, 1)$, must be wrong for image **x**, because the unseen image is assigned to seen class labels. But the result shares some visual similarity with the unseen ground truth. The classification probability p measures the degree of the similarity. Let $y(\mathbf{x}, t)$ denote the t^{th} highest classification result of **x**, then $p(y(\mathbf{x}, t)|\mathbf{x})$ is the classification probability. The model calculates the convex combination of the corresponding seen classes' prototypes, i.e. $s(y(\mathbf{x}, t))$, with the top T probabilities as the weights. Formally,

$$f(\mathbf{x}) = \frac{1}{Z} \sum_{t=1}^{T} p(y(\mathbf{x}, t)|x) \cdot s(y(\mathbf{x}, t)) \ , \tag{2}$$

where $Z = \sum_{t=1}^{T} p(y(\mathbf{x}, t)|x)$ is the normalization factor.

The convex combination result, $f(\mathbf{x})$, is the representing vector of the unseen image in the semantic space. At last, we search for the nearest-neighbour prototype in the semantic embedding space as the unseen image's zero-shot prediction result. Formally,

$$y_{unseen}(\mathbf{x}) = \arg\max_{y' \in Y'} sim(f(\mathbf{x}),\ s(y')) \ , \tag{3}$$

where $sim(\cdot)$ indicates the similarity metric in the model. As usual, we adopt cosine similarity here.

3.2. Encyclopedia Enhanced Semantic Embedding

In the proposed EESE model, we extract the TF-IDF key words of each image class encyclopedia article to form prototypes. As shown in Equation 4, the TF-IDF value of a word in an article is composed of term frequency and inverse document frequency.

$$TF\text{-}IDF_{id} = \frac{n_{id}}{\sum_k n_{kd}} \cdot \log \frac{|D|}{|\{d \in D : i \in d\}|} \ . \tag{4}$$

We divide the number of word i in article d by the number of all words in d as i's term frequency. Then we calculate the inverse fraction of the articles that contain word i, which is obtained by dividing the total number of articles, $|D|$, by the number of articles containing word i. Then we take the logarithm of that quotient as the inverse document frequency. Finally, the TF-IDF value of word i in article d is computed by multiplying the above two statistics. TF-IDF value reflects how important a word is to an article in the whole corpus. It is sensible to capture the global knowledge of a category from the corresponding encyclopedia article using its TF-IDF key words.

Distinguished from the ConSE baseline taking word vectors as categories' prototypes, for an object category w, the EESE model extracts the top m TF-IDF key words i_t ($t \in \{1, 2, 3, \ldots, m\}$) of its article in the encyclopedia. Then the model calculates the convex combination of the key words' word vectors weighted by their corresponding TF-IDF values. Formally,

$$S_{EESE}(w) = \frac{1}{U} \sum_{t=1}^{m} s(i_t) \cdot W_{TF\text{-}IDF}(i_t) \ , \tag{5}$$

where $s(i_t)$ is the word vector of key word i_t, and $U = \sum_{t=1}^{m} W_{TF\text{-}IDF}(i_t)$ is the normalization factor. The resulting $S_{EESE}(w)$ serves as the prototype of object category w in the EESE space.

Then we combine the EESE model with the zero-shot learning framework in Section 3.1 to solve the ZSL problem.

$$\begin{gathered} f_{EESE}(\mathbf{x}) = \frac{1}{Z} \sum_{t=1}^{T} p(y(\mathbf{x}, t)|x) \cdot S_{EESE}(y(\mathbf{x}, t)) \ , \\ y_{EESE}(\mathbf{x}) = \arg\max_{y' \in Y'} sim(f_{EESE}(\mathbf{x}),\ S_{EESE}(y')) \ . \end{gathered} \tag{6}$$

As shown in Equation 6, the final zero-shot classification result of the unseen test image **x** is determined by its nearest neighbour search result in the EESE space.

4. EXPERIMENTS

4.1. Setup

Datasets In the experiments, we train the models on ILSVRC 2012 1K dataset and test them on the 1,548-category ImageNet Fall 2011 1k2hop dataset. The 1k2hop dataset is a subset of the ImageNet Fall 2011 dataset, containing the classes that are within 2-tree-hop of the ILSVRC 2012 1K classes according to the ImageNet hierarchy. There are more than 1.3 million unseen images in the test set.

Visual and semantic features In order to compare with the previous work fairly, we use the same CNN model and word vectors as in [22]. We train a GoogLeNet CNN model on the ILSVRC 2012 1K dataset. For word vectors, we train a skip-gram model [17] on the latest Wikipedia dump corpus. We set the training window size as 5 and extract the 500-dim word vectors. In EESE, we also extract the TF-IDF key words from the Wikipedia articles to acquire the global knowledge of each category.

Evaluation metric We adopt flat hit@K as evaluation metric to compare different models. Flat hit@K indicates the percentage of the unseen test images for which the model returns the true label in its top k predictions. Thus the flat hit@1 is the classification accuracy as typically used.

4.2. Results

Due to the huge diversity of unseen image classes, there are only a few published papers focusing on the large scale zero-shot classification task, such as [14, 16, 20, 21, 22]. The models which test themselves on the challenging ImageNet Fall 2011 dataset are even fewer. In this paper, we choose ConSE model [20] as the experiment baseline. The ConSE framework utilizes the word vectors' semantic and syntactic property sufficiently. The concise ConSE model is a faultless framework to test the effectiveness of different semantic embeddings. Besides ConSE, we also compare the EESE model with the Synthesized Classifiers model (SynC) in [22], which tackles the ZSL problem from the perspective of manifold learning.

The comparison between proposed EESE model and the previously published state-of-the-art models is shown in Table 1. As we can see, the EESE model achieves much better performance than the ConSE and SynC model on large scale zero-shot classification task.

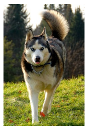
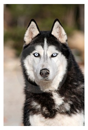
Table 1. Comparative results (%) on ImageNet 1k2hop

Method	Flat hit@K			
	1	2	5	10
ConSE [20]	9.4	15.1	24.7	32.7
ConSE by us	15.5	24.7	37.4	45.4
SynC$^{\text{struct}}$ [22]	9.8	15.3	25.8	35.8
SynC$^{\text{o-vs-o}}$ [22]	10.5	16.7	28.6	40.1
EESE	**18.8**	**28.0**	**41.0**	**49.2**

In Table 1, the performance of ConSE [20] is achieved by AlexNet rather than GoogLeNet which is used in [22] and this paper. We re-implement the ConSE model with GoogLeNet and the identical word vectors as in EESE. The result is shown as *ConSE by us* in Table 1. In the re-implementation, we sum the word vector of every word as the category prototype if the category label is a phrase. The performance of our re-implementation is improved compared to [20].

The ZSL semantic embedding space is established by the seen and unseen classes' prototypes. Each prototype in EESE is composed of the TF-IDF key words. As is displayed, we show the top-5 key words extracted from Wikipedia articles of several classes in Table 2. In the table, *Eskimo dog, husky* and *Siberian husky* are from seen classes, while *Fore plane* and *Smooth plane, smoothing plane* are examples of unseen classes. The key words are listed in descending order of their weights. It clearly shows that, by using the proposed EESE model, we can obtain quite different key words to build the embedding space even for visually and semantically similar object classes. These key words indicate the global knowledge of each class and make great contributions to zero-shot classification of the object classes that even human can hardly tell apart. The EESE model effectively exploits the global knowledge of each category to enhance the discriminative capability of class prototypes.

Table 2. Examples of the EESE key words

Image Example	Ground Truth	EESE Key Words
	Eskimo dog, husky	husky sled owner arctic admixture
	Siberian husky	siberian husky nome sled chukchi
	Fore plane	jointer plane thickness undulation skim
	Smooth plane, smoothing plane	plane smooth finish wood tearout

5. CONCLUSION

In this paper, we have proposed the encyclopedia enhanced semantic embedding model to solve the zero-shot learning problem. We explore the encyclopedia article to extract the global knowledge of each image category. The proposed EESE model promotes the discriminative capability of prototypes with the categories' global knowledge. The EESE model are easy to be implemented on large scale zero-shot classification problem and also perform much better than the previous models. We achieve the state-of-the-art performance on the challenging ImageNet 1k2hop dataset.

6. ACKNOWLEDGEMENT

This work is funded by the National Natural Science Foundation of China (Grant 61403387 and Grant 61673375), the National Key Research and Development Program of China (Grant 2016YFB1001004 and Grant 2016YFB1001005) and the Projects of Chinese Academy of Sciences (Grant QYZDB-SSW-JSC006 and Grant 173211KYSB20160008).

7. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[3] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” *arXiv preprint arXiv:1512.03385*, 2015.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[7] Irving Biederman, “Recognition-by-components: a theory of human image understanding.,” *Psychological review*, vol. 94, no. 2, pp. 115, 1987.

[8] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling, “Learning to detect unseen object classes by between-class attribute transfer,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 951–958.

[9] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling, “Attribute-based classification for zero-shot visual object categorization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2014.

[10] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng, “Zero-shot learning through cross-modal transfer,” in *Advances in neural information processing systems*, 2013, pp. 935–943.

[11] Yanwei Fu, Timothy M Hospedales, Tao Xiang, Zhenyong Fu, and Shaogang Gong, “Transductive multi-view embedding for zero-shot recognition and annotation,” in *European Conference on Computer Vision*. Springer, 2014, pp. 584–599.

[12] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong, “Transductive multi-view zero-shot learning,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 11, pp. 2332–2345, 2015.

[13] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie, “Improving semantic embedding consistency by metric learning for zero-shot classiffication,” in *European Conference on Computer Vision*. Springer, 2016, pp. 730–746.

[14] Chen Huang, Chen Change Loy, and Xiaoou Tang, “Local similarity-aware deep feature embedding,” in *Advances in Neural Information Processing Systems*, 2016, pp. 1262–1270.

[15] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong, “Zero-shot object recognition by semantic manifold distance,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2635–2644.

[16] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al., “Devise: A deep visual-semantic embedding model,” in *Advances in neural information processing systems*, 2013, pp. 2121–2129.

[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[18] Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele, “Multi-cue zero-shot learning with strong supervision,” *arXiv preprint arXiv:1603.08754*, 2016.

[19] Scott Reed, Zeynep Akata, Bernt Schiele, and Honglak Lee, “Learning deep representations of fine-grained visual descriptions,” *arXiv preprint arXiv:1605.05395*, 2016.

[20] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean, “Zero-shot learning by convex combination of semantic embeddings,” *arXiv preprint arXiv:1312.5650*, 2013.

[21] Li Zhang, Tao Xiang, and Shaogang Gong, “Learning a deep embedding model for zero-shot learning,” *arXiv preprint arXiv:1611.05088*, 2016.

[22] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha, “Synthesized classifiers for zero-shot learning,” *arXiv preprint arXiv:1603.00550*, 2016.