# Word Vector Augmentation by its Definition for Zero-shot Image Classification

Kotaro Kikuchi Naohiro Tawara Tetsunori Kobayashi Yoshihiko Hayashi
Department of Computer Science and Engineering
Waseda University, Tokyo, Japan
{kotaro, tawara}@pcl.cs.waseda.ac.jp koba@waseda.jp yshk.hayashi@aoni.waseda.jp

## 1. Introduction

Zero-shot learning (ZSL) aims to classify an image into one of the most probable novel concepts. Here, a novel concept means a concept for which no training has been conducted. ZSL, often seen as a class of transfer learning, can be accomplished by exploiting already learned vision-concept correspondences. More specifically, the given image can be mapped into a feature space by applying the pre-trained mapping function. Assuming that a novel concept has also been represented in the same feature space as the known concepts, the mapped representation can be compared with the representation of a novel concept, enabling zero-shot classification.

Then the issue would be how/where to effectively obtain visual/semantic features. A previous work [10] achieved the current state-of-the-art accuracy by employing the word embeddings of a concept name as the semantic feature (henceforth, concept name feature), suggesting that semantic feature is vital in ZSL.

The present work extends their framework by further incorporating semantic feature (henceforth, definition feature) provided by the definition of a word given in the WordNet lexical-semantic resource. More specifically, we combine the concept name feature with the definition feature, and learn the correspondences with visual features by a simple MLP. This method, however, could be affected by noisy words (e.g. common words) in a definition. We thus propose to incorporate an attention mechanism into the framework to improve the accuracy. The empirical results shows that our method can achieve comparable accuracy with the current state-of-the-art, while improving the interpretabilitiy thanks to the introduction of the attention mechanism.

## 2. Related Work

Attribute-based visual recognition is a powerful way to predict novel concepts with the interpretability [1, 9]. However, defining attributes are hard for large-scale ZSL due to a diversity of target concepts. There are few attempts to extract the attributive information from the articles of concepts, but they do not achieve good results in the accuracy compared to the attribute-based approach [2, 4]. A defacto approach on the large-scale ZSL is utilizing the current unsupervised learning methods based on the statistics of neighbor co-occurrence in massive texts [10]. This unsupervised method can give dense vector representations to various words and capture the word similarity [7]. In this work, we exploit the composed meaning of the concepts in a dictionary definition. Our approach can be applied if dictionary definitions of target concepts are available, so it is suitable for the large-scale ZSL.

## 3. Method

### 3.1. Problem definition

Let $\mathcal{T}_{\mathrm{tr}}$ be a set of labels, and $\mathcal{D}_{\mathrm{tr}} = \{(\boldsymbol{I}_i, t_i, \mathcal{Y}_{t_i}) | 0 \leq i \leq N\}$ be a training dataset consisting of $N$ triples of: an image $\boldsymbol{I}_i$, its label $t_i \in \mathcal{T}_{\mathrm{tr}}$, and the corresponding auxiliary information $\mathcal{Y}_{t_i}$, which generally provides semantic information for the labeled concept.

Zero-shot image classification, then, is defined so as to estimate the label $t_j$ of a given image $\boldsymbol{I}_j$ by exploiting the auxiliary information $\mathcal{Y}_{t_i}$, where the label $t_j$ is a member of the set $\mathcal{T}_{\mathrm{te}}$ of unseen labels. Note here that $\mathcal{T}_{\mathrm{tr}} \cap \mathcal{T}_{\mathrm{te}} = \emptyset$. We assume that even for an unseen label, corresponding auxiliary information can be obtained.

### 3.2. Baseline method

We use a current state-of-the-art framework proposed in [10] to apply our idea. It uses the distributed representation of the concept name as the auxiliary information. This method embeds distributed representation of a concept name $\varphi(w^{cls})$ by function $F$ as follows[1]:

$$F(\mathcal{Y}_t; W) = f(W_1 \varphi(w^{cls})) \tag{1}$$

[1] Due to parameter reduction, we use a one-layer MLP instead of a two-layer MLP in [10].

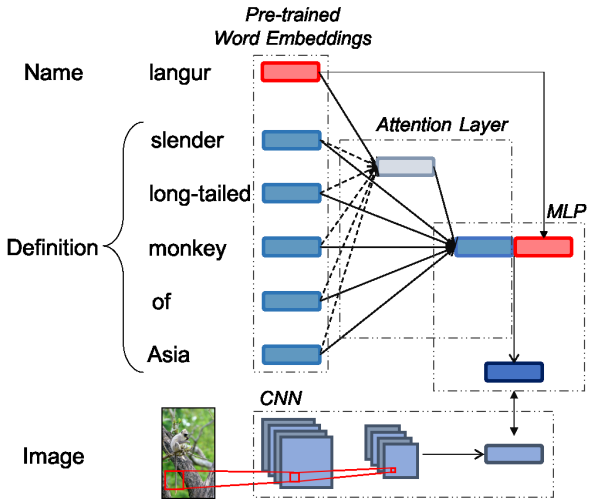

Figure 1. Overview of our proposed method

where $W$ are learnable parameters and $f(.)$ is an activate function. $\varphi(w)$ are $D_\varphi$-dimensional distributed representation of a word. Parameters are learned through back-propagation by minimizing the objective function which consists of mean squared error and $L_2$ regularization as follows:

$$\frac{1}{N}\sum_{i=1}^{N} || \theta(\boldsymbol{I}_i) - F(\mathcal{Y}_{t_i}; W) ||^2 + \lambda || W ||^2 \quad (2)$$

where $\theta(\boldsymbol{I}_i)$ are $D_x$-dimensional visual features and $\lambda$ is the hyper-parameter to adjust the regularization strengths.

During the test phase, most likely label is assigned to a test image $\boldsymbol{I}_j$ in the following:

$$\arg\min_k || \theta(\boldsymbol{I}_j) - F(\mathcal{Y}_{t_k}; W) ||^2 \quad (3)$$

### 3.3. Proposed method

Figure 1 illustrates our proposed method which augments the baseline method by introducing a concept definition as the additional auxiliary information. Extracted features from a concept definition are simply concatenated to the name derived features, so our embedding function is:

$$F(\mathcal{Y}_t; W) = f(W_1[\varphi(w^{cls}); \upsilon(\mathcal{D}; W)]). \quad (4)$$

where $\upsilon(.)$ is the feature extraction function for a concept definition $\mathcal{D}$. Parameters of our embedding function are determined by minimizing Eq. (2).

We use weighted aggregation of word vectors with attention mechanism as the feature extraction from a concept definition. Weighted aggregation of word vectors with attention mechanism is a reasonable representation while emphasizing contributive words and suppressing irrelevant words in a definition. Given distributed representation for a name and each word in a definition, attention weights are estimated through two-layer MLP:

$$s_n = f(W_2 f(W_3[\varphi(w^{cls}); \varphi(w_n^{def})])) \quad (5)$$

$$a_n = \frac{\exp(s_n)}{\sum_l^L \exp(s_l)} \quad (6)$$

$$\upsilon(\mathcal{D}) = \sum_{n=1}^{L} a_n \varphi(w_n^{def}) \quad (7)$$

where $w_n^{def}$ is the $n$-th word in a definition and $L$ is a number of words in a definition. $a_n$ is the attention weight for the $n$-th word.

| | |
|---|---|
| Number of labels in the training set | 1,000 |
| Number of images in the training set | 200,000 |
| Number of labels in the test set | 360 |
| Number of images in the test set | 18,000 |
| Dimension of the visual feature $D_x$ | 2,048 |
| Dimension of the word vector $D_\varphi$ | 1,000 |

Table 1. Experimental settings

## 4. Experiments

### 4.1. Experimental setup

We conduct experiment to evaluate our proposed method following the ILSVRC 2012/2010 setting (see [10] for details). The image dataset of this setting used is ImageNet [3] in which each image is related to a synset defined in WordNet [8]. As a definition, we use a description for each synset excluding example sentences on WordNet, i.e. gloss. We examine the performance of attention mechanism by comparing to the method without the attention mechanism which uses normalizing weights instead of attention weights. Averaged vectors for all words in a synset are used as a name-derived feature $\varphi(w^{cls})$. We use an intermediate representation extracted from the last pooling layer of ResNet-152 [5] as the visual feature $\theta(\boldsymbol{I})$. We use the word vector representation $\varphi(w)$ obtained by Skip-gram model [7]. Adam [6] is applied to optimize all methods in this experiment. The numbers of images and labels we used and parameters for our method are shown in Table 1.

### 4.2. Evaluation

Given a test image, we rank all candidate labels in the test label set $\mathcal{T}_{\text{te}}$ based on the scores calculated by Eq.(3). We create a ranking for each test image, then count correct instances in top $k$. We call this metrics Top@$k$ and use it to evaluate performance for zero-shot image classification.

Label n11950345

Concept mayfly, dayfly, shadfly

Definition

| slender | insect | with | delicate | membranous_wings | having | an | aquatic_larval_stage |
|---|---|---|---|---|---|---|---|
| 0.01 | 50.97 | x | 0.00 | 41.22 | 0.00 | x | 7.80 |

| and | terrestrial | adult | stage | usually | lasting_less_than | two | days |
|---|---|---|---|---|---|---|---|
| x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Figure 2. Attention weights in the success case of proposed method. The each number under a word means an attention weight and x represents a stopword.

| Model | Top@1 | Top@5 |
|---|---|---|
| Zhang *et al*. [10] | 10.89 | **26.42** |
| Ours | **11.06** | 26.22 |
| Ours (w/o attention) | 9.28 | 22.47 |

Table 3. Zero-shot classification accuracy (%) comparison on ILSVRC 2012/2010

### 4.3. Experimental results

Table 2 shows Top@1 and Top@5 results on ILSVRC 2012/2010. While our method is superior to baseline in Top@1 metrics, ours is worse in Top@5 metrics, so the performance of our method can be concluded equivalent to the baseline. Figure 2 shows attention weights for each word in a definition in the case of our method succeeded. In this example, *insect* and *membranous wings* are emphasized. The word *insect* is a hypernym of the concept name *mayfly* and *membranous wings* are a visible part meronym of the *mayfly*. Three trends are observed in the entire result: a hypernym is emphasized only in success cases, a visible part meronym are emphasized in few success cases, and frequent words are almost never emphasized in any cases. These trends indicate the attention mechanism works well to emphasize and suppress words.

## 5. Conclusion

We proposed a novel framework for zero-shot image classification that could successfully extract semantic features acquired from the dictionary definition of a concept by incorporating an attention mechanism. The empirical results shows that our method can achieve comparable accuracy with the current state-of-the-art, while improving the interpretabilitiy benefited from the attention mechanism.

For future work, we plan to assess the performance of our method in various experimental settings. We also explore to incorporate further semantic information that can be mined from lexical-semantic resources, such as WordNet. For instance, we could exploit semantic features acquired from the hypernym of a target concept.

## Acknowledgment

The present work was supported by JSPS KAKENHI Grant Number JP17H01831, and Waseda University Leading Graduate Program for Embodiment Informatics.

## References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *TPAMI*, 38(7):1425–1438, 2016.

[2] J. L. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov. Predicting deep zeroshot convolutional neural networks using textual descriptions. In *Proc. ICCV*, pages 4247–4255, 2015.

[3] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, pages 248–255, 2009.

[4] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proc. ICCV*, pages 2584–2591, 2013.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016.

[6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, pages 3111–3119. 2013.

[8] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[9] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *Proc. CVPR*, pages 49–58, 2016.

[10] L. Zhang, T. Xiang, and S. Gong. Learning a deep embedding model for zero-shot learning. *CoRR*, abs/1611.05088, 2016.