Visually grounded word embeddings for zero-shot learning of visual categories

Hascoet Tristan[1,a)] Yasuo Ariki[1] Tetsuya Takiguchi[1]

Abstract: Traditional object recognition models are bound to a close-world assumptions as they can only discriminate among a finite set of visual classes predefined by the annotated training data available to learn from. This contrasts with the human ability to continuously learn and define new visual classes using natural language. Zero-shot learning (ZSL) models are able to generalize to unseen classes beyond the finite set of class labels on which they have been trained. To do so, ZSL models embed visual class labels in a high-level visual feature space, or semantic space, shared by both training and test classes. The information shared in this label space by training and test classes allows to transfer the visual knowledge learned from the set of training classes to the unseen set of test classes. Early works on ZSL use small domain datasets for which manually annotated visual attributes are used as high-level representations of the visual classes. However, the complexity of designing a finite set of visual primitives able to describe the wide variations in the appearance of naturally occurring object seems untractable so that such manual labeling approach is unlikely to scale to generic object recognition settings. In this paper, we represent visual classes as words and propose to learn their high-level visual feature representations from the co-occurence statistics of their label in large text corpora. Our method can be seen as implicitly learning visually-grounded word embeddings. We perform preliminary experiments on the task of large-scale ZSL using the Imagenet dataset.

1. Introduction

Classification using convolution neural networks (CNN) has become the backbone of modern computer vision systems. Object detection models, for instance, combine the standard classification architecture of CNN with a bounding box regression model. Semantic segmentation models augment the standard classification CNN architecture with a deconvolution path to perform pixel-wise classification. Trained on large amount of high quality labeled data as made available by the Imagenet dataset [1], CNNs have become very accurate at recognizing generic objects in images, with studies reporting higher precision than the human average [2]. However, these systems are fundamentally limited by the underlying close-word assumption of traditional classifiers, i.e., they can only discriminate among the finite set of training classes on which they have been trained.

Although CNNs have become very precise on standard generic object recognition benchmarks, deploying an efficient discriminative model, in practice, requires both a significant amount of training data and expertise to fine-tune a CNN on the specific task at hand, which hinders the adoption of this technology. An ideal generic object recognition system should allow its user to seamlessly specify task-specific visual classes by naming them at inference time, without the need for expensive data collection and model fine-tuning. The ZSL paradigm addresses this problem and holds the promise to deliver truly generic object recognition systems.

Figure 1 illustrates training and inference steps of ZSL models. The key idea behind ZSL is to embed visual classes into a high-level visual feature space. At training time, ZSL models learn a mapping from the raw image input space to the high-level visual feature space using the feature representations of training classes. Given a model f, parameterized by a set of parameters θ and a distance metric $dist$, ZSL models can be trained by minimizing the distance between model outputs and ground truth embeddings with regards to the model's parameters over a set of labelled training data.

$$f_\theta : x \rightarrow y \tag{1a}$$

$$f_\theta : \mathbb{R}^{3\times h\times w} \rightarrow \mathbb{R}^d \tag{1b}$$

$$\forall (x_i, y_i) \in Tr, (x_i, y_i) \in \mathbb{R}^{3\times h\times w} \times \mathbb{R}^d \tag{1c}$$

$$\theta^* = argmin_\theta(\mathbb{E}_{(x_i,y_i)\in Tr} dist(f(x_i), y_i) \tag{1d}$$

In a ZSL setting, the full set of classes is typically split into two disjoint sets of training and test classes, and classes from both sets are embedded into a shared high-level visual feature label space.

$$Y_{train} = \{y_i, i \in [0, N_{train}]\} \tag{2a}$$

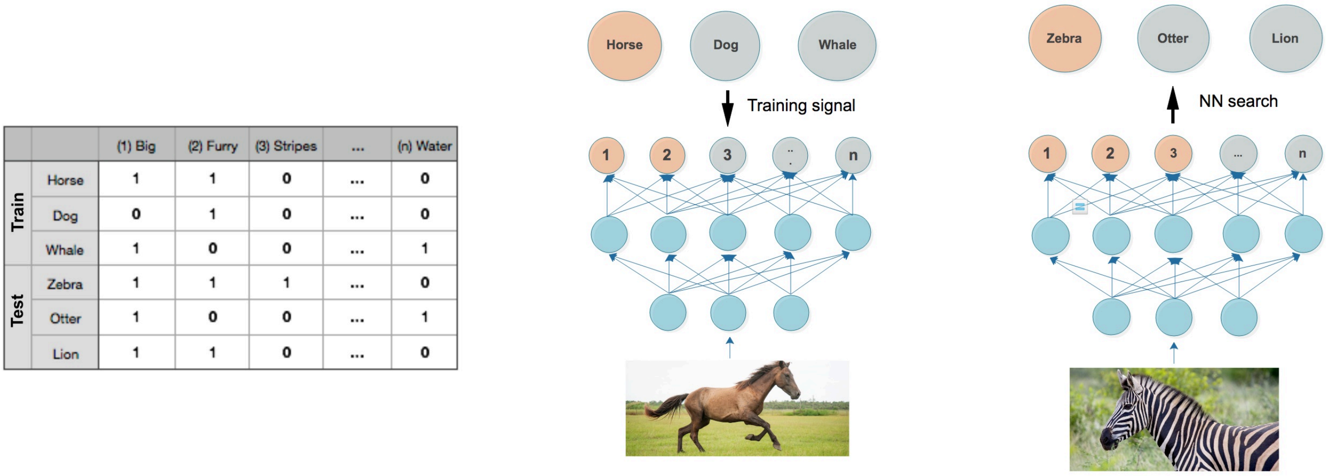
$$Y_{test} = \{y_i, i \in [0, N_{test}]\} \tag{2b}$$

$$Y_{test} \cap Y_{test} = \emptyset \tag{2c}$$

$$\forall y \in Y_{train} \cup Y_{test}, y \in \mathbb{R}^d \tag{2d}$$

At inference, a test image x can be classified seamlessly among training or test classes by nearest neighbor search of $f(x)$ in label space among their respective label sets Y_{train} or Y_{test}.

[1] Kobe University
[a)] tristan.hascoet@me.cs.scitec.kobe-u.ac.jp



		(1) Big	(2) Furry	(3) Stripes	...	(n) Water
Train	Horse	1	1	0	...	0
	Dog	0	1	0	...	0
	Whale	1	0	0	...	1
Test	Zebra	1	1	1	...	0
	Otter	1	0	0	...	1
	Lion	1	1	0	...	0

a. Visual class attributes

b. Training

c. Inference

Fig. 1 Illustration of the ZSL process. A set of visual attributes

$$y^* = argmin_{y \in Y} dist(f(x), y) \qquad (3)$$

ZSL models rely on visually discriminative high-level feature representations of the visual classes y to transfer the discriminative knowledge learned from known training classes Y_{train} to unseen classes Y_{test}. In standard benchmark like AwA [7], class labels are made available as manual annotations of the visual class attributes as illustrated in Figure 1. However, such annotation does not exist for the large-scale generic object recognition problem in which we are interested. In this paper, we propose a method to learn such kind of representation from text data.

Previous works [3] have proposed to embed visual classes in a word embedding space. Word embeddings are the implementation of the distributional hypothesis [4] that states that the meaning of words can be viably defined by the context in which they appear. Word embedding models compute distributed representations of words that have been shown to encode interesting relationships between words as simple vectorial operations. Most famously, word embeddings have shown to efficiently express analogies such as "A man is to a woman what a boy is to a girl" by the relationship

$$v_{man} - v_{woman} \approx v_{boy} - v_{girl}$$

This suggests that word embeddings implicitly learn to represent abstract notions such as gender along specific dimensions of the embedding space.

The capacity to compose abstract notions through simple arithmetic between distributed represetations is very attractive for ZSL. To illustrate the relevance of this idea, consider the zero-shot recognition ability of humans. One might be able to recognize a zebra without having ever seen one if one has been told that a zebra looks very much like a horse covered in black and white stripped patterns. This description requires both low level visual clues such as the black and white stripped patterns, high level visual concept like a horse and the fuzzy notion of similarity between a horse and a zebra's shapes. An ideal visual feature space would be able to capture the abstract notion of a horse-like shape of both animals by an approximate relationship similar to a word embedding analogical relationship as:

$$v_{horse} - v_{color} \approx v_{zebra} - v_{stripes} - v_{black} - v_{white}$$

Word embeddings are learned in an unsupervised manner from words co-occurence statistics in large text corpora, without any visual supervision so the reason why they have proven to provide visually discriminative class representations for ZSL in previous works [3] is not straightforward. In Section 3.1, we briefly present the word2vec [5] skip-gram with negative sampling model and we provide an explanation as to why co-occurence statistics can provide visually discriminative clues in section 3.2. In section 3.3, we build on this explanation and propose a method to learn visually grounded word representations. Section 4 presents preliminary results using the learned word representations on the task of ZSL using the Imagenet dataset.

2. Related work

2.1 Visual class embeddings for ZSL

Most work on ZSL focus on small domain image datasets that provide visual attribute representations of classes. Among the most widely used benchmarks are the AwA dataset[7], which consists of 50 classes of animals annotated with 85 visual attributes and the CUB dataset[8], which consists of 200 bird species annotated with 312 binary attributes. More related to our work, [3] proposed to use word embeddings as visual class representations to perform ZSL in a large scale setting using the Imagenet dataset. Other works have proposed to learn visual class representations from co-occurence statistics of visual classes in natural images [9], text documents [10] or knowledge bases [11].

2.2 Visually grounded word embeddings

As word embeddings have proven useful in computer vision to perform ZSL, computer vision models have similarly found their way in natural language processing applications. In [12], the authors proposed multi-modal word embeddings obtained by concatenation of word embeddings and CNN-extracted visual features. Most related to our work, [13] proposed a multi-modal skip-gram model by augmenting the original word2vec model with visual supervision to compute multimodal word embeddings. They evaluate their model on both semantic similarity and zero-shot learning tasks.

3. Proposed method

3.1 Original word2vec SGNS model

In this section, we give a brief overview of the Skip-gram with negative sampling (SGNG) model to illustrate why word embeddings yield visually discriminative representations. As we omit several important details of the training procedure for brevity, interested readers are referred to the original paper [5] and to the many in-depth analysis and explanations of this model available on-line for rigorous definition of the model. Let V be a vocabulary of n words. Let us assume a text corpus of T words and denote by $p(v)$ the occurrence frequency (unigram distribution), of words $v \in V$ in the corpus. The original word2vec model learns two distinct d-dimensional representation per words: a context embedding c and a word embedding w:

$$\forall v \in V, c_v \in R^d, w_v \in R^d$$

Given a window size s, learning is performed by iterating over the full corpus T, minimizing a cost function l with respect to both representations w and c by stochastic gradient descent:

$$L(T) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=-s}^{s} l(w_t, c_{t+i}) \tag{4}$$

In the SGNS model, the loss function f is defined as:

$$l(w, c) = log(\sigma(w \cdot c)) - \sum_{0}^{k} \mathbb{E}_{c_n \sim p(n)} log(\sigma(w \cdot c_n)) \tag{5}$$

where k is a negative sampling factor parameterizing the model and the esperance term is estimated by randomly drawing a single sample from the unigram distribution.

3.2 Visual clues in words co-occurence statistics

Let us denote by

$$p(i|j) = \frac{p(i, j)}{s \times p(j)}$$

the co-occurrence frequency of word i within the context of word j for a context window of given size s. Given this notation, we can merge equations (4) and (5) into the following formulation:

$$\begin{aligned} L(T) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=-s}^{s} \Big(& log(\sigma(w_t \cdot c_{t+s})) \\ & - \sum_{0}^{k} \mathbb{E}_{n \sim p} log(\sigma(w_t \cdot c_n)) \Big) \end{aligned} \tag{6a}$$

$$\begin{aligned} L(T) = & \frac{1}{T} \sum_{t=1}^{T} \sum_{i=-s}^{s} log(\sigma(w_t \cdot c)) \\ & - \frac{1}{T} \sum_{t=1}^{T} \sum_{i=-s}^{s} \sum_{0}^{k} \mathbb{E}_{n \sim p} log(\sigma(w_t \cdot c_n)) \end{aligned} \tag{6b}$$

$$\begin{aligned} L(T) = & \mathbb{E}_{w_j \sim p(j), c_i \sim p(i|j)} log(\sigma(w_j \cdot c_i)) \\ & - k \times \mathbb{E}_{w_j \sim p(j), c_i \sim p(i)} log(\sigma(w_j \cdot c_i)) \end{aligned} \tag{6c}$$

Equation (6c) highlights the fact that word embeddings are the results of two optimization constraints. The left part of equation (6c) draws word embedding vectors w_i closer (in the sense of the dot product similarity measure) to context vectors c_j with whom they share high probability of co-occurrence $p(i|j)$. The second term pushes word embeddings w_j further apart from context vectors c_i randomly drawn from the corpus unigram distribution $p(i)$.

Table 1 Co-occurrence probabilities of visual class words with visually discriminative words

	horse	zebra	pineapple
leg	$\mathbf{2.3 \cdot 10^{-5}}$	$\mathbf{6.5 \cdot 10^{-3}}$	0.0
hoof	$\mathbf{7.3 \cdot 10^{-5}}$	$\mathbf{3.2 \cdot 10^{-5}}$	0.0
stripes	$2.2 \cdot 10^{-5}$	$\mathbf{1.4 \cdot 10^{-3}}$	0.0
yellow	$1.2 \cdot 10^{-4}$	$2.9 \cdot 10^{-4}$	$\mathbf{7.1 \cdot 10^{-4}}$

Consider the co-occurrence statistics of generic visual class labels, like ”horse”, ”zebra”, or ”pineapple” with visual attributes such as ”hoof”, ”leg”, ”stripes”, or ”yellow”. As shown in Figure 2, horse and zebras have higher co-occurrence probability with their respective attributes ”hoof” and ”leg” than pineapple does. Similarly, pineapple has higher co-occurrence probability with the attribute ”yellow”. Hence, the word embedding of pineapple is drawn closer to the context embedding vector of yellow than the word embeddings of horse and zebras by the optimization process. Both horse and zebras word embeddings are drawn closer than pineapple to the context embedding vector of hoof. The same principle also applies for more fine-grained visual disparities as we can see that stripes has higher co-occurrence probability with zebra than it has with horse.

3.3 Learning Visual Word Embeddings

In the previous paragraph, we showed how word co-occurrence statistics contain visually discriminative information, which explains why word embeddings visually discriminative high-level feature representations of visual classes for ZSL. However, word embeddings are learned in an unsupervised manner, without any visual information. In this paper, we conjecture that learning word embeddings with supervised supervision would yield more visually discriminative visual feature representations.

Our method works by embedding visual features and word representations in a shared embedding space. To do so, each visual class is first associated to a unique word label w (i.e. ”zebra” or ”pineapple”). We denote by $x_{w,i}$ the i-th image sample of class w. For readability, we refer by x_w to any randomly sampled image of class w. We then pretrain a Resnet-50 [6] on a subset of the Imagenet dataset that we use as training classes. We set the dimension of the top Resnet-50 layer to the word embeddings dimension $d = 300$. We use the activation values of this layer as feature representation of raw input images. Given a sample image x_w of class w, we denote by $f(x_w)$ the d-dimensional feature vector extracted as the CNN's top layer activation values.

The starting point of our method is equation (6c), which we first simplify by removing the logarithm and sigmoid forms. We then substitute the word embedding vectors with the visual feature of a randomly sampled image of the corresponding class:

$$\begin{aligned} L(T) &= \mathbb{E}_{w_j \sim p(j), c_i \sim p(i|j)} log(\sigma(w_j \cdot c_i)) \\ &\quad - k \times \mathbb{E}_{w_j \sim p(j), c_i \sim p(i)} log(\sigma(w_j \cdot c_i)) \end{aligned} \tag{7a}$$

$$\begin{aligned} L(T) &= \mathbb{E}_{w_j \sim p(j), c_i \sim p(i|j)} w_j \cdot c_i \\ &\quad - k \times \mathbb{E}_{w_j \sim p(j), c_i \sim p(i)} w_j \cdot c_i \end{aligned} \tag{7b}$$

$$\begin{aligned} L(T) &= \mathbb{E}_{w_j \sim p(j), c_i \sim p(i|j)} f(x_{w_j}) \cdot c_i \\ &\quad - k \times \mathbb{E}_{w_j \sim p(j), c_i \sim p(i)} f(x_{w_j}) \cdot c_i \end{aligned} \tag{7c}$$

Different from the original word2vec model, our model only learns one representation for each word, the context embedding vectors, as we substitute the word embedding term of the original model with the visual supervision signal. Our model learns the context vector representations c from a set of training word labels W_{Tr}. At test time, unseen visual class labels are computed as:

$$w_j = \mathbb{E}_{c_i \sim p(i|j)} c_i \tag{8a}$$

$$w_j = \sum_{i=0}^{n} p(i|j) c_i \tag{8b}$$

Let $W_{test} = \{w_j, j \in [1, N_{test}]\}$ be a set of N_{test} unseen test classes. An input test image x can be classified among W_{test} by retrieving the test class labels with highest similarity according to the dot-product similarity measure:

$$j^* = argmax_{w_j \in W_{test}} (w_j \cdot f(x)) \tag{9}$$

4. Experiments and results

We test our learned representations on the the task of large-scale ZSL. We use the validation split of the ILSVRC2012 classification dataset as test set, which consists of 1000 carefully curated classes for which 50 test images are given. We use the remaining 20,000 classes of the Imagenet dataset as training classes. As mentioned in the previous section, we use a Resnet-50 CNN pretrained on the training set to extract visual features from raw images.

Table 2. presents our results in terms of $top - k$ accuracy as is traditionally reported in ILSVRC challenges. Our results show that our model does learn visually discriminative visual representations as the accuracy scores are way above chance. For time constraints, we could not include comparisons to different models. In future work, we will present in-depth evaluation of our model together with comparisons with baseline models.

Table 2 top-k classification accuracy on the validation split of the ILSVRC2012 image classification dataset

	top-1	top-5	top-10
accuracy	3.1%	9.5%	21.0%

5. Conclusion

In this paper, we argued that visual class word labels share higher co-occurence probability with their visual attributes than they do with random words, which explains why word embeddings can provide visually discriminative representations of visual classes. Based on this idea, we modified the original word2vec algorithm to include visual supervision provided by visual features extracted from raw images by a pretrained CNN. Our model represents words and images in a shared embedding space so that the high level visual feature representations can be directly estimated in the visual space from the co-occurence statistics of its word label. Preliminary results showed that our learned representations are visually discriminative enough to allow for some zero-shot recognition of unseen visual classes. More experiments and in-depth evaluation of our model are still needed to validate our approach.

References

[1] Deng, Jia, et al. ”Imagenet: A large-scale hierarchical image database.” Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.Massachusetts (1993).

[2] http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/

[3] Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. In Advances in neural information pro- cessing systems, 21212129.

[4] Sahlgren, Magnus. ”The distributional hypothesis.” Italian Journal of Disability Studies 20 (2008): 33-53.

[5] Mikolov, Tomas, et al. ”Distributed representations of words and phrases and their compositionality.” Advances in neural information processing systems. 2013.

[6] He, Kaiming, et al. ”Deep residual learning for image recognition.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[7] Lampert, Christoph H., Hannes Nickisch, and Stefan Harmeling. ”Learning to detect unseen object classes by between-class attribute transfer.” Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[8] Wah, Catherine, et al. ”The caltech-ucsd birds-200-2011 dataset.” (2011).

[9] Mensink, T.; Gavves, E.; and Snoek, C. G. 2014. Costa: Co-occurrence statistics for zero-shot classification. In Pro- ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 24412448.

[10] Akata, Z.; Reed, S.; Walter, D.; Lee, H.; and Schiele, B. 2015. Evaluation of output embeddings for fine-grained im- age classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 29272936.

[11] Rohrbach, M.; Stark, M.; and Schiele, B. 2011. Evaluat- ing knowledge transfer and zero-shot learning in a large- scale setting. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 16411648. IEEE.

[12] Kiela, Douwe, and Lon Bottou. ”Learning image embeddings using convolutional neural networks for improved multi-modal semantics.” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.

[13] Lazaridou, Angeliki, Nghia The Pham, and Marco Baroni. ”Combining language and vision with a multimodal skip-gram model.” arXiv preprint arXiv:1501.02598 (2015).