Evaluation of Output Embeddings for Fine-Grained Image Classification

Zeynep Akata*, Scott Reed†, Daniel Walter†, Honglak Lee† and Bernt Schiele*
* Computer Vision and Multimodal Computing
Max Planck Institute for Informatics, Saarbrucken, Germany
† Computer Science and Engineering Division
University of Michigan, Ann Arbor

Image classification has advanced significantly in recent years with the availability of large-scale image sets. However, fine-grained classification remains a major challenge due to the annotation cost of large numbers of fine-grained categories. We show that compelling classification performance can be achieved on such categories even without labeled training data.

Following [1], given a specific input embedding, we derive a prediction by maximizing the compatibility F over SJE as follows:

$$f(x;W) = \arg\max_{y\in\mathcal{Y}} F(x,y;W) = \arg\max_{y\in\mathcal{Y}} \theta(x)^\top W \varphi(y).$$

where $\theta(x)$ is the input embedding and $\varphi(y)$ is the output embedding. The matrix W is learned by enforcing the correct label to be ranked higher than any of the other labels [7], the objective is:

$$\frac{1}{N}\sum_{n=1}^{N} \max_{y\in\mathcal{Y}} \{0, \ell(x_n, y_n, y)\}. \tag{1}$$

where $\ell(x_n,y_n,y) = \Delta(y_n,y) + \theta(x_n)^\top W\varphi(y) - \theta(x_n)^\top W\varphi(y_n)$. For zero-shot learning: we use $\varphi(y)$ of training classes and learn W. For prediction, we project $\theta(x)$ of test images onto the W and search for the nearest φ that corresponds to one of the test classes.

We use state-of-the-art image features [6] and focus on different supervised and unsupervised output embeddings described in the following:

Attributes ($\varphi^{0,1}$ & $\varphi^{\mathcal{A}}$ [3]) model shared characteristics of objects. For instance, for *rat, monkey, whale* and the attribute *big*, $\varphi^{0,1} = [0,0,1] \rightarrow$ *rat* = *monkey* < *whale*, whereas $\varphi^{\mathcal{A}} = [2,10,90] \rightarrow$ *rat* < *monkey* << *whale*.

Word2Vec ($\varphi^{\mathcal{W}}$ [4]) a two-layer neural network is trained to predict a set of target words from a set of context words. The first layer acts as a look-up table to retrieve the embedding for any word in the vocabulary. The second layer predicts the target word(s) via hierarchical soft-max. We use the skip-gram (SG) formulation where words within a local context window are predicted from the centering word.

GloVe ($\varphi^{\mathcal{G}}$ [5]) incorporates co-occurrence statistics of words that frequently appear together within the document. The objective is to learn word vectors such that their inner product equals the co-occurrence probability of these two words.

Weakly-supervised Word2Vec ($\varphi^{\mathcal{W}_{ws}}$) we pre-train the first layer weights using [4] on Wikipedia, and fine-tune the second layer weights using a negative-sampling objective [2] only on the fine-grained text corpus. These weights correspond to the final output embedding. The negative sampling objective is formulated as follows:

$$L = \sum_{w,c\in D_+} \log\sigma(v_c^T v_w) + \sum_{w',c\in D_-} \log\sigma(-v_c^T v_{w'}) \tag{2}$$
$$v_c = \sum_{i\in\text{context(w)}} v_i/|\text{context(w)}|$$

where v_w and $v_{w'}$ are the label embeddings we seek to learn, and v_c is the average of word embeddings v_i within a context window around word w. D_+ consists of context v_c and matching targets v_w, and D_- consists of the same v_c and mismatching $v_{w'}$.

Bag-of-Words ($\varphi^{\mathcal{B}}$) we collect Wikipedia articles that correspond to each object class and build a vocabulary of most frequently occurring words. We then build histograms of these words to vectorize our classes.

Hierarchies ($\varphi^{\mathcal{H}}$) we measure the similarity between two classes by estimating the distance between terms in an ontology such as `WordNet`.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

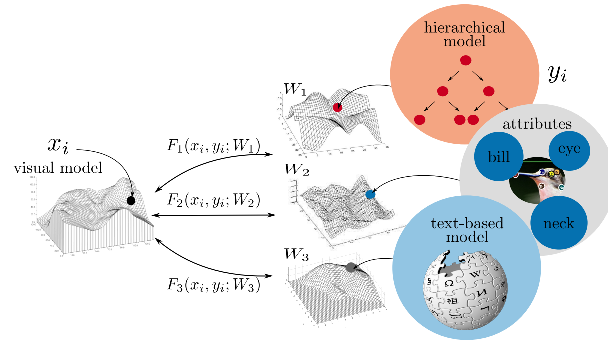

Figure 1: Structured Joint Embedding leverages images (x_i) and labels (y_i) by learning parameters W of a function $F(x_i, y_i, W)$ that measures the compatibility between input ($\theta(x_i)$) and output embeddings ($\varphi(y_i)$).

Combined embeddings to learn a better joint embedding we combine φ:

$$F(x,y;\{W\}_{1..K}) = \sum_k \alpha_k \theta(x)^\top W_k \varphi_k(y) \text{ s.t. } \sum_k \alpha_k = 1. \tag{3}$$

We emphasize the following take-home points: (1) Unsupervised label embeddings learned from text corpora yield compelling zero-shot results (Tab. 1), outperforming previous supervised SoA on AWA and CUB [1].

supervision	source	φ	AWA	CUB	Dogs
unsupervised	text	$\varphi^{\mathcal{W}}$	51.2	**28.4**	19.6
	text	$\varphi^{\mathcal{G}}$	**58.8**	24.2	17.8
	text	$\varphi^{\mathcal{B}}$	44.9	22.1	**33.0**
	WordNet	$\varphi^{\mathcal{H}}$	51.2	20.6	24.3
supervised	human	$\varphi^{0,1}$	52.0	37.8	-
	human	$\varphi^{\mathcal{A}}$	**66.7**	**50.1**	-

Table 1: Zero-shot learning results with SJE w.r.t. supervised and unsupervised output embeddings (Input embeddings: GoogLeNet [6])

(2) In combination, unsupervised output embeddings (w/o supervision) improve zero-shot performance, suggesting that they provide complementary information (Tab. 2).

supervision	method	AWA	CUB	Dogs
unsupervised	SJE (cmb)	60.1	29.9	**35.1**
supervised	SJE (cmb)	**73.9**	**51.7**	–
	SoA [1]	49.4	27.3	–

Table 2: Comparing SJE combined embeddings with SoA.

(3) There is still a large gap between the performance of unsupervised output embeddings and human-annotated attributes on AWA and CUB, suggesting that better methods are needed for learning discriminative output embeddings from text.

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *arXiv:1503.08677*, 2015.

[2] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv:1402.3722*, 2014.

[3] Christoph Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. In *TPAMI*, 2013.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[5] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[6] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[7] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 2005.