# Attribute-Based Classification with Label-Embedding

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, Cordelia Schmid

# Attribute-Based Classification with Label-Embedding

**Zeynep Akata***
Computer Vision Group
Xerox Research Centre Europe

**Florent Perronnin**
Computer Vision Group
Xerox Research Centre Europe

**Zaid Harchaoui**
LEAR group
INRIA

**Cordelia Schmid**
LEAR group
INRIA

## Abstract

Attributes are an intermediate representation whose purpose is to enable parameter sharing between classes, a must when training data is scarce. We propose to view attribute-based image classification as a label-embedding problem: each class is embedded in the space of attribute vectors. We introduce a function which measures the compatibility between an image and a label embedding. The parameters of this function are learned on a training set of labeled samples to ensure that, given an image, the correct class has a higher compatibility than the incorrect ones. Experimental results on two standard image classification datasets are presented, resp. on the Animals With Attributes and on Caltech-UCSD-Birds datasets.

## 1 Introduction

A solution to zero-shot learning [1, 2, 3, 4] that has recently gained in popularity in the computer vision community consists in introducing an intermediate space $\mathcal{A}$ referred to as *attribute* layer [3, 4]. Attributes correspond to high-level properties of the objects which are *shared* across multiple classes, which can be detected by machines and which can be understood by humans. As an example, if the classes correspond to animals, possible attributes include "has paws", "has stripes" or "is black". The traditional attribute-based prediction algorithm requires learning one classifier per attribute. To classify a new image, its attributes are predicted using the learned classifiers and the attribute scores are combined into class-level scores. This two-step strategy is referred to as Direct Attribute Prediction (DAP) in [3].

We note that DAP suffers from several shortcomings. First, a two-step prediction process goes against the philosophy which advocates solving a problem directly rather than indirectly through intermediate problems. Second, we would like an approach which can improve incrementally as new training samples are provided, i.e. which can perform zero-shot prediction if no labeled samples are available for some classes, but which can also leverage new labeled samples for these classes as they become available. Third, while attributes can be a useful source of prior information, other sources of information could be leveraged for zero-shot learning such as semantic hierarchies. Various improvements to DAP have been proposed to address each of these problems separately. However, we do not know of any existing solution which addresses all of them in a principled manner.

This paper proposes a general framework called *Attribute Label Embedding* (ALE). We embed each class $y$ in the space of attribute vectors and introduce a function $F$ which measures the "compatibility" between an image $x$ and a label $y$ (see Figure 1). The parameters of this function are learned on a training set of labeled samples to ensure that, given an image, the correct class has a higher

*Zeynep Akata is also with the LEAR team at INRIA

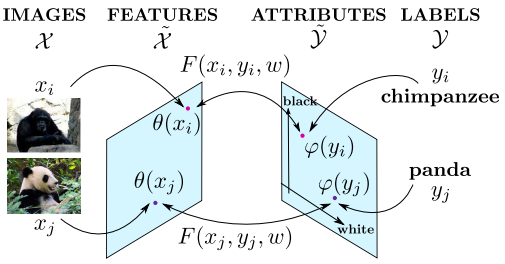


Figure 1: Much work in computer vision has been devoted to image embedding (left): how to extract suitable features from an image? We focus on *label embedding* (right): how to embed class labels in a Euclidean space? We use attributes as side information for the label embedding and measure the "compatibility"' between the embedded inputs and outputs with a function $F$.

compatibility than incorrect ones. Given a test image, recognition consists in searching for the class with the highest compatibility.

ALE addresses in a principled fashion all three problems mentioned previously. First, we do not solve any intermediate problem and learn the model parameters to optimize directly the class ranking. We show experimentally that ALE outperforms DAP in the zero-shot setting. Second, if available, labeled samples can be added incrementally to update the embedding. Third, the label embedding framework is generic and not restricted to attributes. Other sources of prior information can be combined with attributes.

The paper is organized as follows. We first introduce ALE. We then present experimental results on two public datasets: Animals with Attributes (AWA) [3] and Caltech-UCSD-Birds (CUB) [5]. This paper is a shortened version of [6].

## 2 Learning with attributes as label embedding

Given a training set $\mathcal{S} = \{(x_n, y_n), n = 1 \dots N\}$ of input/output pairs with $x_n \in \mathcal{X}$ and $y_n \in \mathcal{Y}$ the goal of prediction is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ by minimizing an empirical risk of the form $\frac{1}{N}\sum_{n=1}^{N} \Delta(y_n, f(x_n))$ where $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ measures the loss incurred from predicting $f(x)$ when the true label is $y$. In what follows, we focus on the 0/1 loss: $\Delta(y, z) = 0$ if $y = z$, 1 otherwise. In machine learning, a common strategy is to use *embedding functions* $\theta : \mathcal{X} \to \tilde{\mathcal{X}}$ and $\varphi : \mathcal{Y} \to \tilde{\mathcal{Y}}$ for the inputs and outputs and then to learn on the transformed input/output pairs.

In what follows, we first describe our model, *i.e.* our choice of $f$. We then explain how to leverage attributes to compute label embeddings. We also discuss how to learn the model parameters. Finally, we show that the label embedding framework is generic enough to accommodate for other sources of side information.

### 2.1 Model

Figure 1 illustrates our model. As is common in structured prediction [7], we introduce a compatibility function $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and define the prediction function $f$ as follows:

$$f(x; w) = \arg\max_{y \in \mathcal{Y}} F(x, y; w) \tag{1}$$

where $w$ denotes the model parameter vector of $F$ and $F(x, y; w)$ measures how compatible is the pair $(x, y)$ given $w$. It is generally assumed that $F$ is linear in some combined feature embedding of inputs/outputs $\psi(x, y)$:

$$F(x, y; w) = w' \psi(x, y) \tag{2}$$

and that the joint embedding $\psi$ can be written as the tensor product between the image embedding $\theta : \mathcal{X} \to \tilde{\mathcal{X}} = \mathbb{R}^D$ and the label embedding $\varphi : \mathcal{Y} \to \tilde{\mathcal{Y}} = \mathbb{R}^E$: $\psi(x, y) = \theta(x) \otimes \varphi(y)$ with $\psi : \mathbb{R}^D \times \mathbb{R}^E \to \mathbb{R}^{DE}$. In this case $w$ is a DE-dimensional vector which can be reshaped into a $D \times E$ matrix $W$. Consequently, we can rewrite $F(x, y; w)$ as a bilinear form:

$$F(x, y; W) = \theta(x)' W \varphi(y) \tag{3}$$

Other compatibility functions could have been considered. For example, the function: $F(x, y; W) = -||\theta(x)'W - \varphi(y)||^2$ is typically used in regression problems. If $D$ and $E$ are large, it might be

advantageous to consider a low-rank decomposition $W = U'V$ to reduce the number of parameters. In such a case, we have: $F(x, y; U, V) = (U\theta(x))' (V\varphi(y))$. For instance, CCA [8] or WSABIE [9] rely on such decompositions.

### 2.2 Attribute label embedding

We now consider the problem of computing label embeddings $\varphi^{\mathcal{A}}$ from attributes which we refer to as Attribute Label Embedding (ALE). We assume that we have $C$ classes, i.e. $\mathcal{Y} = \{1, \dots, C\}$ and that we have a set of $E$ attributes $\mathcal{A} = \{a_i, i = 1 \dots E\}$ to describe the classes. We also assume that we are provided with an association measure $\rho_{y,i}$ between each attribute $a_i$ and each class $y$. These associations may be binary or real-valued if we have information about the association strength. In this work, we focus on binary relevance although one advantage of the label embedding framework is that it can easily accommodate real-valued relevances. We embed class $y$ in the $E$-dim attribute space as follows:

$$\varphi^{\mathcal{A}}(y) = [\rho_{y,1}, \dots, \rho_{y,E}] \tag{4}$$

and denote $\Phi^{\mathcal{A}}$ the $E \times C$ matrix of attribute embeddings which stacks the individual $\varphi^{\mathcal{A}}(y)$'s. We note that in equation (3) the image and label embeddings play symmetric roles. It can make sense to normalize the output vectors $\varphi^{\mathcal{A}}(y)$. In the experiments, we use a binary $\{0, 1\}$ encoding of the attributes and $\ell_2$-normalize the class-embeddings.

### 2.3 Parameter learning

We now turn to the estimation of the model parameters $w$ from a labeled training set $\mathcal{S}$. The simplest learning strategy is to maximize directly the compatibility between the input and output embeddings $\frac{1}{N}\sum_{n=1}^{N} F(x_n, y_n; W)$, with potentially some constraints and regularizations on $W$. This is exactly the strategy adopted in regression or CCA. However, such an objective function does not optimize directly our end-goal which is image classification. Another possibility is to use a multi-class classification objective function as is the case in structured learning [7]. A third possibility is to use a ranking objective function which enforces the correct class to be ranked higher than incorrect ones [9]. We adopt the ranking strategy of the WSABIE algorithm [9] whose training procedure was shown to be particularly scalable.

Let $\ell(x_n, y_n, y) = \Delta(y_n, y) + F(x_n, y; W) - F(x_n, y_n; W)$ and let $r_\Delta(x_n, y_n) = \sum_{y \in \mathcal{Y}} \mathbb{1}(\ell(x_n, y_n, y) > 0)$ be an upper-bound on the rank of label $y_n$ for image $x_n$. WSABIE considers the following ranking objective:

$$R(\mathcal{S}; W, \Phi) = \frac{1}{N}\sum_{n=1}^{N} \gamma_{r_\Delta(x_n, y_n)} \sum_{y \in \mathcal{Y}} \max\{0, \ell(x_n, y_n, y)\} \tag{5}$$

where $\gamma_k$ is a decreasing function of $k$.

**Zero-shot learning.** We adapt the WSABIE objective to zero-shot learning. In such a case, we cannot learn $\Phi$ from labeled data (contrary to WSABIE) but rely on side information. Therefore, the matrix $\Phi$ is fixed and set to $\Phi^{\mathcal{A}}$. We only optimize the objective (5) with respect to $W$.

**Few-shots learning.** We now adapt the WSABIE objective to the case where we have labeled data and side information. In such a case, we want to learn the class embeddings using as prior information $\Phi^{\mathcal{A}}$. We therefore add to the objective (5) a regularizer:

$$R(\mathcal{S}; W, \Phi) + \frac{\mu}{2}||\Phi - \Phi^{\mathcal{A}}||^2 \tag{6}$$

and optimize jointly with respect to $W$ and $\Phi$.

### 2.4 Beyond attributes

While attributes make sense in the label embedding framework, we note that label embedding is more general and can accommodate for other sources of side information. The canonical example is that of structured learning with a taxonomy of classes [7]. Assuming that classes are organized in a

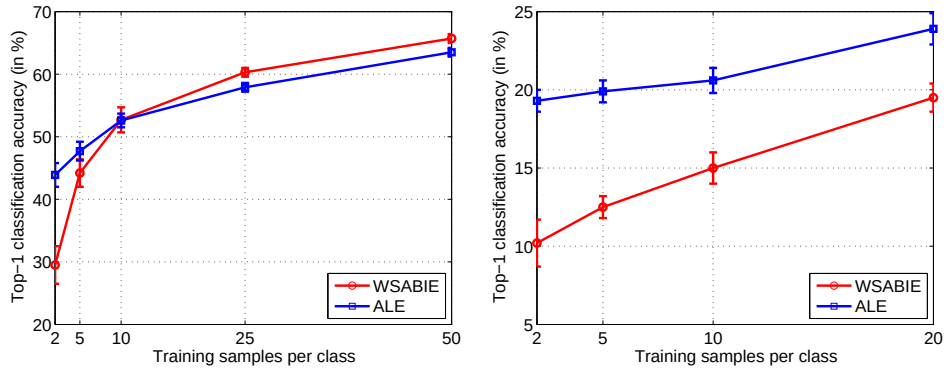


Figure 2: Classification accuracy on AWA (left) and CUB (right) as a function of the number of training samples per class.

tree structure, meaning that we have an ordering operation $\prec$ in $\mathcal{Y}$, we can define $\xi_{y,z} = 1$ if $z \prec y$ or $z = y$. The hierarchy embedding $\varphi^{\mathcal{H}}(y)$ can be defined as the $C$ dimensional vector:

$$\varphi^{\mathcal{H}}(y) = [\xi_{y,1}, \ldots, \xi_{y,C}]. \tag{7}$$

We later refer to this embedding as Hierarchy Label Embedding (HLE) and we compare $\varphi^{\mathcal{A}}$ and $\varphi^{\mathcal{H}}$ as sources of prior information in our experiments. In the case where classes are not organized in a tree structure but form a graph, then other types of embeddings could be used, for instance by performing a kernel PCA on the commute time kernel [10].

Different embeddings can be easily combined in the label embedding framework, *e.g.* through simple early fusion (*i.e.* concatenation) of the different embeddings.

## 3 Experiments

**Datasets.** We report results on two public datasets. Animal With Attributes (AWA) [3] contains roughly 30,000 images of 50 animal classes. Each class was annotated with 85 attributes by 10 students [11] and the result was binarized. CUB-200-2011 [5] contains roughly 11,800 images of 200 bird classes. Each class is annotated with 312 binary attributes derived from a bird field guide website. Hence, there is a significant difference in the number and quality of attributes between the two datasets. We report results in terms of top-1 accuracy (in %) averaged over the classes.

**Features.** We extract local descriptors and aggregate them into an image-level representation using the Fisher Vector (FV) framework which was shown to be a state-of-the-art patch encoding technique [12]. These FVs are our image embeddings $\theta(x)$.

**Zero-shot learning.** We first evaluate the proposed ALE in the zero-shot setting. For AWA, we use the standard zero-shot setup which consists in learning parameters on 40 classes and evaluating accuracy on 10 classes. For CUB, we use 150 classes for learning and 50 for evaluation.

On AWA and CUB, the DAP baselines are resp. 36.1% and 10.5% accuracy. With ALE, we obtain resp. 37.4% and 18.0%. We believe ALE yields superior results to DAP because it optimizes directly the classification end-goal. With HLE (*i.e.* using a class hierarchy as prior information instead of attributes), we obtain resp. 39.0% and 12.0%. When combining ALE and HLE through early fusion (concatenation of the embeddings), we can obtain an improvement on AWA to 43.5% but we did not observe any improvement on CUB.

**Few-shots learning.** We now assume that we have few (*e.g.* 2, 5, 10, etc.) training samples for a set of classes of interest (the 10 AWA and 50 CUB evaluation classes) in addition to all the samples from a set of "background" classes (the remaining 40 AWA amd 150 CUB classes). We compare the proposed ALE with WSABIE which performs parameter sharing through label embedding but which does not include prior information. The results of Figure 2 show that ALE is superior to WSABIE when relevant data is scarce, thus showing the importance of label embedding with prior information in this setting.

**Acknowledgments**

The Computer Vision Group at XRCE is partially funded by the ANR project FIRE-ID. The LEAR group at INRIA is partially funded by the European integrated project AXES.

## References

[1] H. Larochelle, D. Erhan, and Y. Bengio. Zero-data learning of new tasks. In *AAAI*, 2008. 1

[2] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009. 1

[3] Christoph Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 2, 4

[4] A Farhadi, I Endres, D Hoiem, and D Forsyth. Describing objects by their attributes. *CVPR*, 2009. 1

[5] Catherine Wah, Steve Branson, Pietro Perona, and Serge Belongie. Multiclass recognition and part localization with humans in the loop. In *ICCV*, 2011. 2, 4

[6] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013. 2

[7] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 2005. 2, 3

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning (2nd Ed.)*. Springer Series in Statistics. Springer, 2008. 3

[9] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: Learning to rank with joint word-image embeddings. *ECML*, 2010. 3

[10] Marco Saerens, Francois Fouss, Luh Yen, and Pierre Dupont. The principal components analysis of a graph, and its relationships to spectral clustering. In *ECML*, 2004. 4

[11] D. Osherson, J. Stern, O. Wilkie, M. Stob, and E. Smith. Default probability. *Cognitive Science*, 1991. 4

[12] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the Fisher vector: Theory and practice. *IJCV*, 2013. 4