An embarrassingly simple approach to zero-shot learning

Bernardino Romera-Paredes BERNARDINO.ROMERAPAREDES@ENG.OX.AC.UK
University of Oxford, Department of Engineering Science, Parks Road, Oxford, OX1 3PJ, UK

Philip H. S. Torr PHILIP.TORR@ENG.OX.AC.UK
University of Oxford, Department of Engineering Science, Parks Road, Oxford, OX1 3PJ, UK

Abstract

Zero-shot learning consists in learning how to recognise new concepts by just having a description of them. Many sophisticated approaches have been proposed to address the challenges this problem comprises. In this paper we describe a zero-shot learning approach that can be implemented in just one line of code, yet it is able to outperform state of the art approaches on standard datasets. The approach is based on a more general framework which models the relationships between features, attributes, and classes as a two linear layers network, where the weights of the top layer are not learned but are given by the environment. We further provide a learning bound on the generalisation error of this kind of approaches, by casting them as domain adaptation methods. In experiments carried out on three standard real datasets, we found that our approach is able to perform significantly better than the state of art on all of them, obtaining a ratio of improvement up to 17%.

1. Introduction

Automatic classification is arguably the first problem considered in machine learning, thus it has been thoroughly studied and analysed, leading to a wide variety of classification approaches which have been proved useful in many areas such as computer vision and document classification. However, these approaches cannot generally tackle challenging scenarios in which new classes may appear after the learning stage. We find this scenario in lots of real world situations. One happens when dealing with an ever growing set of classes, such as detecting new species of living beings. Another scenario occurs when the granularity of the description of the categories to be distinguished makes it unfeasible to obtain training instances for each of them, e.g. when a user wants to recognise a particular model of dress.

There is an increasing interest in the study of zero-shot learning (ZSL) approaches with the aim of solving this problem. ZSL consists in recognising new categories of instances without training examples, by providing a high-level description of the new categories that relate them to categories previously learned by the machine. This can be done by means of leveraging an intermediate level: the attributes that provide semantic information about the categories to classify. This paradigm is inspired by the way human beings are able to identify a new object by just reading a description of it, leveraging similarities between the description of the new object and previously learned concepts. Similarly, zero-shot learning approaches are designed to learn this intermediate semantic layer, the attributes, and apply them at inference time to predict new classes, provided with their description in terms of these attributes. Hereafter we use the term signature to refer to the attribute description of a class.

Zero-shot learning is inherently a two stage process: training and inference. In the training stage, knowledge about the attributes is captured, and in the inference stage this knowledge is used to categorise instances among a new set of classes. Many efforts have been made to improve the training stage (Hwang et al., 2011; Farhadi et al., 2009; Jayaraman et al., 2014), whereas the inference stage has received little attention (Jayaraman & Grauman, 2014). For example many approaches are unable to fully exploit the discriminative capacity of attributes, and cannot harness the uncertainty of the attribute prediction obtained in the first stage.

We study a framework that is able to integrate both stages, overcoming the general deficiencies previously described. This framework, introduced in (Akata et al., 2013), is based on modelling the relationship between features, attributes, and classes as a (linear) model composed of two layers.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

The first layer contains the weights that describe the relationship between the features and the attributes, and is learned at the training stage. The second layer models the relationship between the attributes and the classes and is fixed using the prescribed attribute signatures of the classes. Given that the training classes and the test classes are different, this second layer is interchangeable.

The main contributions of this paper are:

- An approach, based on the described framework and a principled choice of the regularizer, which has three nice properties: first, it performs comparably or better than the state of the art; second, it is efficient both at training and inference stage; and third, it is extremely easy to implement: one line of code for training and another one for inference (without calling any external functions).
- We provide a bound on the generalisation error of the approaches comprised in this framework. This is done by bridging the gap between zero-shot learning and domain adaptation.

The remainder of the paper is organised as follows. In Sec. 2 we briefly review methods proposed to deal with zero-shot learning. In Sec. 3 we describe the above ZSL framework, and present our method. In Sec. 4 we analyse its learning capabilities. In Sec. 5 we report the results of our experiments on one synthetic and three standard real datasets. Finally in Sec. 6 we discuss the main contributions of this paper and propose several research lines that can be explored.

2. Related work

Attributes learning, defined as the process of learning to recognise several properties from objects, precedes zero-shot learning. Indeed, it is the capability of predicting attributes from instances what drives the possibility of learning new classes based only on their description. The notion of using a description to represent a class dates back to (Dietterich & Bakiri, 1995). The aim was using these binary descriptors as error-correcting codes, although these did not convey any semantic meaning. Recently, there has been an increasing interest in automatic recognition of attributes, partially due to the availability of data containing tags or meta-information. This has proved to be particularly useful for images (Ferrari & Zisserman, 2007; Farhadi et al., 2009; Lampert et al., 2009), as well as videos (Fu et al., 2014; Liu et al., 2011).

Many papers focus on attributes learning, namely the training stage in zero-shot learning methods, putting special emphasis on the need to disentangle the correlations between attributes at the training set, because these properties may not be present in the target data (Jayaraman et al., 2014). For example in (Farhadi et al., 2009) the authors focus on the feature extraction process with the aim of avoiding confusion in the learning process of attributes that often appear together in the training set instances.

With regard to the inference stage in which the predicted attributes are combined to infer a class, many approaches employ 1-nearest neighbour, probabilistic frameworks, or modified versions of either.

Approaches based on 1-nearest neighbour consist in looking in the attribute space for the closest test class signature to the predicted attribute signature of the input instance. It is used in (Farhadi et al., 2009), and in (Palatucci et al., 2009) the authors study risk bounds of this approach when using the Hamming distances between the predicted signature and the signatures of the target classes. Whereas 1-nearest neighbour is an intuitive way for inferring classes from the attributes, it presents several drawbacks. Namely, it treats equally all dimensions of the attribute space, which may be sub-optimal, as some attributes are more important than others for discriminating between classes, and metrics such as Hamming distance ignore quantitative information in the prediction of the attributes.

In (Lampert et al., 2009; 2014) the authors propose a cascaded probabilistic framework in which the predictions obtained in the first stage can be combined to determine the most likely target class. Within this framework two approaches are proposed: directed attribute prediction (DAP), and indirected attribute prediction (IAP). In DAP a probabilistic classifier (e.g. logistic regression model) is learned at training stage for each attribute. At inference stage the previous estimators are used to infer the new classes provided their attributes signatures. In IAP one probabilistic classifier is learned for each training class, whereas at inference stage the predictions are combined accounting for the signatures of both training and test classes. The DAP approach has been widely used by many other methods. In (Suzuki et al., 2014) the authors extend DAP by weighting the importance of each attribute, based on its frequency of appearance. These probabilistic approaches bring a principled way of combining the attribute predictions of a new instance in order to infer its class. However, in addition of being unable to ponder the reliability of the predicted attributes, they introduce a set of independence assumptions that hardly ever hold in real world, for example, when describing animals the attributes “terrestrial” and “farm” are dependent, but are treated as independent in these approaches.

Very recently, the authors of (Jayaraman & Grauman, 2014) proposed an approach that acknowledges unreliability in the prediction of attributes, having mechanisms

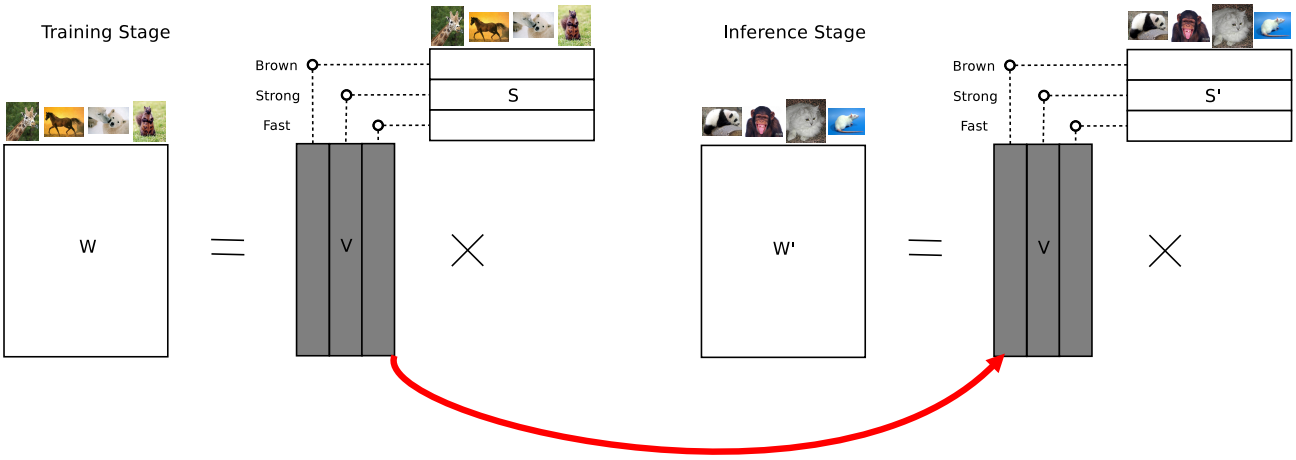

Figure 1. Summary of the framework described in Sec. 3. At training stage we use the matrix of signatures S together with the training instances to learn the matrix V (in grey) which maps from the feature space to the attribute space. At inference stage, we use that matrix V, together with the signatures of the test classes, S', to obtain the final linear model W'.

to deal with it. The approach is based on random forests that classify attribute signatures into the test classes, using a validation partition from the training set. The resultant model empirically proves to be superior to previous inference methods, such as DAP, and it obtains state of the art results in the benchmark datasets. One of the limitations of this model is the need to have the attribute signatures of the test classes at the training stage. In other words, the model learned at training stage is tailored to work with a predefined set of target classes.

The approach we describe in Sec. 3 bypasses the limitations of these methods by expressing a model based on an optimisation problem which relates features, attributes and classes. There are some works which follow a similar strategy. A relevant approach is the one described in (Akata et al., 2013), where the authors propose a model that implicitly learns the instances and the attributes embeddings onto a common space where the compatibility between any pair of them can be measured. The approach we describe in the paper is based on the same principle, however we use a different loss function and regularizer which not only makes the whole process simpler and efficient, but also leads to much better results. Another related approach is proposed in (Hariharan et al., 2012), where the authors use the information regarding the correlations between attributes in both training and test instances. The main differences are that they focus on attribute prediction, and they employ a max-margin formulation that leads to a more complex approach.

Other approaches consider the attributes as latent variables to be learned. For example in (Wang & Mori, 2010) an explicit feature map is designed to model the relationships between features, attributes and classes. Other approaches, such as (Liu et al., 2011; Mahajan et al., 2011), consider different schemes where attributes representations are to be learned.

The approach we describe is grounded on the machine learning areas of transfer learning and domain adaptation. Transfer learning, also known as learning to learn (Lawrence & Platt, 2004) or inductive transfer (Croonenborghs et al., 2008; Raykar et al., 2008; Rückert & Kramer, 2008), shares with zero-shot learning the aim of extracting knowledge from a set of source tasks that can be applied in future tasks. The main difference is that in transfer learning the information about the new tasks is given as a set of labelled instances. An extensive review of these methods can be found in (Pan & Yang, 2010).

The aim of domain adaptation is to learn a function from data in one domain, so that it can be successfully applied to data from a different domain (Ben-David et al., 2007; Daumé III, 2009; Jiang & Zhai, 2007). It resembles transfer learning but there are important differences to note. In transfer learning the input distribution in both source and target tasks is supposed to be the same whereas there are several labelling functions. Domain adaptation makes the reverse assumption, that is, the learning function is the same but the input distributions for source and target tasks are different. The link between our approach and domain adaptation becomes clear in Sec. 4.1.

3. Embarrassingly simple ZSL

Let us start by defining some notation. We assume that at training stage there are z classes, each of them having a signature composed of a attributes. We can represent these signatures in a matrix $S \in [0,1]^{a \times z}$. This matrix may con-

tain boolean entries, when the description of classes is defined as a list of attributes, or more generally, it may contain for each attribute any value in $[0, 1]$ providing a soft link between attributes and classes. Let us denote by $X \in \mathbb{R}^{d\times m}$ the instances available at training stage, where d is the dimensionality of the data, and m is the number of instances. Similarly we use $Y \in \{-1, 1\}^{m\times z}$ to denote the ground truth labels of each training instance belonging to any of the z classes. Note that both matrices Y and S provide enough information so that one can easily obtain the ground truth attributes for each instance. In most cases each row of Y contains only one positive entry indicating the class it belongs to. Nevertheless, the present framework allows an instance to belong to several classes simultaneously.

If we were interested in learning a linear predictor for the z training classes, we would optimise the following problem:

$$\underset{W\in\mathbb{R}^{d\times z}}{\text{minimise}}\, L\left(X^\top W, Y\right) + \Omega\left(W\right), \tag{1}$$

where W contains the parameters to be learned, L is a loss function, and Ω a regularizer. Problem (1) encompasses several approaches, depending on the choice of L and Ω. For example if L is the hinge loss, and Ω is the Frobenius norm, this would lead to a standard SVM, but one can consider other loss functions such as logistic loss, and other regularizers, such as the trace norm, leading to multitask learning methods (Argyriou et al., 2008; Romera-Paredes et al., 2013).

In problem (1) the attributes are not used, and therefore, there is no way to perform knowledge transfer from this set of classes to new classes. Instead, one can introduce the given information about the attributes, S, by replacing $W = VS^\top$, where $V \in \mathbb{R}^{d\times a}$. That leads to the following problem:

$$\underset{V\in\mathbb{R}^{d\times a}}{\text{minimise}}\, L\left(X^\top VS, Y\right) + \Omega\left(V\right). \tag{2}$$

At inference stage we want to distinguish between a new set of z' classes. To do so, we are provided with their attributes signatures, $S' \in [0, 1]^{a\times z'}$. Then, given a new instance, x, the prediction is given by

$$\underset{i}{\text{argmax}}\, x^\top VS'_i.$$

A scheme of this framework is shown in Fig. 1, and it is also used in (Akata et al., 2013). Unlike other zero-shot learning methods, the approach in eq. (2) does not try to minimise explicitly the classification error of the attributes. Instead it minimises the multiclass error, by both learning implicitly how to recognise attributes, and also pondering the importance of each of them in the decision of the class.

There are several points to note from problem (2). Firstly, if the regularizer Ω is of the form $\Omega(B) = \Psi\left(B^\top B\right)$ for some function Ψ, then by using the Representer theorem (Argyriou et al., 2009), it is straightforward to contemplate a kernel version of the problem, where only inner products between instances are used:

$$\underset{A\in\mathbb{R}^{m\times a}}{\text{minimise}}\, L\left(KAS, Y\right) + \Psi\left(S^\top A^\top KAS\right), \tag{3}$$

where $K \in \mathbb{R}^{m\times m}$ is the Gram matrix, $K_{i,j} = \langle\phi(x_i), \phi(x_j)\rangle$, being $\phi(x)$ the representation of x in a given feature space. Secondly, problem (2) and its equivalent problem (3) are convex, and its global optimal solution can be computed efficiently.

3.1. Regularisation and loss function choices

The framework described above comprises several approaches, which vary depending on their regularizer and loss function. Here we design a regularizer which accomplishes the following desiderata:

- The Euclidean norm of the representation of any (training) attribute signature, $s \in [0, 1]^a$, on the feature space, Vs, must be controlled so that ideally the representation of all signatures on the feature space have a similar Euclidean norm. This allows fair comparisons between signatures, and prevents problems that stem from highly unbalanced training sets.
- Conversely, it would be interesting to bound the variance of the representation of all (training) instances X on the attribute space, $V^\top X$. The aim of this point is to make an approach that is invariant enough to the training feature distribution, so that it can generalise to other test feature distributions.

A regularizer that accomplishes the previous points can be written as follows:

$$\Omega\left(V; S, X\right) = \gamma\left\|VS\right\|^2_{\text{Fro}} + \lambda\left\|X^\top V\right\|^2_{\text{Fro}} + \beta\left\|V\right\|^2_{\text{Fro}}, \tag{4}$$

where the scalars γ, λ, β are the hyper-parameters of this regularizer, and $\|\cdot\|_{\text{Fro}}$ denotes the Frobenius norm. The first two terms account for the above points, and we have added one further term consisting in a standard weight decay penalising the Frobenius norm of the matrix to be learned.

Having made these choices, we note that if:

- $L\left(P, Y\right) = \left\|P - Y\right\|^2_{\text{Fro}}$.
- $\beta = \gamma\lambda$

then the solution to problem (2) can be expressed in closed form:

$$V = \left(XX^\top + \gamma I\right)^{-1} XYS^\top \left(SS^\top + \lambda I\right)^{-1}. \tag{5}$$

This, and the corresponding kernel version that can be derived from eq. (3), are the one-line-of-code solutions we mentioned in the introduction.

4. Risk bounds

In this section we provide some theoretical guarantees about our approach, bounding the expected error on the inference stage with respect to the training error. In order to do so, we first transform our problem into a domain adaptation one.

4.1. Simple ZSL as a domain adaptation problem

Let us assume that problem (2) can be expressed in the following way:

$$\underset{V\in\mathbb{R}^{d\times a}}{\text{minimise}} \sum_{i=1}^{m}\sum_{t=1}^{z} \ell\left(x_i^\top V s_t^\top, y_{t,i}\right) + \Omega\left(V\right), \tag{6}$$

where $\ell\left(\cdot,\cdot\right) : \mathbb{R}\times\{-1,1\} \longrightarrow [0,1]$. That implies that one instance may be classified to belong to zero, one, or more than one classes. Such an assumption may be realistic in some cases, for example when there are some instances in the training set that do not belong to any training class. Then, problem (6) can be expressed in a more conventional form:

$$\underset{v\in\mathbb{R}^{da}}{\text{minimise}} \sum_{i=1}^{m}\sum_{t=1}^{T} \ell\left(\widetilde{x}_{t,i}^\top v, y_{t,i}\right) + \Omega\left(v\right), \tag{7}$$

where

$$\widetilde{x}_{t,i} = \text{vec}\left(x_i s_t^\top\right) \in \mathbb{R}^{da}. \tag{8}$$

Note that at inference time, given a new instance, x', the predicted confidence of it belonging to a new class t with attribute signature s_t, is given by $\widetilde{x}_t'^\top v = \text{vec}\left(x' s_t^\top\right)^\top v$. Therefore, even if the original test instances x were sampled from the same distribution as the training instances, the transformation of them using attributes signatures makes the training and test instances come from different distributions. As a consequence, we are facing a domain adaptation problem.

4.2. Risk bounds for domain adaptation

Domain adaptation has been analysed from a theoretical viewpoint in several works (Ben-David et al., 2007; Blitzer et al., 2008). Here we apply these developments to our problem.

In a domain adaptation problem we assume that the training instances are sampled from a source distribution $\mathcal{D}$, and the test instances are sampled from a target distribution $\mathcal{D}'$. Following the definition of (Ben-David et al., 2007), a function h is said to be a predictor if it maps from the feature space to $\{0,1\}$, and f is the (stochastic) ground truth labelling function for both domains, mapping from the feature space to $[0,1]$. Then the expected error of h with respect to the source distribution is defined as:

$$\epsilon(h) = \mathbb{E}_{x\sim\mathcal{D}}\left[|f(x) - h(x)|\right],$$

and the expected error of h with respect to the target distribution, $\epsilon'(h)$, is defined accordingly.

Theorem 1 in (Blitzer et al., 2008) states that given a hypothesis space $\mathcal{H}$ of VC-dimension $\bar{d}$, and sets $\mathcal{U}, \mathcal{U}'$ of $\bar{m}$ instances sampled i.i.d. from $\mathcal{D}$ and $\mathcal{D}'$ respectively, then with probability at least $1-\delta$, for every $h\in\mathcal{H}$:

$$\epsilon'(h) \leq \epsilon(h) + 4\sqrt{\frac{2\bar{d}}{\bar{m}}\left(\log\frac{2\bar{m}}{\bar{d}} + \log\frac{4}{\bar{\delta}}\right)} \tag{9}$$

$$+ \lambda + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}\left(\mathcal{U},\mathcal{U}'\right), \tag{10}$$

where

- λ is an upper-bound of $\inf_{h\in\mathcal{H}}\left[\epsilon(h) + \epsilon'(h)\right]$. In particular if the ground truth function f is contained in $\mathcal{H}$, then $\lambda = 0$.
- $d_{\mathcal{H}}\left(\mathcal{D},\mathcal{D}'\right)$ is known as the $\mathcal{A}$-distance between distributions $\mathcal{D}$ and $\mathcal{D}'$ over the subsets defined in $\mathcal{H}$ (Kifer et al., 2004):

  $$d_{\mathcal{H}}\left(\mathcal{D},\mathcal{D}'\right) = 2\sup_{h\in\mathcal{H}}|P_{\mathcal{D}}(h) - P_{\mathcal{D}'}(h)|,$$

  where $P_{\mathcal{D}}(h)$ denotes the probability of any event in h, under the distribution $\mathcal{D}$. This is equivalent to the expected maximal accuracy achieved by a hypothesis in $\mathcal{H}$ separating the instances generated by the two different distributions $\mathcal{D}$ and $\mathcal{D}'$. In a similar vein, $\hat{d}_{\mathcal{H}}\left(\mathcal{U}_S,\mathcal{U}_T\right)$ is defined as the empirical distance between the samples $\mathcal{U}$ and $\mathcal{U}'$.
- $\mathcal{H}\Delta\mathcal{H}$ is the symmetric difference hypothesis space of $\mathcal{H}$ and it is defined as:

  $\mathcal{H}\Delta\mathcal{H} = \{h(x)\oplus h'(x) : h,h'\in\mathcal{H}\}$,

  being $\oplus$ the XOR operator. That is, a hypothesis f is in $\mathcal{H}\Delta\mathcal{H}$, if for a couple of hypothesis h,h' in $\mathcal{H}$, $f(x)$ is positive if and only if $h(x)\neq h'(x)$ for all x.

In our case $\mathcal{H}$ is the hypothesis space composed of all linear classifiers, $\bar{m} = mz$, and $\bar{d} = da+1$. Let us assume that both train and test instances are sampled from the same distribution, $\mathcal{C}$. When we do the transformation specified in eq. (8) using S and S' for the training and test

instances, we end up having two different distributions, $\mathcal{D}$, and $\mathcal{D}'$ and we are interested in quantifying the $\mathcal{A}$-distance between them over our symmetric difference hypothesis space, $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}, \mathcal{D}')$. The previous assumption may not hold true in many cases, however it can be a fair approximation in the standard case where the contribution of the differences of training and test distributions of the feature spaces is negligible in comparison to the differences between S and S' when quantifying the distance between distributions $\mathcal{D}$ and $\mathcal{D}'$.

We observe two extreme cases. The first one contemplates the trivial scenario where $S = S'$, so that both distributions are similar and thus the distance is 0. In that case, if $\lambda = 0$, the bound given in eq. (10) becomes equivalent to the Vapnik-Chervonenkis bound on a standard classifier. The second case arises when each attribute signature of the training classes is orthogonal to each attribute signature of the test classes, that is, for each $i \in \{1 \dots z\}$, $j \in \{1 \dots z'\}$, $\langle s_i, s'_j \rangle = 0$. In Section A of the supplementary appendix we prove that, in this case, the right hand side term in eq. (10) becomes bigger than one. Thus, the bound is vacuous, implying that no transfer can be done as one would expect. All real scenarios lay between the previous cases. One interesting question is to characterise the value $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}, \mathcal{D}')$ as a function of solely S and S'. We leave this question open in the present paper.

5. Experiments

In order to assess our approach and the validity of the statements we made, we conducted a set of experiments on one synthetic and three real datasets, which comprise a standard benchmark of evaluation of zero-shot learning methods[1].

5.1. Synthetic experiments

First we used synthetic generated data with the aim of both checking the correctness of the described method, which we refer to as ESZSL (embarrassingly simple zero-shot learning), and comparing it with the baseline algorithm DAP on a controlled set up. All hyper-parameters required by these methods were tuned by a validation process using in all cases the range of values 10^b, for $b = -6, -5, \dots, 5, 6$. This set of values was chosen after performing preliminary experiments which empirically showed that the optimal performance for both approaches is found within this interval.

The data were generated as follows. Initially, we created the signatures for the classes by sampling each element of S from a Bernoulli distribution with 0.5 mean. We created the ground truth mapping from the attributes to the features, $V^+ \in \mathbb{R}^{a \times d}$, where we have fixed $a = 100$ and $d = 10$, by sampling every element of it from a Gaussian distribution $\mathcal{G}(0, 1)$. The value of d is intentionally low so that there appear correlations between the attributes, as is usually the case in real data. For each class t, we created 50 instances by first generating their representation in the attribute space by adding Gaussian noise, $\mathcal{G}(0, 0.1)$ to the attribute signature S_t, then we brought them back onto the original feature space by using V^+. Following this process, we generated a training set composed of z classes, and a test and validation set composed of 100 classes each.

[1]The code can be found at `romera-paredes.com/zsl`.

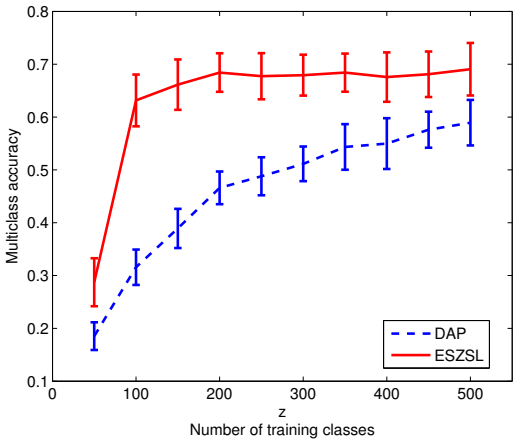

Figure 2. Multiclass accuracy obtained by DAP (Lampert et al., 2009), and ESZSL (Sec. 3.1), when varying the number of training classes, z. Vertical bars indicate ± 1 standard deviation.

In the first experiment, we evaluated how the number of training classes affected the performance of the methods on new classes. To do so, we varied the number of training classes from 50 to 500 in intervals of 50. According to the results shown in Fig 2, we can see that ESZSL significantly outperforms DAP in all cases. It is remarkable that the performance of ESZSL with 100 training classes is superior to the performance of DAP with 500 training classes. We also observe that the performance of ESZSL plateaus when the number of training classes is above 200, possibly because there is no further margin of improvement.

In Sec. 3 we argue that the described approach should be robust to attributes having different discriminative capabilities for characterising the classes. In the second experiment, we assess how the approaches perform in the extreme case where some attributes provide no information at all about the classes at hand. The way we have implemented this is by first, synthesising a dataset just as described above, and second, by randomly selecting a set of ψ attributes (without replacement) so that their information in all signatures is tweaked. The way each of the inputs

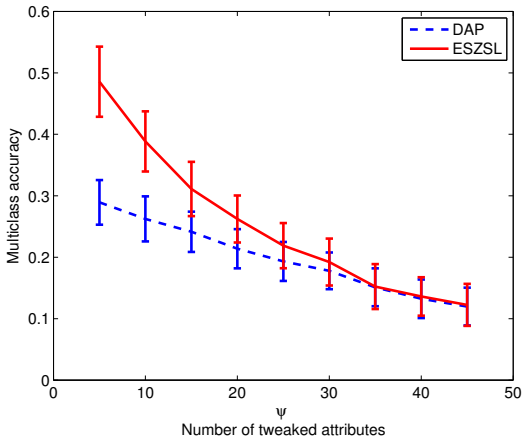

Figure 3. Multiclass accuracy obtained by DAP (Lampert et al., 2009), and ESZSL (Sec. 3.1), when varying the number of corrupted attributes, ψ. Vertical bars indicate ± 1 standard deviation.

of the modified attributes is tweaked is again by sampling from a Bernoulli distribution with 0.5 mean. In this experiments we have tried different values of ψ in the range of 5 to 45 attributes (out of 100), in intervals of 5. The results, reported in Fig 3, support our hypothesis about the robustness of our approach to filter meaningless attributes.

Additional experiments on synthetic data are reported in Sec. B of the appendix.

5.2. Real data experiments

We have tried the same real datasets as the ones reported in (Jayaraman & Grauman, 2014) which are the Animals with Attributes dataset (AwA) (Lampert et al., 2009), the SUN scene attributes database (SUN) (Patterson & Hays, 2012), and the aPascal/aYahoo objects dataset (aPY) (Farhadi et al., 2009). These consist of collections of images comprising a varied set of categories in different scopes: animals, scenes, and objects respectively. AwA dataset contains attribute-labelled classes, which we will use as S in the model. The datasets aPY and SUN are attribute-labelled instances datasets, so the attribute signature of each class is calculated as the average attribute signature of the instances belonging to that class. The characteristics of each of these datasets are summarised in table 1.

In the following we perform two sets of experiments. In the first one, we compare our approach with alike methods that also belong to the framework described in Fig. 1. In the second set of experiments, we compare our approach against the current state of the art. In all cases, in order to tune the hyper-parameters of the methods, we use the following validation procedure. We create the validation set by grouping all instances belonging to 20% of the training classes chosen at random (without replacement). Once the hyper-parameters are tuned, we pool the validation set instances together with the training set instances in order to train the final model. We use the range of values, 10^b for $b = -3, -2, \ldots, 2, 3$ to tune all hyper-parameters.

	AwA	aPY	SUN
Attributes	85	65	102
Training classes	40	20	707
Test classes	10	12	10
Instances	30475	15339	14340

Table 1. Summary of the real datasets employed in the experimental section.

Training instances	(Akata et al., 2013)	ESZSL
500	32.30%	**33.09%**
1000	38.57%	**42.44%**
2000	40.21%	**44.82%**

Table 2. Comparison between the approach in (Akata et al., 2013) and ESZSL, using the AwA dataset.

Preliminary experiments Here we present an experiment comparing our approach to (Akata et al., 2013). We used the recently provided DECAF features of the AwA dataset. We utilised the best configuration reported on (Akata et al., 2013), using different training set sizes of 500, 1000, and 2000 instances. The results, shown in Table 2 show that our approach clearly outperforms (Akata et al., 2013). It is also worth mentioning that the latter approach took more than 11 hours to run the scenario with 2000 training instances, whereas ours only took 4.12 seconds.

One of the main differences between our approach and that of (Akata et al., 2013), is that we use a more elaborated regularizer. Because of that, in the second preliminary experiment we aim to assess the importance of regularising the representation of both attributes and instances in each other's space, as done in eq. (4). In order to do so, we compare this regularizer with a version where only the Frobenius norm of the weights is penalised, that is, where $\gamma = \lambda = 0$ in eq. (4). For this experiment we use the DECAF features as before, and all training instances. ESZSL achieved 50.37% classification performance, whereas the described modified version obtained 45.02%. This supports the hypothesis that the use of these two regularizers leads to a critical difference in the performance.

Comparison with the state of the art In order to make our approach easily comparable with the state of the art, we

Method/Dataset	AwA	aPY	SUN
DAP	40.50	18.12	52.50
ZSRwUA	43.01 ±0.07	26.02 ±0.05	56.18 ±0.27
ESZSL	**49.30 ±0.21**	15.11 ±2.24	**65.75 ±0.51**
ESZSL-AS	–	**27.27 ±1.62**	61.53 ±1.03

Table 3. Multiclass accuracy obtained by DAP (Lampert et al., 2009), ZSRwUA (Jayaraman & Grauman, 2014), the method described in Sec. 3.1 ESZSL, and its modification ESZSL-AS, on the three real datasets described in Table 1.

used the set of standard features provided by the authors of the data (Jayaraman & Grauman, 2014; Lampert et al., 2009; Patterson & Hays, 2012), including SIFT (Lowe, 2004), and PHOG (Bosch et al., 2007). We used combined χ^2-kernels, one for each feature channel[2], following the procedure explained in (Lampert et al., 2009; Jayaraman & Grauman, 2014). In all cases, we used the same attributes signatures, and the same standard partitions between train and test classes, as the ones employed in (Jayaraman & Grauman, 2014).

In these experiments we compare 4 methods: DAP (Lampert et al., 2009), ZSRwUA (Jayaraman & Grauman, 2014), ESZSL (Sec. 3.1), and a small modification of the latter that we call ESZSL All Signatures (ESZSL-AS).

ESZSL-AS can be applied in attribute-labelled instances datasets (aPY and SUN), and consists in treating each training attribute signature as a class in its own right. That is effectively done by removing Y in eq. (5), where now $S \in \mathbb{R}^{a \times m}$ contains as many signatures as the number of training instances. The inference process remains the same, and the class signatures are used to predict the category.

For each dataset we ran 20 trials, and we report the mean and the standard deviation of the multiclass accuracy in Table 3. Overall we notice that the approaches described in Sec. 3 significantly outperform the state of the art.

In the AwA dataset, ESZSL achieves an improvement over 14.6% over the state of the art. Even more surprising, this performance is better than state of the art approaches applied to discovered (non-semantic) attributes, which according to (Jayaraman & Grauman, 2014) is 48.7. Let us recall that this dataset contains attribute-labelled classes, and so, ESZSL-AS cannot be applied here.

Regarding the aPY dataset, the standard ESZSL approach has struggled and it is not able to outperform the DAP baseline. One hypothesis is that the small number of classes in comparison to the number of attributes has probably affected negatively the performance. In contrast we see that ESZSL-AS obtains state of the art results, achieving a 4.8% of improvement over the previous best approach. Its success can be explained by reversing the previous reasoning about why standard ESZSL failed. Indeed, ESZSL-AS effectively considers as many training classes as the number of training instances.

Finally, in the SUN dataset both ESZSL approaches obtain extremely good results, significantly outperforming the current state of the art. ESZSL leads the table, achieving an improvement ratio of 17%. We note that here the number of training classes is much bigger than the number of attributes, therefore the advantages obtained by ESZSL-AS in the previous experiment vanish.

In Sec. C in the appendix, we report further experiments carried out on real data.

6. Discussion

In this paper we have described an extremely simple approach for ZSL that is able to outperform by a significant margin the current state of the art approaches on a standard collection of ZSL datasets. It combines a linear model together with a principled choice of regularizers that allow for a simple and efficient implementation.

We have also made explicit a connection between ZSL and domain adaptation. In particular, we have expressed the framework described in Sec. 3 as a domain adaptation problem. As a consequence, we are able to translate theoretical developments from domain adaptation to ZSL.

Given the simplicity of the approach, there are many different research lines that can be pursued. In this work we focus on semantically meaningful attributes, but the development of similar ideas applied to word embeddings as in (Frome et al., 2013), is both promising and straightforward within this framework. Another interesting research line is to study the addition of non-linearities and more layers into the model, leading to a deep neural network where the top layer is fixed and interchangeable, and all the remaining layers are learned.

As a concluding comment, we acknowledge that many problems require complex solutions, but that does not mean that simple baselines should be ignored. On the contrary, simple but strong baselines both bring light about which paths to follow in order to build more sophisticated solutions, and also provide a way to measure the quality of these solutions.

[2]Available at `www.ist.ac.at/~chl/ABC`.

Acknowledgements

Financial support provided by EPSRC, Leverhulme Trust and ERC grants ERC- 2012-AdG 321162-HELIOS and HELIOS-DFR00200.

We thank Prof. Lampert, and Dinesh Jayaraman for kindly providing the real datasets used in this paper.

References

Akata, Zeynep, Perronnin, Florent, Harchaoui, Zaid, and Schmid, Cordelia. Label-embedding for attribute-based classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 819–826. IEEE, 2013.

Argyriou, Andreas, Evgeniou, Theodoros, and Pontil, Massimiliano. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

Argyriou, Andreas, Micchelli, Charles A, and Pontil, Massimiliano. When is there a representer theorem? vector versus matrix regularizers. *The Journal of Machine Learning Research*, 10:2507–2529, 2009.

Ben-David, Shai, Blitzer, John, Crammer, Koby, Pereira, Fernando, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

Blitzer, John, Crammer, Koby, Kulesza, Alex, Pereira, Fernando, and Wortman, Jennifer. Learning bounds for domain adaptation. In *Advances in neural information processing systems*, pp. 129–136, 2008.

Bosch, Anna, Zisserman, Andrew, and Munoz, Xavier. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pp. 401–408. ACM, 2007.

Croonenborghs, Tom, Driessens, Kurt, and Bruynooghe, Maurice. Learning relational options for inductive transfer in relational reinforcement learning. In *Inductive Logic Programming*, pp. 88–97. Springer, 2008.

Daumé III, Hal. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

Dietterich, Thomas G. and Bakiri, Ghulum. Solving multiclass learning problems via error-correcting output codes. *arXiv preprint cs/9501101*, 1995.

Farhadi, Ali, Endres, Ian, Hoiem, Derek, and Forsyth, David. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1778–1785. IEEE, 2009.

Ferrari, Vittorio and Zisserman, Andrew. Learning visual attributes. In *Advances in Neural Information Processing Systems*, pp. 433–440, 2007.

Frome, Andrea, Corrado, Greg S, Shlens, Jon, Bengio, Samy, Dean, Jeff, Mikolov, Tomas, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pp. 2121–2129, 2013.

Fu, Yanwei, Hospedales, Timothy M, Xiang, Tao, and Gong, Shaogang. Learning multimodal latent attributes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(2):303–316, 2014.

Hariharan, Bharath, Vishwanathan, SVN, and Varma, Manik. Efficient max-margin multi-label classification with applications to zero-shot learning. *Machine learning*, 88(1-2):127–155, 2012.

Hwang, Sung Ju, Sha, Fei, and Grauman, Kristen. Sharing features between objects and their attributes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1761–1768. IEEE, 2011.

Jayaraman, Dinesh and Grauman, Kristen. Zero-shot recognition with unreliable attributes. pp. 3464–3472, 2014.

Jayaraman, Dinesh, Sha, Fei, and Grauman, Kristen. Decorrelating semantic visual attributes by resisting the urge to share. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 1629–1636. IEEE, 2014.

Jiang, Jing and Zhai, ChengXiang. Instance weighting for domain adaptation in nlp. In *ACL*, volume 7, pp. 264–271, 2007.

Kifer, Daniel, Ben-David, Shai, and Gehrke, Johannes. Detecting change in data streams. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pp. 180–191. VLDB Endowment, 2004.

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

Lampert, Christoph H, Nickisch, Hannes, and Harmeling, Stefan. Attribute-based classification for zero-shot visual object categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(3):453–465, 2014.

Lawrence, Neil D and Platt, John C. Learning to learn with the informative vector machine. In *Proceedings of the*

twenty-first international conference on Machine learning, pp. 65. ACM, 2004.

Liu, Jingen, Kuipers, Benjamin, and Savarese, Silvio. Recognizing human actions by attributes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3337–3344. IEEE, 2011.

Lowe, David G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

Mahajan, Dhruv, Sellamanickam, Sundararajan, and Nair, Vinod. A joint learning framework for attribute models and object descriptions. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 1227–1234. IEEE, 2011.

Palatucci, Mark, Hinton, Geoffrey, Pomerleau, Dean, and Mitchell, Tom M. Zero-Shot Learning with Semantic Output Codes. *Neural Information Processing Systems*, pp. 1–9, 2009.

Pan, Sinno Jialin and Yang, Qiang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

Patterson, Genevieve and Hays, James. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2751–2758. IEEE, 2012.

Raykar, Vikas C, Krishnapuram, Balaji, Bi, Jinbo, Dundar, Murat, and Rao, R Bharat. Bayesian multiple instance learning: automatic feature selection and inductive transfer. In *Proceedings of the 25th international conference on Machine learning*, pp. 808–815. ACM, 2008.

Romera-Paredes, Bernardino, Aung, Hane, Bianchi-Berthouze, Nadia, and Pontil, Massimiliano. Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1444–1452, 2013.

Rückert, Ulrich and Kramer, Stefan. Kernel-based inductive transfer. In *Machine Learning and Knowledge Discovery in Databases*, pp. 220–233. Springer, 2008.

Suzuki, Masahiro, Sato, Haruhiko, Oyama, Satoshi, and Kurihara, Masahito. Transfer learning based on the observation probability of each attribute. In *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, pp. 3627–3631. IEEE, 2014.

Wang, Yang and Mori, Greg. A discriminative latent model of object classes and attributes. In *Computer Vision–ECCV 2010*, pp. 155–168. Springer, 2010.