Augmented Attribute Representations

Viktoriia Sharmanska[1], Novi Quadrianto[2], and Christoph H. Lampert[1]

[1] IST Austria, Klosterneuburg, Austria
[2] University of Cambridge, Cambridge, UK

Abstract. We propose a new learning method to infer a mid-level feature representation that combines the advantage of semantic attribute representations with the higher expressive power of non-semantic features. The idea lies in augmenting an existing attribute-based representation with additional dimensions for which an autoencoder model is coupled with a large-margin principle. This construction allows a smooth transition between the zero-shot regime with no training example, the unsupervised regime with training examples but without class labels, and the supervised regime with training examples and with class labels. The resulting optimization problem can be solved efficiently, because several of the necessity steps have closed-form solutions. Through extensive experiments we show that the augmented representation achieves better results in terms of object categorization accuracy than the semantic representation alone.

Key words: Discriminative Autoencoder, Hybrid Representations

1 Introduction

Representations in terms of semantic attribute have recently gained popularity in computer vision, where they were used mainly for two different tasks: to solve classification problems based on class descriptions instead of training data (zero-shot learning) [1, 2], and to automatically create (textual) descriptions of images [3, 4]. In this work we build on the first of these aspects and we extend it transparently to the case when few training examples are given (*small shot*), either with class annotation (*supervised*), or without it (*unsupervised*). The underlying idea is to extend the attribute representation with additional mid-level features, which are not necessarily semantic by themselves, but that augment the semantic features minimally in the sense that they offer additional representative power where necessary, and only there.

Figure 1 illustrates this concept: assume we are given semantic representations $(a_1, \ldots, a_5)$ for three object classes, *zebra*, *white tiger* and *elephant*. As zebras and white tigers differ only in one entry in this representation, they will easily be confused when performing zero-shot classification with an imperfect, image-based attribute predictor. The representation of elephants, on the other hand, is clearly distinct from the other two classes, and classification errors are unlikely for them. The topic of our work is to reduce the total risk of misclassifications by learning an augmentation of the attributes with features $(b_1, \ldots, b_3)$

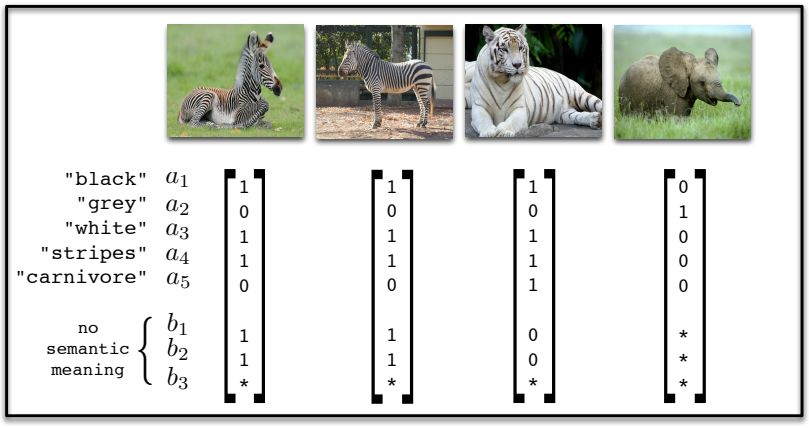

Fig. 1. Proposed hybrid representation: a fixed semantic $(a_1, \ldots, a_5)$ part is augmented by a non-semantic $(b_1, \ldots, b_3)$ part, where the latter is learned by enforcing a large margin separation criterion between classes. See Section 1 for a detailed description.

that are learned automatically, even if this causes them to not necessarily be semantic anymore. Specifically, we obtain values $(b_1, \ldots, b_3)$ for each image by enforcing a large-margin criterion: the distance between representations of any pair of images of different classes should differ by at least a constant (here 3). As a result, different values are chosen for (b_1, b_2) for the zebra images than for the white tiger image. For the elephant, the semantic representation alone is already sufficient to enforce the distance criterion to the other classes. Therefore, no specific values for (b_1, b_2) are enforced. Similarly, the value of b_3 can be chosen freely for all examples, which allows satisfying other criteria of the model, in our case a reconstruction criterion. Note that, contrary to the above description, the method we propose learns all entries of b jointly, not by an iterative reasoning as used above for the purpose of illustration.

To implement the above intuition, we rely on two recently successful concepts for learning of image representations: the *autoencoder* framework [5] and the *large margin* concept [6]. The autoencoders follow a generative approach to learning an intermediate representation by identifying features that allow reconstruction of the image representation with only a small error. In the large margin nearest neighbor framework, we learn a representation in a discriminative way by trying to reduce the nearest neighbor classification on a training set in a robust way. In the rest of the manuscript, we formalize these concepts and formulate them as a joint optimization problem over projection matrices. We show how to solve the optimization problem using alternating optimization in which some parts have efficient closed form solutions. We perform an experimental evaluation on the Animals with Attribute dataset that shows that the learned hybrid representations improve over the representation purely in terms of semantic attributes when additional training data is available.

2 Learning to Augment Features

For the rest of the manuscript we will assume the following situation: we are given N images in a d-dimensional feature representation, $x_1, \ldots, x_N$, for example a *bag-of-visual-words*, from which we form a *data matrix* $X = (x_1, x_2, .., x_N) \in \mathbb{R}^{d \times N}$. Each $x_i \in X$ has a known attribute representation $a_i \in \mathcal{A}$, e.g. obtained from an existing set of attribute classifiers, such as [1]. Our goal is to augment a_i with a non-semantic $b_i \in \mathcal{B}$, forming a hybrid $[a_i, b_i] \in \mathcal{AB}$, where $[\cdot]$ denotes the concatenation of vectors and $\mathcal{AB} = \mathcal{A} \times \mathcal{B}$. From the new, hybrid representation we expect better properties than from the semantic part alone with respect to a target task. For simplicity, in this work we consider only a binary representation for the semantic attribute space $\mathcal{A} = \{0, 1\}^n$, and binary or probabilistic representations for the non-semantic space $\mathcal{B} = [0, 1]^m$, and we assume that the target task is nearest-neighbor based object categorization. As it will become clear from the description, generalization of this setup are not hard to obtain, as they only require exchange of a loss function.

In learning the augmented representations, we look at two scenarios: *unsupervised* and *supervised*. The unsupervised case is applicable whenever we have training examples, regardless if we know their labels or not, whereas for the supervised case, we need to know the class labels. Again, it will become clear from the description that a *semi-supervised* case that combines properties of the unsupervised and supervised can easily be obtained by a suitable choice of loss function, but we do not explore this option in this manuscript.

2.1 Unsupervised Learning of a Feature Space Augmentation

As main idea in learning the augmenting features in an unsupervised way we use the *autoencoder* principle. In general, an autoencoder aims for finding a latent representation for a set of data that 1) is low-dimensional and therefore compact, and 2) preserves as much of the information in the original input signal as possible. This is achieved by forming a two-layered construction, in which a first layer *encodes* the input data into the latent representation, and a second layer *decodes* this representation back into the original data space. Each of the layers is parametrized, and training the autoencoder means to identify parameters for both layers such that the overall reconstruction error for a set of training examples is minimized. Intuitively, a representation that allows good reconstruction of the input sample has captured more of the contained information than one that doesn't.

In our case, we are interested not in any ad-hoc latent representation, but we want to augment the existing semantic attributes. We achieve this by making the attribute vector a_i, a fixed part of the latent representation for any x_i, and learning an encoding only for the second part, b_i. In the decoding layer, we try to reconstruct x_i from the joint $[a_i, b_i]$, which has the effect that the b_i representation only needs to encode the information that a_i lacks, see Figure 2. Consequently, we have found a simple way to factorize the information in x_i into a semantic part in $\mathcal{A}$, and an additional, potentially non-semantic, part in $\mathcal{B}$.

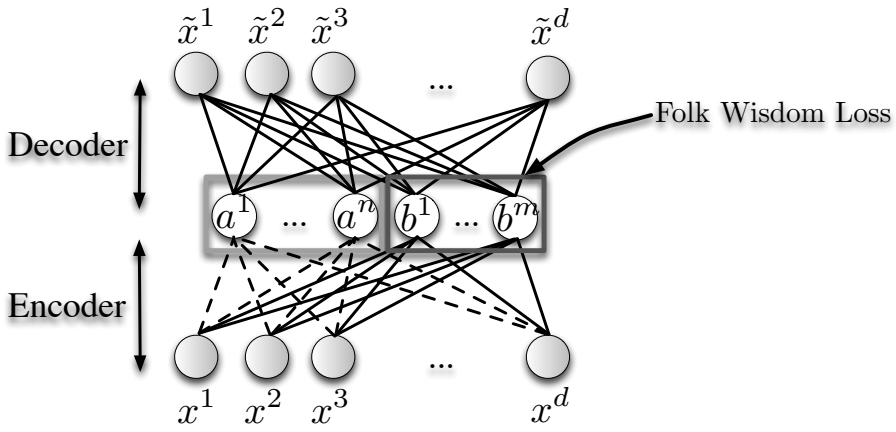

Fig. 2. Autoencoder model for learning hybrid representations: input image $x \in \mathbb{R}^d$, (encoded) hybrid representation $[a, b] \in \mathbb{R}^{n+m}$, (decoded) reconstructed image $\tilde{x} \in \mathbb{R}^d$. The reconstruction error guides the learning, and a folk wisdom principle influences good discrimination between classes in the latent attribute space.

Encoding function. As described above, the encoder function, e, maps an input $x \in \mathbb{R}^d$ to the latent space $\mathcal{AB}$. As the first, semantic, component $a \in \mathcal{A}$ is obtained from a separate method for attribute-prediction, we only parametrize the second, non-semantic component as $b = \sigma_e(W_B x)$, where $W_B \in \mathbb{R}^{m \times d}$ contains all parameters, and $\sigma_e(z) = \frac{1}{1+\exp(-z)}$ is a sigmoid non-linearity that we apply component-wise to ensure that the latent layer takes values in a range comparable to the binary-valued a. Together, we write

$$e(x) \;=\; [a, b] \;=\; [a,\, \sigma_e(W_B x)] \tag{1}$$

Decoding function. The decoder function $g : \mathcal{AB} \to \mathbb{R}^d$ aims at reconstructing the image in its original input space $\mathcal{X}$ from the latent space $\mathcal{AB}$. We assume the following linear form:

$$g([a, b]) \;=\; U[a, b] \tag{2}$$

parametrized by a matrix, $U \in \mathbb{R}^{d \times (n+m)}$, that we decompose as $U = [U_A, U_B]$ with $U_A \in \mathbb{R}^{d \times n}, U_B \in \mathbb{R}^{d \times m}$. To simplify notation, we denote the result of first encoding x then decoding it again by $\tilde{x}$. For the complete data X we can write this as

$$\tilde{X} = U_A A + U_B B \tag{3}$$

where $A \in \mathcal{A}^N$, and $B \in \mathcal{B}^N$ are the encoded representations of the data X.

Reconstruction loss. The reconstruction loss measures the loss incurred by mapping the input data to the latent space and then reconstructing the input

from the latent space. As such, it can be used to judge the quality of a choice of parameters W_B and U. We follow the usual choice for real-valued $x \in \mathbb{R}^d$ and use a squared error loss [7] that has the form

$$L_R = \sum_{i=1}^{N} \|x_i - \tilde{x}_i\|^2 \;=\; \|X - \tilde{X}\|_{Fro}^2 \tag{4}$$

where $\|\cdot\|_{Fro}$ denotes Frobenius norm of a matrix.

2.2 Supervised Learning of a Feature Space Augmentation

If we have access to ground truth annotation during the learning phase we can improve the augmented representation by adding an additional loss term that more directly reflects the object categorization task than the reconstruction loss.

Folk Wisdom Loss. This loss term is inspired by the intuitive principle "stay close to your friends and run away from your enemies". We can incorporate this loss for learning in the latent space $\mathcal{AB}$, because in a supervised setup a natural friendship (enemy) relation between samples is given by having the same (different) class labels. The folk wisdom loss [8] then directly reflects the idea that we would like to make few mistakes in nearest neighbor classification.

The idea of preserving the friendship while projecting the data to the latent space was earlier described in [6], where Weinberger and Saul showed how to learn a linear transformation over the data such that k-nearest neighbor belong to the same class while examples from different classes are separated by a large margin. In our work we rely on the *large margin nearest neighbor (LMNN)* formulation that they propose. First, for each sample we identify a set of friends and non-friends based on their class label. We use the notation $i \sim j$ to indicate that x_i and x_j are friends, and the notation $i \not\sim k$ to indicate that x_i and x_k are non-friends. The folk wisdom loss can then be formalized as:

$$L_{FW} = \sum_{i \sim j} d^2_{\mathcal{AB}}(x_i, x_j) + \sum_{\substack{i \sim j, \\ i \not\sim k}} \max\left\{0, C + d^2_{\mathcal{AB}}(x_i, x_j) - d^2_{\mathcal{AB}}(x_i, x_k))\right\} \tag{5}$$

where $d_{\mathcal{AB}}$ denotes the Euclidean distance in the $\mathcal{AB}$ space, i.e. $d^2_{\mathcal{AB}}(x_i, x_j) = \|[a_i, b_i] - [a_j, b_j]\|^2 = \|a_i - a_j\|^2 + \|b_i - b_j\|^2$. The first term in (5) penalizes large distances between objects of the same class. The second term penalizes small distances between objects of different classes, i.e. each sample is enforced to be C-units further from its non-friends than from its friends, where C is a application dependent parameter, that we set to be the median of the square distance between classes in the $\mathcal{A}$ space.

2.3 Regularization Risk Functional

To avoid overfitting, especially in the regime when little data is available, we introduce regularizers on all parameters:

$$\Omega(W_B) = \|W_B\|^2_{Fro}, \quad \Omega(U_A) = \|U_A\|^2_{Fro}, \quad \Omega(U_B) = \|U_B\|^2_{Fro} \tag{6}$$

In combination, we obtain the following regularized risk functional for learning the hybrid attribute representations

$$L(W_B, U) = L_R(W_B, U) + \eta L_{FW}(W_B) + \alpha\Omega(U_A) + \beta\Omega(U_B) + \gamma\Omega(W_B) \quad (7)$$

where we have made the dependence of the loss terms on the unknowns W_B and U explicit. The objective function expresses the properties we expect from the latent representation: 1) it should be compact (automatic, because $\mathcal{A}$ and $\mathcal{B}$ are low-dimensional), 2) it should retain as much information as possible from X (enforced by L_R), 3) it should have higher discriminative power than $\mathcal{A}$ alone (enforced by the folk wisdom loss L_{FW}), and 4) it should generalize from X to unseen data (enforced by the regularization). The trade-off variables η, α, β, and γ control the relative influence of the aspects 2)–4). Setting $\eta = 0$ we obtain a formulation that allows unsupervised feature learning, because only the folk wisdom loss requires knowledge of labels (through the definition of friends and non-friends). Even though we do not enforce property 3) in this case, we can still hope for better classification performance, because property 2) will cause additional information to be present in the hybrid representation that in the semantic one alone.

3 Optimization

Minimizing the expression (7) is a non-convex optimization problem. The reconstruction loss is non-convex with respect to the weight matrix W_B due to the nonlinear transformation in the encoder function (1). The folk wisdom loss is non-convex with respect to the weight matrix W_B when optimizing the non-friends relation part, i.e. the second term in (5). One potential approach to solve the optimization problem is to use alternating optimization with respect to one weight matrix at the time while fixing the others.

The key observation is that when the weight matrices W_B, U_B in (7) are fixed we can obtain the closed form solution for updating the matrix U_A by solving a ridge regression problem. The closed form solution to:

$$\min_{U_A \in \mathbb{R}^{d \times n}} \|U_A A + U_B B - X\|^2_{Fro} + \alpha \|U_A\|^2_{Fro} \quad (8)$$

for fixed X, and U_B, A, B is:

$$U_A = (X - U_B B)A^T(AA^T + \alpha I_n)^{-1} \quad (9)$$

where I_n is the identity matrix of size n, and αI_n reflects the regularization on the matrix U_A. Essentially U_A aims to capture the information, which was lost by decoding from the latent space $\mathcal{B}$, i.e. $X - U_B B$. By analogy, for fixed X, and U_A, A, B we obtain the closed from solution for updating the matrix U_B:

$$U_B = (X - U_A A)B^T(BB^T + \beta I_m)^{-1} \quad (10)$$

where I_m is the identity matrix of size m, and βI_m regularizes the matrix U_B.

Algorithm 1 Learning Feature Augmentation

Input Training set X with attribute representation A
Input Regularization constants α, β, γ
Input If *supervised*: training labels Y, regularization constant η
repeat
  $U_A \leftarrow$ update from closed form solution (9)
  $U_B \leftarrow$ update from closed form solution (10)
  if *supervised* **then**
    Randomly pick friend and non-friend pairs based on class label Y
    $W_B \leftarrow \text{argmin}_{W_B} \quad L_R(W_B) + \eta L_{FW}(W_B) + \gamma\Omega(W_B)$
  else
    $W_B \leftarrow \text{argmin}_{W_B} \quad L_R(W_B) + \gamma\Omega(W_B)$
  end if
until convergence, or for a maximal number of iterations
Return W_B, U_A, U_B

For W_B the non-linearity of encoding prevents a closed form expression. After updating U_A, U_B several existing optimization solvers can be used for updating the matrix W_B. In this work we use Broyden-Fletcher-Goldfarb-Shanno gradient descent method with limited-memory variation (L-BFGS). Note, we do not need to run full L-BFGS procedure at each pass to update the matrix W_B, few steps only. To speed up the training, we use a step size of 2 in our experiments. While training the autoencoder, because A is fixed whereas B is learned, we expect U_A to vary less strongly, so we can accelerate the optimization by updating the matrix U_A less frequent, e.g. at every t-th iteration. The proposed training procedure is summarized in the Algorithm 1.

4 Related Work

While *semantic attributes* [9] have received a lot of attention recently, most of the existing work studies either *zero-shot learning* with no training examples [1, 2], or the more classical case of many training examples, that allow training of discriminative probabilistic or maximum-margin classifiers [10, 11]. Our interest lies on the case inbetween, where some, but few examples per class are available. It appears wasteful to use zero-shot learning in this case, but it has also been observed previously that discriminative techniques tend to fail in this regime [12, 13], unless specific transfer learning techniques can be incorporated [14, 15].

A characteristic aspect of our work is that we want to extend the set of semantic attributes. Prior approaches aimed at preserving the property that all attributes have a semantic meaning. Therefore, they required additional human knowledge, obtained either by the analysis of textual sources [16], or by interaction with human users [17]. By adopting a hybrid model in which semantic and non-semantic attributes occur together, we get away without such an additional source of human input.

Our approach of using an autoencoder to find a useful feature representation follows the recent trend of learning feature representations in an unsupervised way [18, 5, 19]. Splitting the feature representation of the autoencoder into heterogeneous groups has been discussed in [20], [21] among others. However, in our case factorization of the autoencoder is different due to asymmetry of the semantic and non-semantic parts. The semantic part reflects the human-interpretable attributes and is fixed, whereas the non-semantic part is learned to overcome shortcomings of the semantic attributes at the expense of not being interpretable. To our knowledge, our work is the first that explores the idea of autoencoders jointly with the large margin nearest neighbor principle [6]. Other approaches to preserve class structure during feature learning exist, however. For example, [22] trains a deep network and afterwards uses Neighborhood Component Analysis (NCA) to improve the k-NN classification accuracy. NCA is also the basis of [23], which conceptually is most related to our method. It aims at learning a feature representation which is suitable for the object categorization especially with a small training set. Its focus, however, does not lie on leveraging existing attribute annotation, but to make optimal use of the available training examples by constructing many virtual training sets. We compare to their results in our experimental evaluation.

5 Experiments

We use the *Animals with Attributes (AwA)*[3] dataset introduced in [1]. The dataset consists of 30475 images. Each of the image is attached with a category label which corresponds to the animal class. There are 50 animals classes in this dataset. The dataset also contains semantic information in the form of an 85-dimensional Osherson's [24] attribute vector for each animal class. Following the studies of [1], we use 24295 images from 40 classes to learn the semantic attribute predictors. From the remaining 10 classes, we take 4680 images for training the autoencoder model, and use the rest of 1500 images, i.e. 150 from each class, to test the performance of the model. We repeat the whole procedure of training and test 5 times to get better statistics of the performance. In our experiments, we use the representation by SURF descriptors [25] provided with the dataset and referred to as original feature representation. We further normalize the features to have zero mean and unit standard deviation for each dimension.

Algorithms. We analyze two variants of our proposed method: the first variant is where the hybrid representation is learned unsupervisedly via the autoencoder architecture while minimizing *only* the reconstruction loss; and the second is where the hybrid image representation is learned with additional supervision via the folk wisdom principle. The supervision comes from friendship and non-friendship relations based on class label. In this experiment, we define friends to be samples coming from the same class and non-friends to be from different

[3] `http://attributes.kyb.tuebingen.mpg.de/`

classes. To keep the terms in balance we sample the pairs such that the cardinality of the non-friends set has the same order as the friends set. We find that 3 friends and 3 non-friends for each sample is a good balance between computational efficiency and accuracy performance. Further, we stochastically change the pairs of friends and non-friends as the optimization solver cycles through the steps.

Evaluation metric. We use k-nearest neighbor classification accuracy as the evaluation metric with $k = 3$. We compare the classification performances of our proposed unsupervised and supervised hybrid representations to baselines using original bag-of-visual-words image representation and pure semantic attribute representation of [1]. The semantic attribute representation is currently the only method that is able to predict a class label without seeing any examples of the class and thus this attribute representation shows significant advantage in the small shot setting over bag-of-visual-words representation. However the latter, in principle, can benefit from the availability of more training data points.

Model selection. For the semantic attribute baseline, we learn a predictor for all of the 85 attributes based on samples and semantic information on the set of 40 animal classes. We use an ℓ_2 -regularized logistic regression model with the 2000 dimensional bag-of-visual-words image representations. We perform a cross validation model selection for choosing the regularization trade-off variable.

We also perform a cross validation model selection approach in choosing the regularization parameters for our unsupervised learning variant, α, β and γ, and then for our supervised variant η given the trade-off parameters of the unsupervised from the previous model selection.

Results. We demonstrate the performance in a small shot setting, when we have $2, 4, 6, 8, 10, 20, 30$ number of training samples per class. These samples are the only ones used to train the autoencoder model and to assess k-nearest neighbor performance. We randomly select the required number of training samples from the available training samples per class, which in total is 4680 images. We are interested in exploring how the latent attribute space $\mathcal{AB}$ benefits when augmenting the $\mathcal{A}$ with only few dimensions, and up to the case when we double the size of the latent space representation compare to semantic attribute space $\mathcal{A}$. Guided by this interest, we augment the semantic attribute space $\mathcal{A}$ with a $m = 10, 20, 50, 85$ dimensional space $\mathcal{B}$, and we study the performance of the methods across dimensions. The results are summarized in Figure 3.

Our experiments show that categorization performance of the proposed unsupervised variant is always better or on the par with semantic attribute representation. Further, in a majority of cases we observe that our supervised variant shows an increased improvement over the unsupervised counterpart. As expected, in the small training samples regime, performance of both proposed hybrid models and semantic attribute representation are significantly better than the original representation.

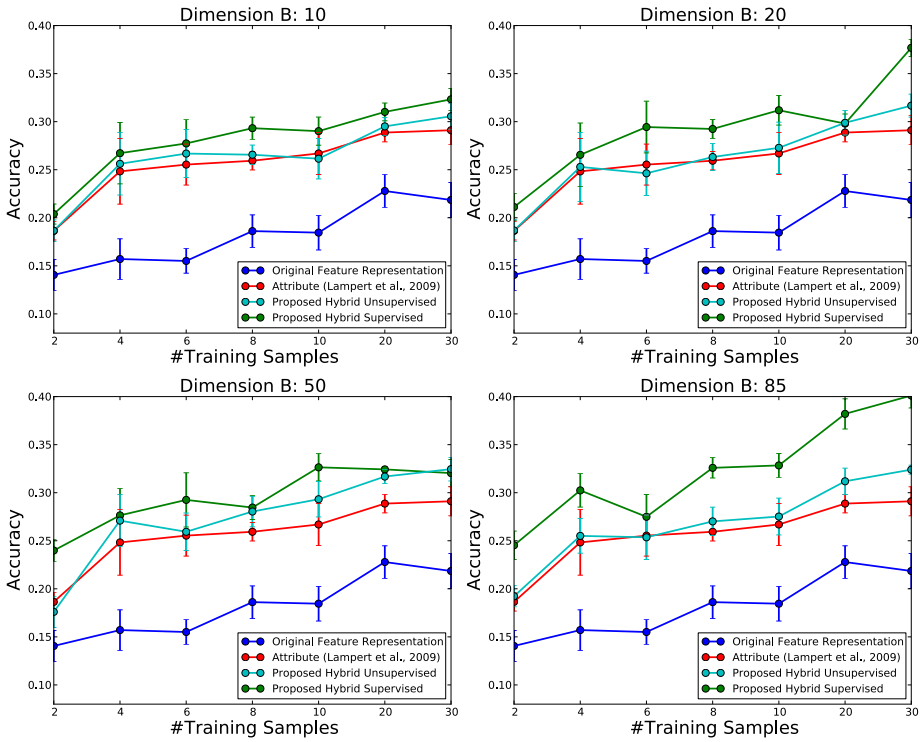

Fig. 3. Augmented attribute representations using proposed hybrid unsupervised and supervised methods. Comparison with baselines methods using original feature representation (SURF features), and predicted attribute representation [1] (mean and standard deviation over 5 runs). View of the classification performance across dimensions of the latent space $\mathcal{B}$.

Looking at Figure 3 more closely for $m = 10$ dimensional space $\mathcal{B}$, we can see that our hybrid unsupervised model shows only minute improvements over semantic attributes representation when augmenting with only few dimensions. This is expected as the effect of few additional dimensions is overwhelmed by a strong semantic prior which is by itself already discriminative enough. Whereas at higher dimensions such as $m = 50$, the unsupervised variant becomes clearly better in comparison to the semantic attribute representation alone. When we double the size of the latent space, i.e. $m = 85$, we observe saturation in the improvements at small number of training samples, due to highly redundant information in the encoded $\mathcal{B}$ space. As number of samples grow, the trend of increased recognition performance continues.

We also observe a more positive effect of incorporating the folk wisdom principle into the learning of latent attribute space when more samples become available. The proposed hybrid supervised representation integrates the knowledge about object classes by enforcing a large margin between images from different

classes. The margin concept helps to improve recognition performance even at low dimension of the $\mathcal{B}$ space. But we note that in some cases the performance of our supervised method only matches the unsupervised counterpart. Such cases can be seen in Figure 3 at dimension $m = 20$, and at dimension $m = 50$. This is caused by sensitivity of the method to the model selection on the trade-off variables between reconstruction loss and folk wisdom loss.

We also look into the case when we are given the *ground truth* semantic attributes of the input images for training the autoencoder model. One could hope that this leads to better results, as it eliminates the effect of noisy predictions at training time. On the other hand, using the ground truth attributes prevents the training stage from learning to compensate for such errors. The results of these experiments for $m = 10$ and $m = 85$ dimensional space $\mathcal{B}$ are shown on Figure 4. Note, because the ground truth attributes are defined per class, the semantic attribute representation of the image directly corresponds to its class representation, and therefore prevents a completely unsupervised setting. Moreover, the nearest neighbor performance using semantic attribute representation (red line) does not gain from more training data because all examples of one class have the same ground truth attribute representation. We observe an advantage of using the hybrid representations with and without folk wisdom loss over the baseline methods for higher dimensional $\mathcal{B}$ space, as for $m = 85$ on Figure 4. Similar to the case with predicted semantic attributes, augmenting the semantic attribute space only with few dimensions, as for $m = 10$ on Figure 4, does not give essential advantage in performance, which highlights again the discrimination power of the semantic attribute representation alone.

We also provide more extensive experimental analysis of the impact of different model components on Figure 5. As we can see in our setting, augmenting the semantic attributes with proposed hybrid unsupervised and supervised methods is clearly better than learning a representation "from scratch" (baselines with $A = 0$). We also illustrate the dominating role of the folk wisdom criterion over the reconstruction criterion in the proposed hybrid supervised model. In this case, the augmented attribute represnetations are learned using the folk wisdom criterion while eliminating the reconstruction term in (7).

Comparison to related work. Earlier work on object categorization for the Animals with Attributes dataset followed different experimental setups than the one we use, so numeric results are not directly comparable to ours. For completeness, we nevertheless give an overview here: the original work of [1] reports classification accuracies of 40.5% in a zero-shot setup with *direct attribute prediction (DAP)*, and 27.8% with *indirect attribute prediction (IAP)*. However, the work makes use of a multiple-kernel learning strategy with six different feature types, whereas we rely only on a single feature type. Note that the "Attribute" baseline we report in Figure 4 corresponds approximately the DAP model. [26] performed experiments on 1– and 10–nearest neighbor classification accuracy. For the same SURF features that we use, the authors achieve 11.7% accuracy for the ℓ_2-norm and 16.4% accuracy for the ℓ_1-norm, which is comparable to the "Original Feature Representation" we report in Figure 3 and Figure 4. [23]

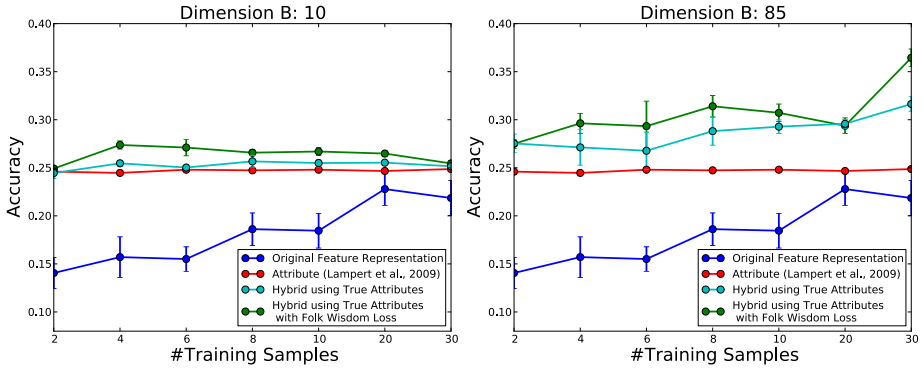

Fig. 4. Learning hybrid representations with and without folk wisdom loss using ground truth semantic attributes. Comparison with baselines methods using original feature representation (SURF features), and ground truth attribute representation [1] (mean and standard deviation over 5 runs). View across dimensions of the latent space $\mathcal{B}$.

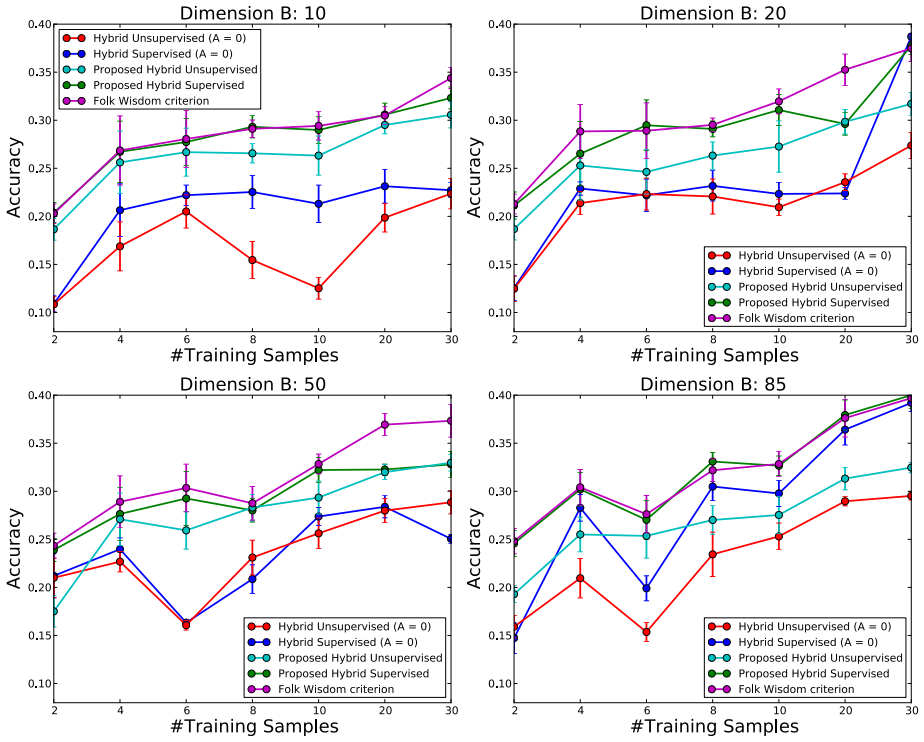

Fig. 5. In-depth analysis of the model components. Comparison of the proposed methods to augment semantic attribute representations with learning feature representations without semantic attributes ($A = 0$). Role of the Folk Wisdom criterion alone in the proposed hybrid supervised method. Mean and standard deviation over 5 runs. View across dimensions of the latent space $\mathcal{B}$.

learned feature representations in a one-shot setup. Using the same combination of 6 different feature descriptors as [1], the authors report 23.7% for linear representations, 27.2% for non-linear, and 29.0% for a combinations of non-linear with semantic attribute features.

6 Discussion and Conclusion

In this work we introduced a method to augment a semantic attribute representation by additional, non-semantic, mid-level features. The main idea is to learn only the non-semantic part of the representation by an autoencoder in combination with an (optional) maximum-margin loss term, while keeping the semantic part fixed. The effect is that the additional feature dimension overcome shortcomings of the semantic original ones, but do not copy their behavior. We interpret the result as an orthogonal decomposition of the image features into semantic, and non-semantic information.

Our experiments showed that the additional flexibility offered by the hybrid features improve the nearest neighbor classification accuracy over the purely semantic representation. In particular, they allow for a smooth transition between the zero-shot case (no training images), the unsupervised case (training images without labels) and the supervised case (training images including their labels).

A drawback of the setup we chose is that it requires regularization, and therefore the choice of regularization parameters. We used standard cross-validation for this, but if the number of training examples is small – and this is exactly the case of interest to us– this step can become unreliable. Instead, it could be promising to decide on free parameters using a Bayesian criterion that does not require splitting the available data into parts. A second task we plan to address is how to make use of the learned representation beyond classification itself. Because a significant part of the hybrid representation is semantic, we expect that techniques, e.g., for generating *image descriptions* are still applicable. In this respect is that very useful that the modular setup of our method allows replacing the folk wisdom with any other suitable loss. We plan to explore this and the questions mentioned previously in future work.

Acknowledgments. NQ is supported by a Newton International Fellowship.

References

1. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by betweenclass attribute transfer. In: CVPR. (2009) 951–958
2. Palatucci, M., Pomerleau, D., Hinton, G., Mitchell, T.: Zero-shot learning with semantic output codes. In: NIPS. (2009) 1410–1418
3. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.A.: Describing objects by their attributes. In: CVPR. (2009) 1778–1785
4. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A., Berg, T.: Baby talk: Understanding and generating simple image descriptions. In: CVPR. (2011) 1601–1608

5. Hinton, G., Salakhutdinov, R.: Reducing the dimensionality of data with neural networks. Science **313** (2006) 504 – 507
6. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. JMLR **10** (2009) 207–244
7. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. JMLR **11** (2010) 3371–3408
8. Quadrianto, N., Lampert, C.H.: Learning multi-view neighborhood preserving projections. In: ICML. (2011) 425–432
9. Ferrari, V., Zisserman, A.: Learning visual attributes. In: NIPS. (2008) 433–440
10. Wang, Y., Mori, G.: A discriminative latent model of object classes and attributes. In: ECCV. (2010) 155–168
11. Mahajan, D., Sellamanickam, S., Nair, V.: A joint learning framework for attribute models and object descriptions. In: ICCV. (2011) 1227–1234
12. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. PAMI **28** (2006) 594–611
13. Wolf, L., Hassner, T., Taigman, Y.: Descriptor based methods in the wild. In: ECCV Workshop on Faces in Real Life Images. (2008)
14. Tommasi, T., Orabona, F., Caputo, B.: Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In: CVPR. (2010) 3081–3088
15. Rohrbach, M., Stark, M., Szarvas, G., Gurevych, I., Schiele, B.: What helps where - and why? semantic relatedness for knowledge transfer. In: CVPR. (2010) 910–917
16. Berg, T., Berg, A., Shih, J.: Automatic attribute discovery and characterization from noisy web data. ECCV (2010) 663–676
17. Parikh, D., Grauman, K.: Interactively building a discriminative vocabulary of nameable attributes. In: CVPR. (2011) 1681–1688
18. Welling, M., Rosen-Zvi, M., Hinton, G.: Exponential family harmoniums with an application to information retrieval. In: NIPS. (2005)
19. Ranzato, M., Huang, F., Boureau, Y., LeCun, Y.: Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: CVPR. (2007) 1–8
20. Gregor, K., LeCun, Y.: Emergence of complex-like cells in a temporal product network with local receptive fields. CoRR **abs/1006.0448** (2010)
21. Hinton, G., Krizhevsky, A., Wang, S.: Transforming auto-encoders. In: ICANN. (2011) 44–51
22. Salakhutdinov, R., Hinton, G.: Learning a nonlinear embedding by preserving class neighbourhood structure. In: AISTATS. (2007)
23. Tang, K., Tappen, M., Sukthankar, R., Lampert, C.: Optimizing one-shot recognition with micro-set learning. In: CVPR. (2010) 3027–3034
24. Osherson, D.N., Stern, J., Wilkie, O., Stob, M., Smith, E.E.: Default probability. Cognitive Science **15** (1991) 251–269
25. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). CVIU (2008) 346–359
26. Ebert, S., Larlus, D., Schiele, B.: Extracting structures in image collections for object recognition. In: ECCV. (2010) 720–733