Visually-Driven Semantic Augmentation for Zero-Shot Learning

Abhinaba Roy[1,2]
abhinaba.roy@iit.it

Jacopo Cavazza[1]
jacopo.cavazza@iit.it

Vittorio Murino[1,3]
vittorio.murino@iit.it

[1] Pattern Analysis and Computer Vision
Istituto Italiano di Tecnologia
Genova, Italy

[2] Department of Naval, Electrical,
Electronic and Telecommunications
Engineering
University of Genova, Italy

[3] Department of Computer Science
University of Verona, Italy

Abstract

In this paper, we tackle the zero-shot learning (ZSL) classification problem and analyse one of its key ingredients, the semantic embedding. Despite their fundamental role, semantic embeddings are not learnt from the visual data to be classified, but, instead, they either come from manual annotation (attributes) or from a linguistic text corpus (distributed word embeddings, DWEs). Hence, there is no guarantee that visual and semantic information could fit well, and as to bridge this gap, we propose to augment the semantic information of attributes/DWEs with semantic representations directly extracted from visual data by means of soft labels. When combined in a novel ZSL paradigm based on latent attributes, our approach achieves favourable performances on three public benchmark datasets.

1 Introduction

Zero-Shot Learning (ZSL) refers to the problem of transferring a classification model trained on a set of *seen* classes and deploying it on a set of completely different classes - the *unseen* ones [6, 15, 33]. To do so, ZSL approaches take advantage of *semantic embeddings* which act as a sort of “bridge” between seen and unseen classes. Depending on the nature of semantic embedding, ZSL approaches can be classified in two categories, one based on attributes and the other based on distributed word embeddings.

Attribute-based methods leverage human defined attributes to describe the classes to be discriminated. Specifically, attributes are binary vectors in which each entry denotes the presence/absence of a particular feature characterizing the "object" or class. For instance, in the case of animal classification [15], a model trained on *zebras* is able to recognize *horses* since informed that both have four legs/hoofs, are mammals, have a mane and they both eat grass. Crucially, since attribute annotation is an expensive process, *distributed word embeddings* (DWEs) - such as word2vec [19] or GloVE [22] - are used as surrogates. They are learnt from a deep neural network that creates continuous embeddings from a text corpus by imposing that two words have nearby embeddings if they often occur close to each other in a sentence.

Each one of these two types of semantic embeddings has drawbacks. Finding an exhaustive list of attributes by manual annotations is usually expensive and difficult; on the other hand, DWE-based approaches are not easily interpretable. More importantly, the semantic information provided by attributes/word-embeddings is typically unable to encode semantic patterns in visual data. For instance, still in the example of the *zebra*-to-*horse* transfer above quoted, a great boost to ZSL would be provided by noting that, in addition to sharing attributes, zebras look extremely similar to horses - apart from the stripes.

Recently, a few papers [11, 12, 18, 23] have tried to incorporate visual information in the semantic embeddings by aligning the geometry of *semantic space* (made of either attributes or DWEs) onto the *visual space* produced by the visual feature representation, usually provided by fully connected layers of a convolutional neural network (CNN). These works leverage the implicit assumption that, among the visual and the semantic spaces, the former is preferable and the latter should be modified accordingly. Differently, in our work, we posit that semantic and visual spaces are equally important since providing complementary sources of information. Therefore, as opposed to modifying the semantic embedding on the basis of visual patterns, we propose to **augment** it by using visual semantic information extracted from the data itself. This *augmentation* is semantic in nature since it exploits the class similarity information obtained from a deep neural network in the form of *soft labels* [28], which are finally jointly considered with the semantic attributes/DWEs. This is performed by devising an optimization process in which the latent attributes are inferred in the resulting *visually-driven* augmented space, which globally includes the semantic embedding, the visual features and the soft labels.

Differently from the hard labels (e.g., one-hot encoding), which only describe the correct class, soft labels [28] estimate the likelihood probability distribution for an arbitrary instance to belong to every class. Therefore, if two classes are visually similar to each other, we expect this similarity to be captured by soft labels, and we claim that this fact can boost performance in ZSL methods.

In summary, we present visually-driven semantic augmentation (VdSA), a novel ZSL pipeline in which we learn a set of latent attributes to fuse semantic information captured by attributes/word-embeddings with the one conveyed by soft labels. To the best of our knowledge, our learning pipeline is the first to use visual data to augment the semantic embeddings (attributes/DWEs) usually exploited in ZSL. More specifically, the main contributions of this paper are the following:

1. We propose a visually-driven semantic augmentation (VdSA) method, a novel ZSL approach that augments the semantic information coming from attributes/DWEs with that of the visual patterns embedded in a deep network's soft labels.

2. We provide an experimental ablation study to thoroughly certify that the usage of soft labels is indeed beneficial for ZSL, no matter which semantic space is considered (either manually defined attributes, distributed word embeddings or both).

3. In a broad experimental analysis on aP&Y [6], AwA [15] and CUB-200 [29] benchmark datasets, we assess the superiority of our proposed paradigm in terms of (improvements with respect to) the state-of-the-art performance in ZSL.

The rest of the paper is organised as follows. In Section 2, we briefly review previous related works in the ZSL literature. In Section 3, we present our visual-driven attribute augmentation approach, and, in Section 4, we assess the validity of our method by reporting the results of a broad experimental testing phase. Finally, Section 5 draws conclusions and sketches the future work.

2 Related Work

After the seminal works of Lampert et al. [15] and Farhadi et al. [6], attribute based classifiers [17, 30, 31] have become popular among zero-shot learning methods. These methods learn an embedding between *seen* data and attributes so as to carry out prediction on unseen classes via regression, ranking model or neural networks [1, 12, 16, 24].

A recent class of methods ([11, 12, 18] and especially [23]) attempt to impose a congruity constraint between visual and semantic embeddings by aligning geometrical properties of the latter onto the former. Similarly, our work also makes semantic and visual embedding more congruent, but we do not change intrinsic properties of the semantic representation, but rather we enrich it by extracting visual cues directly from the data.

Our approach is closely related to the works that learn an intermediate latent (attribute) space where both visual features and attributes are projected [10, 21, 40]. With either probabilistic graphical models [40] or dictionary-based methods [10, 21], these works take advantage of such latent space to modify the semantic embedding for the sake of ZSL. In stark contrast to the above mentioned approaches, we do not attempt to modify either visual or semantic embedding since we believe that they both provide useful information. However, since relying completely only on semantic embedding is not enough to obtain a reliable classification,we instead propose to *augment* the classification capabilities of this embedding space by exploiting visual cues extracted from seen (class) data in the form of soft labels [28]. Other related works are briefly discussed in §4.3, where we compare the empirical performance of our approach with the state-of-the-art methods in the recent literature.

3 The VdSA Method

In this Section, we present the core technical contribution of our work, consisting of a novel optimization pipeline, called Visually-driven Semantic Augmentation (VdSA), to augment semantic embeddings by means of visual cues extracted from soft labels. Before digging into the details of the method (in §3.2), some technical background on ZSL is provided (in §3.1).

3.1 ZSL background

Let X denote a generic instance data to be classified (in this work, an image). In ZSL, the task is to train a model with full supervision using instances belonging to a given set of seen classes $\mathcal{Y}_{\text{seen}}$. During testing, such model is transferred on a *different* set of unseen classes $\mathcal{Y}_{\text{unseen}}$. As usually done in ZSL [33], one assumes that seen and unseen classes are disjoint sets, that is, $\mathcal{Y}_{\text{seen}} \cap \mathcal{Y}_{\text{unseen}} = \emptyset$.

For each instance X, we compute its visual signature $f_X \in \mathbb{R}^m$ in a visual embedding space whereas, for each class y, we compute a semantic representation $s_y \in \mathbb{R}^n$ in semantic embedding space. In other words, one can think of f_X as a deep feature from a CNN [14, 27], and s_y can be a distributed word embedding (such as word2vec [19]).

As a learning process, ZSL can be framed in stages. In training, we *only* use the seen classes $\mathcal{Y}_{\text{seen}}$ to learn a transformation Φ taking visual signature f_X as input and outputs a semantic representation $\Phi(f_X)$. In testing, we predict the unseen class of a testing instance $\widetilde{X}$ by selecting $\widetilde{y} \in \mathcal{Y}_{\text{unseen}}$ according to the criterion

$$\widetilde{y} = \arg\min_{y \in \mathcal{Y}_{\text{unseen}}} \|\Phi(f_{\widetilde{X}}) - s_y\|_2, \tag{1}$$

where $\|\cdot\|_2$ stands for the Euclidean norm.

3.2 Visually-driven Semantic Augmentation for ZSL

Our proposed Visually-driven Semantic Augmentation (VdSA) builds upon the previous setup of learning a transformation Φ from visual to semantic embeddings (that is, from f, computed from X, to s, computed from y).

Our approach is to augment semantic information through the representation p_X extracted from the softmax output vector of a deep neural network (trained for image classification tasks on the seen classes only), fed with X. Let us note that since we are learning a mapping $\Phi\colon f \to s$ where the semantic representation s is the output (and not the input), one cannot simply concatenate soft labels to attributes/DWEs features and learn a (supposedly) better function Φ. Therefore, the way of inserting soft labels in a ZSL pipeline is not trivial a priori.

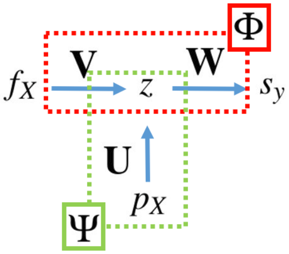

Figure 1: Overview of the proposed ZSL method. In the first stage, the visual embedding f_X is first mapped into the latent attribute space z and, afterwards into the semantic embedding s_y: this mapping is performed by Φ, which depends by parameters $\mathbf{V}$ and $\mathbf{W}$. The second stage is accomplished by the auxiliary mapping Ψ, which depends by parameters $\mathbf{U}$.

In our work, we tackle this issue and propose the following novel ZSL pipeline. The key idea consists of introducing a intermediate layer z represented by latent attributes, instead of having a direct mapping Φ from f_X to s_y [9, 10, 13, 21, 34, 40]. The rationale is that we take advantage of z in order to do a compression while fusing the visual embedding f_X and soft-labels p_X, ultimately integrating the semantic cues which are extracted by softmax operators from visual data directly.

Formally, let F_{seen} denote the visual data, i.e., $m \times d$ matrix which stacks by columns all the visual features f_{X_k} computed from training data instances X_k, $k = 1, \ldots, d$, belonging to the set of seen classes . Similarly, let S_{seen} be the $n \times d$ matrix whose k^{th} column gives the semantic representation s_{y_k} relative to the seen training class y_k to which X_k belongs to. As a baseline, we consider the following model based on latent attributes [9, 10, 13, 21, 34, 40]:

$$\min_{\mathbf{W},\mathbf{V},Z} \|S_{\text{seen}} - \mathbf{W}Z\|_F^2 + \alpha \|Z - \mathbf{V}F_{\text{seen}}\|_F^2, \tag{2}$$

where $\alpha > 0$, $\|\cdot\|_F$ stands for the Frobenius norm, Z stacks by columns all the latent attributes, and the two sets of parameters $\mathbf{W}$ and $\mathbf{V}$ represent the mapping Φ (see Figure 1)[1].

In order to make latent attributes aware of semantic information distilled from visual data by means of soft labels, we introduce the auxiliary function Ψ which enriches z using p_X. Therefore, our proposed ZSL framework rewrites as follows

$$\boxed{\min_{\mathbf{W},\mathbf{V},\mathbf{U},Z} \|S_{\text{seen}} - \mathbf{W}Z\|_F^2 + \alpha \|Z - \mathbf{V}F_{\text{seen}}\|_F^2 + \beta \|Z - \mathbf{U}P_{\text{seen}}\|_F^2} \tag{3}$$

[1] Although in principle the map Φ can be arbitrary, here we consider it to be a composition of linear functions as commonly done in several mainstream ZSL approaches such as [1, 7, 10, 13, 25].

Dataset	No. of instances	No. of attributes	No. of seen/unseen classes
aP&Y [6]	15339	64	20/12
AwA [15]	30475	85	40/10
CUB-200 [29]	11788	312	150/50

Table 1: Description of the ZSL benchmark datasets used in experiments

where $\alpha, \beta > 0$ and P_{seen} stacks all soft labels by columns. In (3), the usage of the auxiliary mapping Ψ helps Φ in optimizing z as to 1) map visual into semantic embeddings, and 2) take advantage of the auxiliary semantic information extracted by means of soft labels. This constitutes a novel approach, never investigated in previous ZSL methods [6, 10, 15, 17, 21, 30, 31, 40], in which semantic information driven by visual data is actually disregarded.

Optimization.[2] The objective function (3) is not jointly convex with respect to all the variables $\mathbf{W}, \mathbf{V}, \mathbf{U}, Z$. But it becomes convex as long as one optimizes over one variable while fixing the others. In fact, if one uses alternating optimization to solve (3), when either optimizing over $\mathbf{W}$ (resp. $\mathbf{V}$ or $\mathbf{U}$), only the first (resp. second or third) term in (3) is considered. More importantly, solving for $\mathbf{W}, \mathbf{V}$ and $\mathbf{U}$ separately is a least square fitting for which a closed-form solution exists (due to the usage of the Frobenius norm, one can use normal equations [2]).

Similarly, the optimization for Z can be done in closed-form by the following change of variables: Z can be found by minimizing the objective $\|A - BZ\|_F^2$ (while freezing $\mathbf{W}, \mathbf{V}$ and $\mathbf{U}$) where the matrices A and B are given as the 3×1 block-column matrices composed of $S_{\text{seen}}, \mathbf{V}F_{\text{seen}}, \mathbf{U}P_{\text{seen}}$ and $\mathbf{W}, I, I$, respectively (I denotes the identity matrix of suitable size). We impose a normalization on our trainable parameters; each column in $\mathbf{W}, \mathbf{V}, \mathbf{U}$ or Z has $\|\cdot\|_2$-norm upper bounded by 1. Even with such constraint, we still achieve closed-form solution in our alternated optimization thanks to Lagrangian multipliers [26]. In the inference stage, predictions are done using eq. (1).

4 Experimental Results

In this Section, we empirically demonstrate the effectiveness of the VdSA approach. More precisely, after providing technical details for reproducibility (in §4.1), we provide an ablation study to assess the effect of soft labels (in §4.2) and, finally, we report a comparison with the state-of-the-art methods on three benchmark datasets (in §4.3).

4.1 Details for reproducing the experiments

We validate our proposed semantic augmentation by performing experiments on three ZSL benchmark datasets for the task of object recognition and image classification: aPascal & aYahoo (aP&Y) [6], Animals with Attributes (AwA) [15], Caltech-UCSD Birds-200-2011 (CUB-200) [29]. Details of these datasets are shown in Table 1 in terms of number of samples/attributes. To partition seen and unseen classes, we adopt the splitting criteria commonly used in the literature [33].

Visual Embedding. As done in [11, 23, 38], we encode each image with the 4096-dimensional `fc7` feature vector extracted from a VGG-19 model [27] pre-trained on ImageNet [5].

[2] *Code available at https://github.com/elchico1990/Visually-Driven-Semantic-Augmentation-for-Zero-Shot-Learning.git*

Binary attributes annotated by humans			
Dataset	baseline (2) + **A**	VdSA + **A** + **H**	VdSA + **A** + **S**
aP&Y	50.6	51.1	**51.7**
AwA	78.2	**81.0**	78.4
CUB-200	55.0	55.1	**56.7**

Continuous distributed word embeddings (DWEs) learnt from a text corpus			
Dataset	baseline (2) + **W**	VdSA + **W** + **H**	VdSA + **W** + **S**
aP&Y	41.7	42.4	**43.2**
AwA	51.6	54.1	**56.7**
CUB-200	29.8	33.9	**34.8**

Combination of binary attributes and continuous DWEs			
Dataset	baseline (2) + **A** + **W**	VdSA + **A** + **W** + **H**	VdSA + **A** + **W** + **S**
aP&Y	49.1	49.7	**53.6**
AwA	76.1	79.8	**80.6**
CUB-200	54.6	56.7	**59.7**

Table 2: Results of our ablation study. We present the multi-class classification accuracies (in percentage %) for the unseen classes used in testing - best result for each row in boldface. We compare a baseline latent attribute model (2) with our proposed model (3) for visually-driven semantic augmentation. In both cases, we evaluate with different semantic embeddings: either binary manually annotate attributes (**A**) or distributed word embeddings (**W**). Also, we compare augmentation by exploiting both hard labels (**H**) - given as ground truth - and soft-labels (**S**) - estimated from the softmax operator of VGG-19 [27].

Semantic Embedding. We consider binary attributes - manually annotated - provided with each dataset. In addition, we also use continuous distributed word embeddings (DWEs). To this end, we use word2vec [19], exploiting a pre-trained model[3] to cast each class name into a 300-dimensional vectorial representation.

Soft labels. We generate soft labels by extracting softmax outputs generated after fine-tuning AlexNet [14]. To do so, we run ADAM optimizer for 5000 iterations with a fixed learning rate of 0.001 and dropout regularization in the AlexNet fully connected layers (with a dropout rate of 0.5). In each dataset, we setup soft labels for both seen and unseen classes, but, in order to follow a fair ZSL protocol, we supervise back-propagation on the soft labels *only* for seen classes with *only* seen class data. Therefore the entries of soft labels which correspond to unseen classes are not directly optimized with supervision but, rather, we expect that the network itself will populate them implicitly. In this way, the network will mine the similarities among different classes by itself, ultimately facilitating the transfer of knowledge (more details in §4.2).

Latent attributes. For the alternate optimization of our objective function (3), we used a uniform random initialization (in the range $[-1/2, 1/2]$) for all parameters. We did not notice any remarkable difference in the results of the optimization depending on the order with which variables are optimized - and therefore we optimized in the order **W**, **V**, **U**, Z. For the latent attributes, we fixed their number to be 300 for AwA and 500 for CUB-200 and aP&Y respectively. The values of α and β as well as the number of latent attributes are determined after five fold cross validation, using seen classes only.

[3] https://github.com/chrisjmccormick/word2vec_matlab

4.2 Ablation Study

In this Section, we present an ablation study to evaluate the effect of our visually driven semantic augmentation. To do so, we compare our latent attribute augmentation (3) with the baseline (2), without augmentation [9, 34]. We consider different semantic embeddings: binary attributes - **A**, word2vec distributed word embeddings - **W**, as well as a concatenation of the two (**A** + **W**). In addition, we also compare our proposed visual augmentation based on soft labels with another approach which, instead of the prediction given by softmax operators, directly takes into account ground truth labels in the form of one-hot encodings.
The results of our ablation study are reported in Table 2.

Discussion. We register a common trend in all three datasets when using either **A**, **W** or **A** + **W**. That is, the ZSL testing accuracies always increase when switching from the baseline (2) - second column - to our proposed visually driven semantic augmentation (3) using soft labels **S** - fourth column. For instance, +4% on CUB-200 when using **W** and +5% in the **A**+**W** case. This clearly states that our proposed augmentation 1) extracts semantic patterns from visual data and 2) combines it with the semantic information of attributes/DWEs.

Hard vs. soft labels. In principle, one can expect that hard labels **H** are better than soft ones **S** since those one-hot vectors are given as ground truth annotations (for the seen classes). On the contrary, soft labels **S** are just predictions and therefore they can be wrong. On the contrary, when switching from hard to soft labels - Table 2, third and fourth columns respectively, accuracies grow systematically. This can be explained by the fact that soft labels are more informative with respect to hard ones since they convey the confidence with which a given instance is estimated to belong to each class [28]. In our work, we build upon this idea to show that such concept can be favourably embodied in ZSL: soft labels, even if trained on the seen classes only, implicitly learns similarity patterns between seen and unseen classes and ultimately boost the transfer in between.

Combining various semantic information. From the results in Table 2, one can see how combining attributes and DWEs in (2) is not straightforward. This is clear from the fact that DWEs (such as word2vec) are surrogates for the attributes used to conveniently circumvent manual annotation. However, in terms of performance, **A** is arguably better than **W** and our experimental findings confirm that. Moreover, a concatenation of the two does not enrich the semantic information exploited by the ZSL model to transfer from seen to unseen classes. On the contrary, the performance systematically deteriorates: when switching from **A** to **A**+**W**, the baseline (2) drops from 50.6% to 49.1% on aP&Y and (3) drops from 81.0% to 79.8% on AwA.
Remarkably, our proposed semantic augmentation based on soft labels shows a completely different behaviour: when concatenating **A** to **W**, we sharply improve with respect to using **A** only. Precisely, in the 4^{th} column of Table 2, our method achieves an improvement on +1.9% on aP&Y, +2.2% on AwA and +3% on CUB-200.

As the consequence of the solid potential showed by our method in this analysis, in the next Section, we will take advantage of it to challenge the state-of-the-art in ZSL.

4.3 Comparison with the state-of-the-art in ZSL

To evaluate our proposed visually-driven semantic augmentation against the state of the art while making comparisons fair, we split methods on the basis of the adopted semantic embedding, either 1) attributes **A**, 2) DWE (word2vec) **W**, or 3) a combination of the two **A**+**W**.

1) Among the methods which use annotated attributes, we compare with the probabilistic graphical model of Direct Attribute Prediction (DAP) [15]. We also consider DeVise [7], SJE

Method	Semantic Embedding	Datasets		
		AwA	CUB-200	aP&Y
DAP [16]	**A**	57.2	-	38.16
DeVise [7]	**A**	56.7	33.5	-
SJE [1]	**A**	66.7	50.1	-
Kodirov et al. [12]	**A**	73.2	39.5	-
ESZSL [25]	**A**	75.3	47.2	24.2
SSE [38]	**A**	76.3	30.4	46.2
MTL [35]	**A**	63.7	32.3	-
SynC [4]	**A**	72.9	54.7	-
Bucher et al. [3]	**A**	77.3	43.3	**53.2**
JLSE [40]	**A**	80.4	42.1	50.4
LAD [10]	**A**	81.0	55.1	51.1
JSLA [21]	**A**	**82.8**	49.8	-
Visually-driven Semantic Augmentation, VdSA (ours)	**A**	*78.4*	***56.7***	*51.7*
DeVise [7]	**W**	50.4	-	-
MTL [35]	**W**	55.3	-	-
ConSE [20]	**W**	46.8	23.1	21.8
SynC [4]	**W**	56.7	21.5	28.5
LatEm [32]	**W**	50.8	16.5	19.8
VAWE [23]	**W**	**61.2**	27.4	35.2
Visually-driven Semantic Augmentation, VdSA (ours)	**W**	*56.7*	***34.8***	***43.2***
Fu et al. [8]	**A + W**	66.0	-	-
SJE [1]	**A + W**	73.9	51.7	-
Kodirov et al. [12]	**A + W**	75.6	40.6	-
Visually-driven Semantic Augmentation, VdSA (ours)	**A + W**	***80.6***	***59.7***	***53.6***

Table 3: Benchmarking our proposed Visually-driven Semantic Augmentation (***VdSA***) with the state-of-the-art in ZSL. We report classification performance (measured in percentage %) obtained on the unseen classes, following the usual training/testing splits (see Table 1). The performance of our method is in italic, for each of the semantic embeddings (attributes **A**, word2vec **W** and the concatenation **A+W**) we highlight in bold the best performance. Globally, the highest classification scores in the Table are underlined.

[1] and Embarrassing Simple ZSL (ESZSL) [25] that use bi-linear compatibility functions to map visual into semantic information. We include the Similarity Semantic Embedding (SSE) [38], and the dictionary learning-based method of Kodirov et al. [12], which both apply unsupervised domain adaptation methods to ZSL. We also consider the neural network based approach of [35] based on multi-task learning (MTL) and the metric-learning paradigm by Bucher et al. [3]. We also compare against the synthetic classifier (SynC) [4] which expresses images and semantic class embeddings as a mixture of seen class proportions. Finally, we evaluate our proposed augmentation scheme against the methods JLSE [40], LAD [10], JSLA [21] which are all based on latent attributes.

2) In the case of DWEs (here word2vec [19]), we additionally consider the convex combination of nearest neighbours predictors ConSE [20], the latent embedding framework LatEm which takes advantage of a structured-output support vector machine [32], and the VAWE approach[4] [23] that re-aligns the topology of the semantic embeddings by using the one of

[4] Since VAWE is not technically a full method but just a pruning techniques for DWEs, we reported the combination of WAVE with ConSE as published in Qiao et al. [23].

the visual embedding.

3) Finally, we also consider the geometrical method proposed by Fu et al. [8], which deploys ZSL on a manifold with geodesic distances.

Let us clarify that, we do not compare against transductive approaches [36, 39] since those methods use unseen classes for training, while we do not. Moreover, we do not compare with methods such Zhang et al. [37] or Kodirov et al. [13] since they take advantage of end-to-end learning to explicitly train a deep feature representation for ZSL, whereas, we use pre-computed visual features to codify the input images[5].

We report the quantitative results of our comparison with state-of-the-art methods in Table 3 and we discuss the results in the rest of this Section.

Discussion. When using **A**, we outperform ESZSL on AwA, CUB-200 and aP&Y by +3%,+7% and +27%, respectively, and SJE in CUB-200 by +6%. With respect to DAP, we achieve a 21% improvement in AwA, and 13% improvement in aP&Y. We improve over SSE and SynC by +2% in AwA and CUB-200 and +5% in aP&Y. We outperform MTL by more than 15% in AwA and by more than 26% in CUB-200.

In general, on aP&Y and CUB-200, our method is superior to all reported methods. The only exception is Bucher et al. [3]: on aP&Y, in fact metric learning [3] seems better than attribute augmentation - but, on either AwA and CUB-200, our approach is still superior. Finally, apart from AwA, our method is able to systematically improve other latent-attribute-based methods: we improve LAD by +1.6% on CUB-200 and by +0.6% on aP&Y, we outperform JSLA by +6.9% on CUB-200 and we are +1.3% and +14.6% better than JLSE on aP&Y and CUB-200, respectively. This empirically proves the claim that augmenting is preferable to modify semantic representations.

When comparing with methods that exploit **W**, our augmentation scheme scores again a solid performance: it is on par with respect to SynC on AwA (56.7%) and it outperforms DeVise, MTL, ConSE and LatEm by large margin. For instance, +18.3% on CUB-200 with respect to LatEm and +21.4% on aP&Y with respect to ConSE.

VAWE is worth a dedicated discussion. Despite in AwA our method is gapped by −4.5%, in either CUB-200 or aP&Y, our approach improves over the performance of VAWE by +7.4% and +8.0%, respectively. Overall, this is an empirical evidence that (in most cases) visually-driven augmentation is preferable to visually-driven re-alignment.

Moreover, when combining **A + W**, our method improves Fu et al. [8], SJE and Kodirov et al. [12] by +5.0% on AwA and by +8.0% on CUB-200 - in either Table 2 or 3, this is the most favorable setup for our augmentation.

In general, our method performs extremely well on CUB-200 which is a fine-grained dataset of birds (as opposed to AwA and aP&Y which tackle regular object recognition). Thus, our experimental findings seem to suggest that visual semantics augmentation is particularly effective in the case of fine-grained ZSL.

5 Conclusions & future work

In this paper, we propose a novel optimization approach for ZSL which is based on augmenting the semantic information conveyed by attributes/ distributed word embeddings with visual patterns which are distilled from data by means of soft labels. In a broad experimental validation, we thoroughly analyse the benefits of casting such augmentation in the form of a learning pipeline where latent attributes do bottleneck compression to fuse multiple sources

[5] In spite of that, it's worth saying that, when using **A + W**, our visual augmentation is still able to improve [37] by +1.4% on CUB-200.

of semantic information. As the results certify, our pipeline score a solid performance with respect to the state-of-the-art performance on public benchmark datasets.

As future works, we will try to explore other complementary source of semantic information to boost ZSL and we also consider different applicative benchmarks (such as zero-shot action recognition).

References

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 2927–2936. IEEE, 2015.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *European Conference on Computer Vision*, pages 730–746. Springer, 2016.

[4] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5327–5336, 2016.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

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

[7] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[8] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot object recognition by semantic manifold distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2635–2644, 2015.

[9] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. Video2vec embeddings recognize events when examples are scarce. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10):2089–2103, 2017.

[10] Huajie Jiang, Ruiping Wang, Shiguang Shan, Yi Yang, and Xilin Chen. Learning discriminative latent attributes for zero-shot classification. In *ICCV*, 2017.

[11] Huajie Jiang, Ruiping Wang, Shiguang Shan, Yi Yang, and Xilin Chen. Learning discriminative latent attributes for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4223–4232, 2017.

[12] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, 2015.

[13] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, 2017.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*. 2012.

[15] CH. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

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

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

[18] Thomas Mensink, Efstratios Gavves, and Cees G. M. Snoek. COSTA: co-occurrence statistics for zero-shot classification. In *CVPR*, 2014.

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

[20] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[21] Peixi Peng, Yonghong Tian, Tao Xiang, Yaowei Wang, and Tiejun Huang. Joint learning of semantic and latent attributes. In *European Conference on Computer Vision*, pages 336–353. Springer, 2016.

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

[23] Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. Visually aligned word embeddings for improving zero-shot learning. *arXiv preprint arXiv:1707.05427*, 2017.

[24] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58, 2016.

[25] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015.

[26] W. Rudin. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1976.

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[28] Christian Thiel, Stefan Scherer, and Friedhelm Schwenker. Fuzzy-input fuzzy-output one-against-all support vector machines. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 156–165. Springer, 2007.

[29] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[30] Xiaoyang Wang and Qiang Ji. A unified probabilistic approach modeling relationships between attributes and objects. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2120–2127. IEEE, 2013.

[31] Yang Wang and Greg Mori. A discriminative latent model of object classes and attributes. In *European Conference on Computer Vision*, pages 155–168. Springer, 2010.

[32] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 69–77, 2016.

[33] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning - the good, the bad and the ugly. *CVPR*, 2017.

[34] Xun Xu, Timothy M Hospedales, and Shaogang Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *European Conference on Computer Vision*, pages 343–359. Springer, 2016.

[35] Yongxin Yang and Timothy M Hospedales. A unified perspective on multi-domain and multi-task learning. 2015.

[36] Meng Ye and Yuhong Guo. Zero-shot classification with discriminative semantic representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7140–7148, 2017.

[37] Li Zhang, Tao Xiang, Shaogang Gong, et al. Learning a deep embedding model for zero-shot learning. 2017.

[38] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4166–4174, 2015.

[39] Ziming Zhang and Venkatesh Saligrama. Zero-shot recognition via structured prediction. In *European conference on computer vision*, pages 533–548. Springer, 2016.

[40] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6034–6042, 2016.