Hard Negative Mining for Metric Learning Based Zero-Shot Classification

Maxime Bucher[1,2], Stéphane Herbin[1], Frédéric Jurie[2]

[1]ONERA - The French Aerospace Lab, Palaiseau, France
[2]Normandie Univ, UNICAEN, ENSICAEN, CNRS, Caen, France

Abstract. Zero-Shot learning has been shown to be an efficient strategy for domain adaptation. In this context, this paper builds on the recent work of Bucher *et al.* [1], which proposed an approach to solve Zero-Shot classification problems (ZSC) by introducing a novel metric learning based objective function. This objective function allows to learn an optimal embedding of the attributes jointly with a measure of similarity between images and attributes. This paper extends their approach by proposing several schemes to control the generation of the negative pairs, resulting in a significant improvement of the performance and giving above state-of-the-art results on three challenging ZSC datasets.

Keywords: domain adaptation, zero-shot learning, hard negative mining, bootstrapping

1 Introduction

Among the different image interpretation methods exploiting some kind of knowledge transfer in their design, Zero Shot Classification (ZSC) can be considered as a domain adaptation problem where the new target domain is defined using an intermediate level of representation made of human understandable *semantic attributes*. The source domain is defined by an annotated image database expected to capture the relation between data and attribute based representation of classes.

Most of the recent approaches addressing ZSC [2–8] rely on the computation of a similarity function in the semantic space. They learn the semantic embedding, either from data or from class description, and compare the embedded data using standard distance.

Recently, [1] proposed to add a metric learning (ML) step to adapt empirically the similarity distance in the embedding space, leading to a multi-objective criterion optimizing both the metric and the embedding. The metric is learned using an empirical optimized criterion on random but equally sampled pairs of similar (positive) and dissimilar (negative) data.

In this paper, following observations from the active learning community (see *e.g.*, [9]), we show that a careful choice of the negative pairs combined with the multi-objective criterion proposed in [1] leads to above state of the art results.

2 Improved ZSL by efficient hard negative mining

In a recent work, Bucher *et al.* [1] introduced a metric learning step in their zero-shot classification pipeline. Their model is trained from pairs of data, where positive (resp. negative) pairs are obtained by taking the training images associated with their own provided attribute vector (resp. by randomly assigning attribute vector of another image) and are assigned to the class label '1' (resp. '-1'). The set of positive pairs are denoted $\mathbf{D}_+$ in the following, and $\mathbf{D}_-$ for the negatives. This paper investigates improved ways to select negative pairs.

2.1 Bucher *et al.* metric learning framework for zero-shot classification

This section summarizes Bucher *et al.* [1] paper, to which the reader should refer for more details. Zero-shot classification problem is cast into an optimal framework of the form:

$$\mathbf{Y}^* = \underset{\mathbf{Y}\in\mathcal{Y}}{\arg\min}\, S(\mathbf{X}, \mathbf{Y}),$$

where $\mathbf{X}$ is an image, $\mathbf{Y}$ a vector of attributes and $\mathbf{S}$ a parametric similarity measure. A metric matrix denoted as $\mathbf{W}_A$ transforms the attribute embedding space into a space where the Euclidean distance can be used. The similarity between images and attributes is computed as:

$$S(\mathbf{X}, \mathbf{Y}) = \left\|(\hat{\mathbf{A}}_X(\mathbf{X}) - \mathbf{Y})^T \mathbf{W}_A\right\|_2 \tag{1}$$

where $\hat{\mathbf{A}}_X$ embeds the $\mathbf{X}$ modality into the space of $\mathbf{Y}$ using a linear transformation combined with a ReLU type transfer function:

$$\hat{\mathbf{A}}_X(\mathbf{X}) = \max(0, \mathbf{X}^T \mathbf{W}_X + \mathbf{b}_X). \tag{2}$$

The role of learning is to estimate jointly the two matrices $\mathbf{W}_X$ and $\mathbf{W}_A$. The empirical learning criterion used is the sum of 3 terms:

(i) a term for the metric $\mathbf{W}_A$:

$$l_H(\mathbf{X}_i, \mathbf{Y}_i, Z_i, \tau) = \max\left(0, 1 - Z_i(\tau - S(\mathbf{X}_i, \mathbf{Y}_i)^2)\right). \tag{3}$$

where Z_i states that the two modalities are consistent ($Z_i = 1$) or not ($Z_i = -1$). τ is the threshold separating similar and dissimilar examples.

(ii) a quadratic loss for the linear attribute prediction $\mathbf{W}_X$ (only applied to positive pairs):

$$l_A(\mathbf{X}_i, \mathbf{Y}_i, Z_i) = \max(0, Z_i).\left\|\mathbf{Y}_i - \hat{\mathbf{A}}_X(\mathbf{X}_i)\right\|_2^2. \tag{4}$$

(iii) a quadratic penalization to prevent overfitting:

$$R(\mathbf{W}_A, \mathbf{W}_X, \mathbf{b}_X) = \|\mathbf{W}_X\|_F^2 + \|\mathbf{b}_X\|_2^2 + \|\mathbf{W}_A\|_F^2 \tag{5}$$

The overall objective function can then now be written as the sum of the previously defined terms:

$$\mathcal{L}(\mathbf{W}_A, \mathbf{W}_X, \mathbf{b}_X, \tau) = \sum_i l_H(\mathbf{X_i}, \mathbf{Y_i}, Z_i, \tau) + \lambda \sum_i l_A(\mathbf{X_i}, \mathbf{Y_i}, Z_i) \\ +\mu R(\mathbf{W}_A, \mathbf{W}_X, \mathbf{b}_X) \tag{6}$$

In this context, ZSC is achieved by finding the most consistent attribute description from a set of exclusive attribute class descriptors $\{\mathbf{Y}_k^*\}_{k=1}^C$ given the image where k, is the index of a class:

$$k^* = \operatorname*{arg\,min}_{k \in \{1 \ldots C\}} S(\mathbf{X}, \mathbf{Y}_k^*) \tag{7}$$

2.2 Hard negative mining

In a metric learning problem, while the set of positive pairs $\mathbf{D}_+$ is fixed and given by the training set with one pair per positive image, the set of negative pairs $\mathbf{D}_-$ can be chosen more freely; indeed, as there are many more ways of being different than being equal, the number of negative and positive pairs may not be identical. Moreover, we will see that increasing the size of $\mathbf{D}_-$ compared to that of $\mathbf{D}_+$ by some factor n leads to better overall results.

In the following, we explore three different strategies to sample the distribution of negative pairs using several learning epochs: we first present a variant of the method of [1] and then describe two iterative greedy schemes.

Random In [1], negative pairs are obtained by associating a training image with an attribute vector chosen randomly among those of other seen classes, with one negative pair for each positive one. As a variant, we propose to generate randomly n negative pairs (instead of one) for each positive pair, chosen as in [1], *i.e.*, by randomly sampling the set of attribute vectors from the other classes. We include in the objective function a penalization to compensate for the unbalance between positive and negative pairs (see section 2.3).

Uncertainty This strategy is inspired by hard mining for object detection [9–12] and consists in selecting the most informative negative pairs and iteratively updating the scoring function given by Eq. (1). We denote by $S_t(\mathbf{X}, \mathbf{Y})$ this score at time t. During training, each time step t corresponds to a learning epoch. At the first epoch, $S_1(\mathbf{X}, \mathbf{Y})$ is learned using the random negative pairs of [1]. At each time t each pair of training image $\mathbf{X}_i$ and candidate annotation $\mathbf{Y}$ coming from different (but seen) classes is ranked according to the uncertainty score:

$$u_t(\mathbf{Y}|\mathbf{X}_i) = \exp(-(S_t(\mathbf{X}_i, \mathbf{Y}) - S_t(\mathbf{X}_i, \mathbf{Y}^*))) \tag{8}$$

where $\mathbf{Y}^*$ is the true vector of attributes of $\mathbf{X}_i$. The vector of attributes which are most similar to the actual one while coming from different classes are the most relevant for improving the model. We define a probability of generating the pair based on this similarity score and sample this distribution.

Uncertainty/Correlation We propose to improve the previous approach by taking into account the intra-class correlation. The underlying principle governing the selection is that the most correlated vectors of attribute, in a given class, are the most useful ones to consider. The correlation can be measured by:

$$q(\mathbf{Y}) = \exp\left(\frac{-1}{|\mathcal{Y}_k|} \sum_{\mathbf{Y}' \in \mathcal{Y}_k} \|\mathbf{Y} - \mathbf{Y}'\|_2\right) \tag{9}$$

where k is the true class index of $\mathbf{Y}$ and $\mathcal{Y}_k$ is the set of attribute vector representations.

A trade-off between uncertainty and correlation is obtained globally by using the following scoring function:

$$p_t(\mathbf{Y}|\mathbf{X}_i) = u_t(\mathbf{Y}|\mathbf{X}_i) * q(\mathbf{Y}) \tag{10}$$

where each image attribute vector $\mathbf{Y}$ at epoch t has a score of p_t to be associated with $\mathbf{X}_i$. The current set of negative pairs $\mathbf{D}_-$ at epoch t is obtained by iteratively increasing the set with new data sampled according to Eq. (10).

2.3 Adaptation of the objective function

The original learning criterion (6), such as defined in [1], assumes that negative and positive pairs are evenly distributed. This is not the case in the proposed approach: the criterion must be adapted to compensate for the imbalance between positive and negative pairs, by weighting the positive and negative pairs according to their frequencies:

$$\begin{aligned}\mathcal{L}(\mathbf{W}_A, \mathbf{W}_X, \mathbf{b}_X, \tau) = \frac{1}{|\mathbf{D}_+|}\left(\sum_{i \in \mathbf{D}_+} l_H(\mathbf{X_i}, \mathbf{Y_i}, Z_i, \tau) + \lambda l_A(\mathbf{X_i}, \mathbf{Y_i}, Z_i)\right)\\ + \frac{1}{|\mathbf{D}_-|}\left(\sum_{j \in \mathbf{D}_-} l_H(\mathbf{X_j}, \mathbf{Y_j}, Z_j, \tau)\right) + \mu R(\mathbf{W}_A, \mathbf{W}_X, \mathbf{b}_X)\end{aligned} \tag{11}$$

This criterion is updated at each new epoch when learning the model.

3 Experiments

Datasets In this section we evaluate the proposed hard mining strategy on different challenging zero-shot learning tasks, by doing experiments on the 4 following public datasets: aPascal&aYahoo (aP&Y) [13], Animals with Attributes (AwA) [14], CUB-200-2011 (CUB) [15] and SUN attribute (SUN) [16] datasets. They have been designed to evaluate ZSC methods and contain a large number of categories (indoor and outdoor scenes, objects, person, animals, *etc.*) described using various semantic attributes (shape, material, color, part name *etc.*). To make comparisons with previous works possible, we used the same training/testing splits as [13] (aP&Y), [14] (AwA), [3] CUB and [17] (SUN).

Table 1: Zero-shot classification accuracy (mean ± std) on 5 runs. We report results with VGG-verydeep-19 [18] features. unc./cor. = *Uncertainty/Correlation* method. The unc./cor. method can't be apply to the AwA dataset since all images of the same class have the same attributes, contrarily to the aP&Y, CUB and SUN datasets.

Feat.	Method	aP&Y	AwA	CUB	SUN
VGG-VeryDeep [18]	Lampert *et al.* [2]	38.16	57.23	-	72.00
	Romera-Paredes *et al.* [4]	24.22±2.89	75.32±2.28	-	82.10±0.32
	Zhang *et al.* [5]	46.23±0.53	76.33±0.83	30.41±0.20	82.50±1.32
	Zhang *et al.* [6]	50.35±2.97	80.46±0.53	42.11±0.55	83.83±0.29
	Wang *et al.* [7]	-	78.3	**48.6±0.8**	-
	Bucher *et al.* [1]	53.15±0.88	77.32±1.03	43.29±0.38	84.41±0.71
	Ours 'random'	54.41±1.47	83.48±0.99	43.79±0.68	85.98±1.14
	Ours 'uncertainty'	56.01±0.58	**86.55±1.07**	45.41±0.10	**86.21±0.88**
	Ours 'unc./cor.'	**56.77±0.75**	-	45.87±0.34	86.10±1.09

Image features For each dataset, we used the VGG-VeryDeep-19 [18] CNN models, pre-trained on imageNet (without fine tuning) and extract the fully connected layer (*e.g.*, FC7 4096-d) for representing the images.

Hyper-parameters To estimate the three hyper-parameters (λ, the dimensionality of the metric space (m) and μ) we apply a grid search validation procedure by randomly keeping 20% of the training classes. $\mathbf{W}_A$ and $\mathbf{W}_X$ are randomly initialized with normal distribution and optimized with stochastic gradient descent.

3.1 Zero-shot classification

The experiments follow the standard ZSC protocol: during training, a set of images from known classes is available for learning the model parameters. At test time, images from unseen classes have to be assigned to one of the possible classes. Classes are described by a vector of attributes. Performance is measured by mean accuracy and std over the classes.

Tables 1 and 2 show the performance given by our hard-mining approach, which outperforms previous methods on 3 of the 4 datasets by more than 3% on average (+9% on AwA). The smart selection of negative pairs plays a role on the decision boundaries especially where classes have close attribute descriptions. We did not compare our results with [3] or [8] as they use different image features. The 2 alternatives explored in the paper (*Uncertainty* vs *Uncertainty/Correlation*) give similar performance, but, as shown in the next section *Uncertainty/Correlation* is faster.

Table 2: Zero-shot classification accuracy (mean ± std) on aP&Y dataset as a function of the ratio of positive/negative pairs.

Method / #neg. pair	1	10	50	100
Random	53.15±0.88	53.98±0.79	54.41±1.47	54.37±1.05
Uncertainty	55.47±1.00	55.84±1.09	55.48±1.37	56.01±0.58
Uncertainty/Correlation	**56.08±0.41**	**56.05±0.54**	**56.69±1.78**	**56.77±0.75**

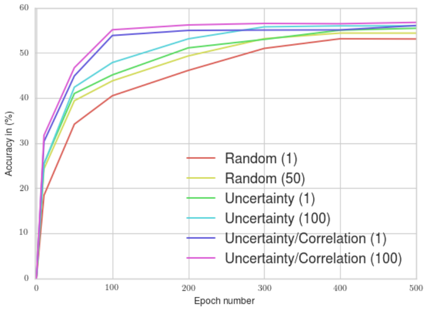

Fig. 1: Evolution of the performance as a function of the number of epochs, on the aP&Y dataset, with a neg/pos ratio of 1 and 100.

3.2 Performance as a function of the ratio of positive/negative pair

Table 2 give accuracy performances on aP&Y dataset for the three methods in function of the number of negative examples for each positive pair. Bucher *et al.* [1] configuration corresponds to random method with one negative example per positive one. Our new negative pair selection method have a strong impact on the performance with a noticeable mean improvement of 3%. Augmenting the ratio of negative pairs over positive ones has a positive influence on the accuracy.

3.3 Convergence

We also made experiments to evaluate the impact of the hard mining selection on the convergence during training. Figure 1 shows that *Uncertainty/Correlation* converges around 4 times faster than the *Uncertainty* and *Random* methods. This confirms the fact that more informative (negative) pairs are selected with this strategy. The negative/positive ratio has a (small) positive impact on the convergence.

4 Conclusions

This paper extended the original work of Bucher *et al.* [1] by proposing a novel hard negative mining approach used during training. The proposed selection strategy gives close or above state-of-the-art performance on four standard benchmarks and has a positive impact on convergence.

References

1. Bucher, M., Herbin, S., Jurie, F.: Improving Semantic Embedding Consistency by Metric Learning for Zero-Shot Classification. In: ECCV 2016, Amsterdam, Netherlands (October 2016)
2. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-Based Classification for Zero-Shot Visual Object Categorization. IEEE Trans Pattern Anal Mach Intell **36**(3) (2014) 453–465
3. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of Output Embeddings for Fine-Grained Image Classification. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
4. Romera-Paredes, B., Torr, P.H.: An embarrassingly simple approach to zero-shot learning. In: ICML. (2015) 2152–2161
5. Zhang, Z., Saligrama, V.: Zero-Shot Learning via Semantic Similarity Embedding. In: IEEE International Conference on Computer Vision (ICCV). (2015)
6. Zhang, Z., Saligrama, V.: Zero-shot learning via joint latent similarity embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 6034–6042
7. Wang, Q., Chen, K.: Zero-shot visual recognition via bidirectional latent embedding. arXiv preprint arXiv:1607.02104 (2016)
8. Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., Schiele, B.: Latent embeddings for zero-shot classification. arXiv preprint arXiv:1603.08895 (2016)
9. Fu, Y., Zhu, X., Li, B.: A survey on instance selection for active learning. Knowledge and Information Systems **35**(2) (2013) 249–283
10. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. arXiv preprint arXiv:1604.03540 (2016)
11. Li, X., Snoek, C.M., Worring, M., Koelma, D., Smeulders, A.W.: Bootstrapping visual categorization with relevant negatives. IEEE Transactions on Multimedia **15**(4) (2013) 933–945
12. Canévet, O., Fleuret, F.: Efficient sample mining for object detection. In: ACML. (2014)
13. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). (2009)
14. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). (2009)
15. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Technical report (July 2011)
16. Patterson, G., Xu, C., Su, H., Hays, J.: The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding. International Journal of Computer Vision **108**(1-2) (2014) 59–81
17. Jayaraman, D., Grauman, K.: Zero-shot recognition with unreliable attributes. In: Conference on Neural Information Processing Systems (NIPS). (2014)
18. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: ICLR. (2014)