Towards Recognizing Unseen Categories in Unseen Domains

—

Supplementary Material

Massimiliano Mancini[1,2[0000−0001−8595−9955]], Zeynep Akata[2[0000−0002−1432−7747]], Elisa Ricci[3,4[0000−0002−0228−1147]], and Barbara Caputo[5,6[0000−0001−7169−0158]]

[1] Sapienza University of Rome, [2] University of Tübingen, [3] University of Trento, [4] Fondazione Bruno Kessler, [5] Politecnico di Torino, [6] Italian Instituite of Technology
mancini@diag.uniroma1.it

1 Hyperparameter choices

In this section, we will detail the hyperparameter choices and validation protocols that, for lack of space, we did not include in the main paper.

ZSL. For each dataset, we use the train, validation and test split provided by [11]. In all the settings we employ features extracted from the second-last layer of a ResNet-101 [6] pretrained on ImageNet as image representation, *without* end-to-end training. For CuMix , we consider f as the identity function and as g a simple fully connected layer, performing the mixing directly at the feature-level while applying our alignment loss in the embedding space (*i.e.* $\mathcal{L}_{\text{M-IMG}}$ and $\mathcal{L}_{\text{M-F}}$ coincide in this case and are applied only once.) All hyperparameters have been set dataset-wise following [11], using the available validation sets. For all the experiments, we use SGD as optimizer with an initial learning rate equal to 0.1, momentum equal to 0.9, a weight-decay set to 0.001 for all settings but AWA, where is set 0. The learning-rate is downscaled by a factor of ten after 2/3 of the total number of epochs and $N = 30$. In particular, for CUB and FLO we train our model for 90 epochs, setting $\beta_{\max} = 0.8$ and $\eta_{\text{I}} = \eta_{\text{F}} = 10.0$ for CUB, and $\beta_{\max} = 0.4$ and $\eta_{\text{I}} = \eta_{\text{F}} = 4.0$ for FLO. For AWA, we train our network for 30 epochs, with $\beta_{\max} = 0.2$ and $\eta_{\text{I}} = \eta_{\text{F}} = 1.0$. For SUN, we train our network for 60 epochs, with $\beta_{\max} = 0.8$ and $\eta_{\text{I}} = \eta_{\text{F}} = 10$. In all settings, the batch-size is set to 128.

DG. We use as base architecture a ResNet-18 [6] pretrained on ImageNet. For our model, we consider f to be the ResNet-18, g to be the identity function and ω will be a learned, fully-connected classifier. We use the same training hyperparameters and protocol of [7], setting $\beta_{\max} = 0.6$, $\eta_{\text{I}} = 0.1$, $\eta_{\text{F}} = 3$ and $N = 10$.

ZSL+DG. For all the baselines and our method we employ as base architecture a ResNet-50 [6] pretrained on ImageNet, using SGD with momentum as

optimizer, with a learning rate of 0.001 for the ZSL classifier and 0.0001 for the ResNet-50 backbone, a weight decay of $5 \cdot 10^{-5}$ and momentum 0.9. We train the models for 8 epochs (each epoch counted on the smallest source dataset), with a batch-size containing 24 sample per domain. We decrease the learning rates by a factor of 10 after 6 epochs. For our model, we consider the backbone as f and a simple fully-connected layer as g. We set $N = 2$, $\eta_{\mathrm{I}} = 10^{-3}$ for all the experiments, while $\beta_{\max}$ in $\{1, 2\}$ and η_{F} in $\{0.5, 1, 2\}$ depending on the scenario.

2 ZSL+DG: analysis of additional baselines

In Table 3 of the main paper, we showed the performance of our method in the new ZSL+DG scenario on the DomainNet dataset [8], comparing it with three baselines: SPNet [10], simple *mixup* [14] coupled with SPNet and SPNet coupled with EpiFCR [7], an episodic-based method for DG. We reported the results of these baselines to show 1) the performance of a state-of-the-art ZSL method (SPNet), 2) the impact of *mixup* alone (*mixup*+SPNet) and 3) the results obtained by coupling state-of-the-art models for DG and for ZSL together (EpiFCR+SPNet). We chose SPNet and EpiFCR as state-of-the-art references for ZSL and DG respectively due to their high performances on their respective scenarios, plus because they are very recent approaches.

In this section, we motivate our choices by showing that other baselines of ZSL and DG achieve lower performances in this new scenario. In particular we show the performances of two standard ZSL methods, ALE [1] and DEVISE [4] and a standard DG/DA method, DANN [5]. We choose DANN since it is a strong baseline for DG on residual architectures, as shown in [7]. As in the main paper, we show the performances of the ZSL methods alone, ZSL methods coupled with DANN, and with EpiFCR. For all methods, we keep the same training hyperparameters, tuning only the method-specific ones. The results are reported in Table 1. As the table shows, CuMix achieves superior performances even compared to these new baselines. Moreover, these baselines achieve lower results than the EpiFCR method coupled with SPNet, as expected. This motivates our choices of the main paper. It is also worth highlighting how coupling ZSL methods with DANN for DG achieves lower performances than the ZSL methods alone in this scenario. This is in line with the results reported in [8], where standard domain alignment-based methods are shown to be not effective in the DomainNet dataset, leading also to negative transfer in some cases [8].

Finally, we want to underline that coupling EpiFCR with any of the ZSL baselines, is not a straightforward approach, but requires to actually adapt this method, re-structuring the losses. In particular, we substitute the classifier originally designed for EpiFCR with the classifier specific of the ZSL method we apply on top of the backbone. Moreover, we additionally replace the classification loss with the loss devised for the particular ZSL method. For instance, for EpiFCR+SPNet, we use as classifier the semantic projection network, using the cross-entropy loss in [10] as classification loss. Similarly, for EpiFCR+DEVISE and EpiFCR+ALE, we use as classifier a bi-linear compatibility function [11]

Table 1. ZSL+DG scenario on the DomainNet dataset with ResNet-50 as backbone.

Method		Target Domain					
DG	ZSL	clipart	infograph	painting	quickdraw	sketch	avg.
-	DEVISE [4]	20.1	11.7	17.6	6.1	16.7	14.4
	ALE [1]	22.7	12.7	20.2	6.8	18.5	16.2
	SPNet [10]	26.0	16.9	23.8	8.2	21.8	19.4
DANN [5]	DEVISE [4]	20.5	10.4	16.4	7.1	15.1	13.9
	ALE [1]	21.2	12.5	19.7	7.4	17.9	15.7
	SPNet [10]	25.9	15.8	24.1	8.4	21.3	19.1
EpiFCR [7]	DEVISE [4]	21.6	13.9	19.3	7.3	17.2	15.9
	ALE [1]	23.2	14.1	21.4	7.8	20.9	17.5
	SPNet [10]	26.4	16.7	24.6	9.2	**23.2**	20.0
CuMix		**27.6**	**17.8**	**25.5**	**9.9**	22.6	**20.7**

Table 2. Results on DomainNet dataset with *Real-Painting* as sources and ResNet-50 as backbone.

Method/Target	Clipart	Infograph	Sketch	Quickdraw	Avg.
SPNet	21.5±0.6	14.1±0.2	17.3±0.3	4.8±0.4	14.4
Epi-FCR+SPNet	22.5±0.5	14.9±0.7	18.7±0.6	**5.6**±0.4	15.4
MixUp img only	21.2±0.4	14.0±0.7	17.3±0.3	4.8±0.1	14.3
MixUp two-level	22.7±0.3	16.5±0.4	19.1±0.4	4.9±0.3	15.8
CuMix reverse	22.9±0.3	15.8±0.2	18.2±0.3	4.8±0.5	15.4
CuMix	**23.7**±0.3	**17.1**±0.2	**19.7**±0.3	5.5±0.3	**16.5**

coupled with a pairwise ranking objective [4] and with a weighted pairwise ranking objective [1] respectively.

3 ZSL+DG: ablation study

In order to further investigate our design choices on the ZSL+DG setting, we conducted experiments on a challenging scenario where we consider just two domains as sources, i.e. Real and Painting. The results are shown in Table 3. On average our model improves SPNet by 2% and SPNet + Epi-FCR by 1.1%. Our approach without curriculum largely outperforms standard image-level mixup [14] (more than 2%). Applying mixup at both feature and image level but without curriculum is effective but achieves still lower results with respect to our CuMix strategy (as in Tab. 2). Interestingly, if we apply the curriculum strategy but switching the order of semantic and domain mixing (CuMix reverse), this achieves lower performances with respect to CuMix, which considers domain mixing harder than semantic ones. This shows that, in this setting, it is important to correctly tackle intra-domain semantic mixing before including inter-domain ones.

Table 3. ZSL results.

Method	CUB	SUN	AWA1	FLO
ALE [1]	54.9	58.1	59.9	48.5
SJE [2]	53.9	53.7	65.6	53.4
SYNC [3]	56.3	55.6	54.0	-
GFZSL [9]	49.3	60.6	**68.3**	-
SPNet [10]	56.5	60.7	66.2	-
Baseline	52.4	58.2	62.5	58.4
CuMix	**60.4**	**62.4**	64.0	**59.7**

4 ZSL results

In this section, we report the ZSL results in tabular form. The results are shown in Table 3. With respect to Figure 3 of the main paper, in the table, we also report the results of a baseline which uses just the cross-entropy loss term (similarly to [10]), without the mixing term employed in our CuMix method. As the table shows, our baseline is weak, performing below most of the ZSL methods in all scenarios but FLO. However, adding our mixing strategy allows to boost the performances in all scenarios, achieving state-of-the-art performances in most of them. We also want to highlight that in Table 3, as in the main paper, we do not report the results of methods based on generating features of unseen classes for ZSL [12, 13]. This choice is linked to the fact that these methods can be used as data augmentation strategies to improve the performances of any ZSL method, as shown in [12]. While using them can improve the results of all the baselines as well as CuMix , this falls out of the scope of this work.

References

1. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for attribute-based classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 819–826 (2013)
2. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2927–2936 (2015)
3. Changpinyo, S., Chao, W.L., Gong, B., Sha, F.: Synthesized classifiers for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5327–5336 (2016)
4. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Advances in neural information processing systems. pp. 2121–2129 (2013)
5. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research **17**(1), 2096–2030 (2016)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

7. Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y.Z., Hospedales, T.M.: Episodic training for domain generalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1446–1455 (2019)
8. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1406–1415 (2019)
9. Verma, V.K., Rai, P.: A simple exponential family framework for zero-shot learning. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 792–808. Springer (2017)
10. Xian, Y., Choudhury, S., He, Y., Schiele, B., Akata, Z.: Semantic projection network for zero-and few-label semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8256–8265 (2019)
11. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence **41**(9), 2251–2265 (2018)
12. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5542–5551 (2018)
13. Xian, Y., Sharma, S., Schiele, B., Akata, Z.: f-vaegan-d2: A feature generating framework for any-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10275–10284 (2019)
14. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018)