Interpretable Open-Set Domain Adaptation via Angular Margin Separation

Xinhao Li[1], Jingjing Li[1,2](✉), Zhekai Du[1], Lei Zhu[3], and Wen Li[1]

[1] University of Electronic Science and Technology of China
Mc1114207918@outlook.com, lijin117@yeah.net, dzk1996411@163.com,
liwenbnu@gmail.com
[2] Institute of Electronic and Information Engineering of UESTC in Guangdong
[3] Shandong Normal University
leizhu0608@gmail.com

A Appendix

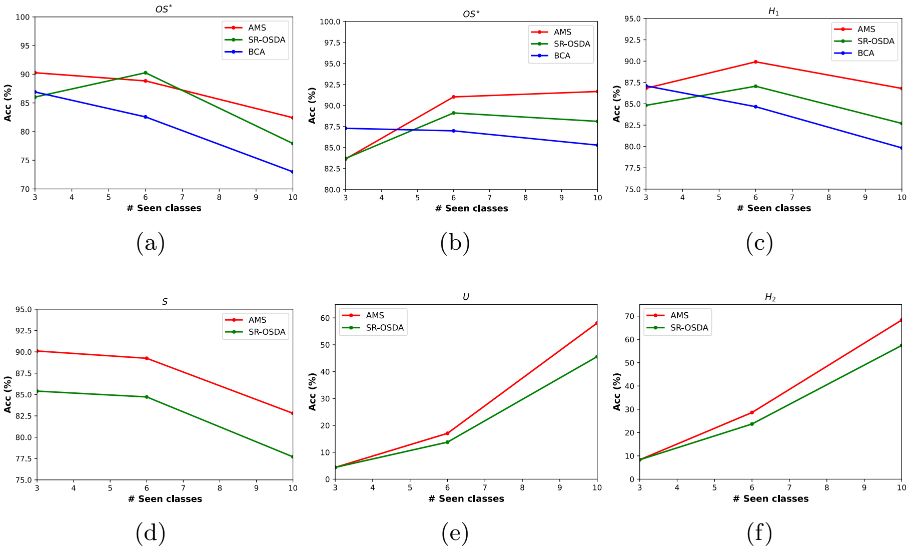
A.1 Openness Analysis

Since the relation between the number of seen and unseen classes could be uncertain in reality, it is necessary to further verify that our proposed AMS can perform robustly under different openness: $\mathbb{O} = 1 - \frac{|C_s|}{|C_t|}$. To achieve this, we set up three different configurations of task P→R, in which the number of seen classes is 3, 6, and 10, respectively, and the corresponding number of unseen classes is 14, 11, and 7. Therefore, the openness value in each scenario is 0.82, 0.64, and 0.41. For performance on open-set domain adaptation, we take the recent SR-OSDA framework and the state-of-the-art OSDA method BCA as baselines. As is shown in Fig. 1 (a), (b), and (c), AMS achieves superior or competitive results on OS^*, $OS^\diamond$, and H_1 under all openness values. For performance on semantic recovery, we only compare with SR-OSDA since BCA is a pure OSDA method. Fig. 1 (d), (e), and (f) show that our method still consistently outperforms SR-OSDA by a notable margin in terms of all metrics.

A.2 Qualitative Evaluation of Semantic Recovery

We qualitatively evaluate the capability of AMS to interpret novelties by showcasing its semantic recovery results in Fig. 2. To emphasize the interpretation of *unseen* novelties, we test a model trained in task R→A with samples from the *unseen* classes in the AwA2 dataset. Moreover, the rightmost two images in the last row are from classes *outside* the R and A domains, which are used to further confirm that AMS can reasonably interpret completely new novelties encountered in deployment. From Fig. 2 we can see that AMS can not only correctly recover salient attributes of unseen classes (notes in black), but also reasonably and flexibly predict attributes for individual samples (notes in green), which corroborates that AMS is a practical step towards interpretable OSDA.

We also compare the semantic recovery results of AMS and SR-OSDA by visualizing their confusion matrices in task R→A. From Fig. 3 we can observe



(a) (b) (c)

(d) (e) (f)

Fig. 1: Performance in task P→R under three different openness values: 0.82, 0.64, and 0.41, with the number of seen classes set to 3, 6, and 10, respectively. Best viewed in color.

that AMS is better at recovering semantic attributes for both seen and unseen classes, indicating that it is more capable of object interpretation.

A.3 Parameter Analysis

Since the functionality of AMS is the joint result of partial alignment, angular margin separation, and visual-semantic projection, it is desirable for us to investigate the role and behavior of each component separately. In this regard, we conduct parameter analysis for our 4 hyper-parameters: λ_1, λ_2, λ_3, and m, respectively. To fully disentangle different factors, when testing each hyper-parameter, we keep the other ones fixed at the values specified in section 4.1.

Fig. 4 (a) shows that the performance on all metrics is stable when λ_1 is no greater than 0.1. When λ_1 is larger than 0.1, both the recognition result $OS^\diamond$ and semantic recovery result U of *unseen* class start to drop rapidly. We argue that when partial alignment becomes too potent, it could forcefully align unseen classes with seen classes, damaging the performance of the former who have no explicit supervision information.

Fig. 4 (b) shows that the performance is relatively stable when λ_2 is between 0.05 and 0.3. Out of that range, diverse unseen classes are either under-separated or over-separated, leading to an accuracy drop in their semantic recovery result U.

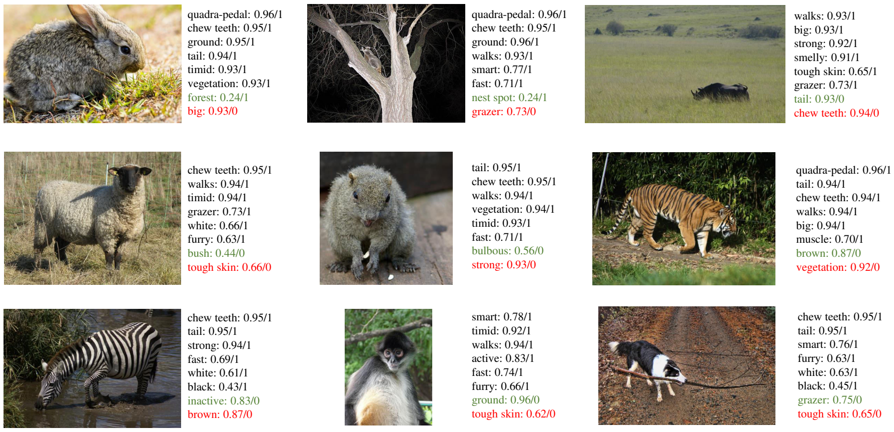


Fig. 2: Selected samples from the AwA2 dataset and attributes recovered by our model trained in task R→A. The black ones are correctly recovered attributes, green ones are reasonable predictions, and red ones are incorrect predictions. X/Y denotes *predicted value/class-prototypical value*. Best viewed in color.

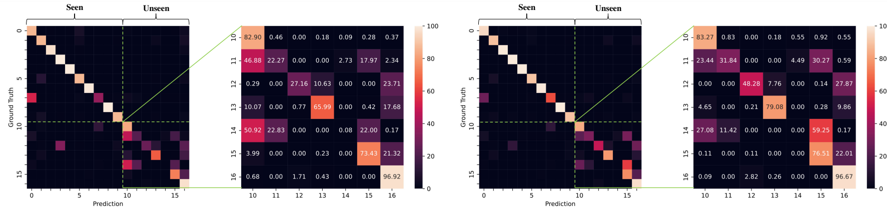
Table 1: Davies–Bouldin index (DBI) on D2AwA and I2AwA (calculated using cosine distance).

tasks	A→P			A→R			P→A			P→R			R→A			R→P			I→A		
method	$\lvert C_s\rvert+1$	$\lvert C_s\rvert+\lvert C_t\rvert$	$\lvert C_t\rvert$	$\lvert C_s\rvert+1$	$\lvert C_s\rvert+\lvert C_t\rvert$	$\lvert C_t\rvert$	$\lvert C_s\rvert+1$	$\lvert C_s\rvert+\lvert C_t\rvert$	$\lvert C_t\rvert$	$\lvert C_s\rvert+1$	$\lvert C_s\rvert+\lvert C_t\rvert$	$\lvert C_t\rvert$	$\lvert C_s\rvert+1$	$\lvert C_s\rvert+\lvert C_t\rvert$	$\lvert C_t\rvert$	$\lvert C_s\rvert+1$	$\lvert C_s\rvert+\lvert C_t\rvert$	$\lvert C_t\rvert$	$\lvert C_s\rvert+1$	$\lvert C_s\rvert+\lvert C_t\rvert$	$\lvert C_t\rvert$
w/o fine-tuning	7.95	6.78	5.85	4.13	3.41	2.20	3.41	2.88	1.59	4.11	3.43	2.22	3.41	2.88	1.59	7.82	6.74	5.93	3.43	5.14	3.69
fine-tuning	5.32	5.24	5.46	2.36	2.46	1.99	2.33	2.29	1.55	2.42	2.49	2.02	1.94	1.95	1.44	4.55	4.43	4.43	2.56	4.38	2.69

Fig. 4 (c) shows that semantic recovery result on *unseen* classes U first rises as λ_3 increases from 0.01 to 0.8, and starts to drop when λ_3 surpasses 1.5. We analyze that such phenomenon is because when λ_3 is small, the semantic recovery objective is insufficiently optimized, while when λ_3 is too large, the visual-semantic projection ϕ is overfitting to *seen* classes. Interestingly, when λ_3 becomes too large, recognition result on *unseen* classes $OS^{\diamond}$ also drops. We argue that this is because in multi-modality training, the performance on the semantic modality could also affect that on the visual modality, and vice versa.

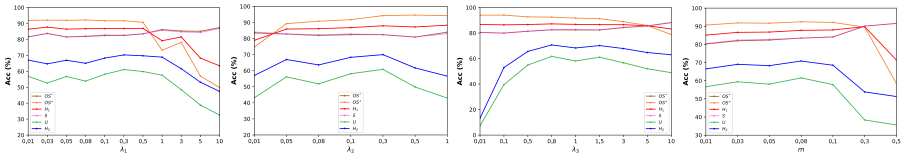
Fig. 4 (d) shows that performance on all metrics remains stable when the angular margin m is no greater than 0.1. Otherwise, performance on both the recognition result $OS^{\diamond}$ and semantic recovery result U of *unseen* class begins to drop due to brute-force alignment with *seen* classes.

The above parameter analysis not only unveils the role and behavior of each constituent of our proposed AMS but also demonstrates that it can perform stably and robustly with each relevant hyper-parameter varying in a range, which confirms that its superior performance is the result of algorithmic novelty but hyper-parameters tuning.



(a) (b)

Fig. 3: Confusion matrix in task R→A. (a) SR-OSDA. (b) AMS. The unseen classes are zoomed in for better visualization. Best viewed in color.



(a) (b) (c) (d)

Fig. 4: Parameter analysis in task P→R. (a) λ_1. (b) λ_2. (c) λ_3. (d) Margin m. Best viewed in color.

A.4 Cluster Property Analysis

After the proposed fine-tuning phase, the features of seen classes tend to become angularly discriminative and intra-class compact, and we could also expect the features of unseen classes to have better clustering properties as well and not fall into the area of seen classes. By this, we primarily mean that the overall unseen class composed of all unseen classes is more compact and separate from seen classes, compared with under naive Softmax. On top of that, we could also expect different unseen classes inside the overall unseen class to also cluster better. Conceptually speaking, thanks to the angular prototype regularization, the decision boundary is *less overfitting* to seen classes, leaving out an open space, wherein rather than falling dispersedly into areas of seen classes, unseen classes are tightly bounded and better stay together. Hence, the overall unseen class is more compact and farther away from all seen classes. Besides, experimental evidence in the metric-based few-shot learning literature shows that optimizing prototype-based metrics could facilitate class-discriminative features when generalizing to unseen classes. Thus, we could expect that our angular prototype objective also brings such merit. For concrete evidence, we compare the Davies–Bouldin index (DBI), a popular metric for cluster quality evaluation, between with and without our fine-tuning phase in table 1, where we achieve smaller (better) DBI in terms of target domain "seen classes+1 overall unseen class", "seen classes+unseen classes", "unseen classes", verifying that indeed the overall unseen class clusters better against seen classes, and different unseen classes also cluster better against each other.

A.5 More Implementation Details

In addition to the specifications of neural network architecture and hyper-parameter values mentioned in our paper, we set the initial learning rate η_0 of the backbone to 0.001 and all other networks to 0.01, and the learning rate is adjusted by $\eta_p = \eta_0(1 + \alpha p)^{-\beta}$, where p is the training progress changing from 0 to 1, and $\alpha = 10$, $\beta = 0.75$. We adopt mini-batch SGD using momentum 0.9, weight decay 0.001, and Nesterov accelerated gradient. We implement our codes with PyTorch on two NVIDIA GeForce RTX 2080Ti GPUs and average results from 5 random runs. Please note that [2] reports OS: Avg Acc on $|C_s| + 1$ classes rather than the H_1 in our work which is more appropriate according to [1]. OS simply regards all unseen classes as one class, whose Acc is averaged into many seen classes, and thus fails to give unseen classes discovery due significance. That's one reason for seeing different figures in our report to those in [2]. For OS, our method is also notably superior for generally higher Acc on both seen and unseen classes. Besides, [2] has not released code, and our re-implementation in fact achieves higher or comparable H_1 in 5 out of 7 tasks, H_2 in 6 out of 7 tasks (we use classwise rejection for fair comparison with our method), which makes this baseline even stronger. Codes for our paper and our re-implementation of [2] are available at AMS.

References

1. Bucci, S., Loghmani, M.R., Tommasi, T.: On the effectiveness of image rotation for open set domain adaptation. In: European Conference on Computer Vision. pp. 422–438. Springer (2020)
2. Jing, T., Liu, H., Ding, Z.: Towards novel target discovery through open-set domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9322–9331 (2021)