Open-World Learning Under Dataset Shift

Ponhvoan Srey
Nanyang Technological University
ponhvoan.srey@gmail.com

Yuhui Zhang
Tokyo Institute of Technology
zhang.y.av@m.titech.ac.jp

Takafumi Kanamori
Tokyo Institute of Technology/RIKEN
kanamori@c.titech.ac.jp

***Abstract*—Conventional classification models in machine learning are imposed with strict constraints, limiting their implementation in real-world scenarios. Datasets encountered in the wild naturally contain instances belonging to both classes seen in the training data, and unseen novel classes. Further, test data are frequently drawn from a different distribution compared to the training data. In this paper, with the aim of developing a more universal classifier, we explore the challenging but practical setting where we allow for both the open-world nature and a shift in distribution between training and test data. At the core, we build upon the vanilla open-world semi-supervised model by introducing a novel weighted variance minimization term for the unlabeled data. This regularization term improves generalization by encouraging the model to be *cautiously* confident, *i.e.* to output more confident predictions for instances of known classes, but not for those of unknown classes. We demonstrate the effectiveness of our method, especially for seen classes, on benchmark datasets. Our simple approach can be easily integrated in other existing open-world frameworks, and even beyond semi-supervised learning. Most importantly, our work advocates for more research that improves open-world models by looking into the rich literature of domain adaptation, thereby developing reliable open-world learning systems that are resilient to distributional shifts.**

***Index Terms*—open-world, semi-supervised learning, domain adaptation**

I. Introduction

Remarkable performance has been achieved in deep learning through the use of large quantities of labeled data in a multitude of tasks, but the vast majority of data available is unlabeled. The cost of manually labelling is prohibitive potentially due to privacy issues or the need for expert knowledge. In semi-supervised learning (SSL), both labeled and unlabeled data are incorporated into the training procedure, with the goal of surpassing the performance when trained in either a purely supervised or purely unsupervised fashion.

In this paper, we are concerned with classification under the SSL framework. In particular, we work under the transductive learning assumption, where the model learns from the combined dataset and aims to classify the unlabeled samples. Here, we focus on open-world classification, where novel classes unseen in the training data may appear in the test data. Standard classification models are ill-equipped to tackle the open-world setting because the labeled and unlabeled sets are assumed to overlap. When dealing with data in-the-wild, the closed-set assumption rarely holds. A robust and reliable model needs to be able to account for novel classes as well.

To construct a universal classifier, we further relax the fatal assumption in conventional machine learning models that the training and test domains are independent and identically distributed (*i.i.d.*). Our primary objective is to improve the learning of models for open-world classification under the SSL paradigm, specifically in the challenging case when there is a shift between the training and test domains. We propose Open-world Domain Adapted Classification (ODAC), a unified training method to handle this setting. The key to our approach is the introduction of a novel weighted variance minimization term that enforces the model to be *cautiously confident*, and output more certain predictions for data that belong to classes seen in the training data, but not for those of unknown novel classes. We find the simple approach to be useful in helping to learn a more compact representation, and achieve better accuracy, especially for the seen classes. We evaluate our method against the vanilla model on benchmark datasets.

II. Related Work

Open-world learning: Open-world learning [1]–[4] operates under the open-world setting where novel classes appear in the test data. The first paper to address the open-world nature [2] focused on preventing the model from overfitting to the seen classes. Subsequent works such as [1], [3] deviated in their goals, such as aiming to learn a compact feature representation, and working on the estimation of number of novel classes. A potential extension would be to combine the strengths of the different methods, such as using [3] as a pretraining protocol to [2]. However, these works have not explicitly considered a dataset shift from training to test data.

Domain adaptation: In domain adaptation, the training and test domains are different, and the model is required to adapt to and perform well on the test data. The common approach is to learn domain invariant representation by minimizing the divergence between the feature distributions of the source (training) data and the target (test) data [5]. Prior methods worked on various definitions of divergence between feature distributions [6]–[8]. Operating under the same setting as our own is universal or open-set domain adaptation [9]–[11] which accounts for novel classes in the test data. Nonetheless, their chief objective is to faithfully align feature distributions, they focus on representation learning, and often require an additional step to train a classifier. Meanwhile, ODAC is an end-to-end framework concerned with classification.

Semi-supervised learning (SSL) SSL methods utilized unlabeled data to improve the performance of models trained only with labeled data. Various works have explored either making the model either more certain [12] or less certain [13] when predicting on the unlabeled data.

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00188

III. Proposed Approach: ODAC

In the open-world setting with dataset shift, we have the source labeled dataset $\mathcal{D}_l = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and the target unlabeled dataset $\mathcal{D}_u = \{(\mathbf{x}_j)\}_{j=1}^m$, each $\mathbf{x} \in \mathbb{R}^k$, usually with $n \ll m$, such that $\mathcal{D}_l$ and $\mathcal{D}_u$ are drawn from different distributions. The task is to assign instances from $\mathcal{D}_u$ to either a seen class $c_s \in \mathcal{C}_s$ or a novel class $c_n \in \mathcal{C}_n$, assuming the number of novel classes is known. The network architecture we adopt consists of (i) a backbone embedding function $f_\theta : \mathbb{R}^k \to \mathbb{R}^d$, mapping inputs $\mathbf{x} \in \mathcal{X}_l \cup \mathcal{X}_u$ to obtain features $\mathbf{z} \in \mathcal{Z}_l \cup \mathcal{Z}_u$, and (ii) a linear classifier with weight W.

To tackle the problem of Open-World Learning under dataset shift, we propose ODAC, which operates on top of ORCA [2]. ORCA trains the model on the combined dataset using a composition of supervised and unsupervised objectives. ODAC combines the vanilla ORCA with a novel regularisation term to minimise the variance of the model predictions on the unlabeled dataset. Encouraged to make more confident predictions on the unlabeled data, the model effectively learns discriminative features from labeled data that result in low-variance predictions for the unlabeled data as well [12]. This advantage would likely be more apparent in our case, as the unlabeled data provides greater variation due to the dataset shift.

However, there are both known and unknown classes in the target unlabeled data. Thus, while we would like our network to be more confident in its predictions for unlabeled data of known classes to reduce overfitting to labeled data, we do not want it to be overconfident for unlabeled data of unknown classes. Therefore, we propose to minimise the variance of predictions only for unlabeled data of known classes to avoid misguiding the model.

Since we do not have information on which instances in the unlabeled dataset belong to the unknown classes, outright rejection of samples would likely propagate error. We avoid this by applying a soft weight on the unlabeled samples: for those samples deemed to belong to an unknown class, their weight should be smaller, and vice versa. Under the mild assumption that if feature $\mathbf{z}$ belongs to a seen class, the uncertainty of the predictions will be low. We model the uncertainty using entropy. Equivalently, if the entropy is high, $\mathbf{z}$ is likely to belong to the unknown class. Here, the classifier only predicts for known classes. To avoid training a separate classifier, we take the first $|\mathcal{C}_l|$ softmax outputs of $\sigma_k(W^T \cdot \mathbf{z})$, where σ is the sigmoid function, and truncate the rest, and normalize these outputs such that their sum is 1.

With this assumption, we can define the weight in relation to the entropy of the classifier:

$$w(\mathbf{z}) = \frac{1}{W} \exp\left(\sum_{k=1}^{|\mathcal{C}_l|} -\hat{p}_k \log \hat{p}_k\right),$$

where $\hat{p}_k = \sigma_k(W^T \cdot \mathbf{z})$, and W is a normalizing constant such that $\sum w(\mathbf{z}) = 1$. We define the additional variance minimization term as

$$\gamma \sum_{\mathbf{z} \in \mathcal{Z}_u} w(\mathbf{z}) \operatorname{Var}(\mathbf{z}),$$

where $\gamma > 0$ is a balancing hyperparameter, $\operatorname{Var}(\mathbf{z}) = 1 - \max_k P(y = k|\mathbf{z}) = 1 - \max_k \sigma_k(W^T \cdot \mathbf{z})$. Additionally, we employ domain-specific batch normalization (DSBN) to address the domain shift [14].

TABLE I: Mean accuracy over three different corruption types for CIFAR10-C, CIFAR100-C, and VisDA datasets

Method	CIFAR10-C			CIFAR100-C			VisDA		
	Seen	Unseen	Overall	Seen	Unseen	Overall	Seen	Unseen	Overall
ORCA	7.1	14.2	11.7	1.88	2.73	1.83	2.0	36.2	24.9
ODAC	19.6	20.5	13.7	1.97	2.12	1.42	10.2	29.8	22.3

IV. Experimental Results

Experimental Setup. We test on three benchmarks:

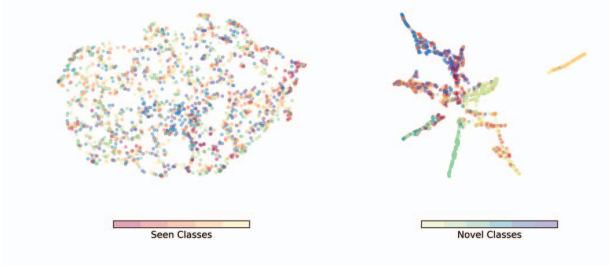
- CIFAR10-C and CIFAR100-C [15]: we consider three corruptions, namely snow, gaussian noise and glass blur.
- VisDA [16]: a large synthetic-to-real dataset of 12 classes.

For all datasets, 50% of the total number of classes are unseen in the labeled dataset, and the unlabeled dataset is the same size as the labeled one. We use ResNet18 [17] as the network and leverage the model weights pretrained on ImageNet. We note that a different pretraining protocol, particularly one that works in the open-world such as [3], could boost performance. Lastly, we apply DSBN by simply forwarding the source and target data in separate mini-batches as in [18].

Evaluation. We compare the seen, unseen, and overall accuracy of ODAC to ORCA in TABLE I. ODAC considerably outperforms the vanilla model on the seen classes, especially for CIFAR10-C which shows an improvement of 176%. This improvement is in line with what is observed in the closed set scenario when we enforce a model to be more confident in its predictions. However, besides CIFAR10-C, ODAC does not perform competitively in the unseen and overall accuracy. We observe that the accuracy is sensitive to the weight hyperparameter γ, and so, ODAC may require more finetuning. Further, we visualize the feature embeddings by ODAC compared to OpenCon [3] on the VisDA dataset in Fig. 1. Under the dataset shift, OpenCon is unable to achieve a compact representation, whereas ODAC shows better embeddings in more distinct clusters.

V. Conclusion

We improve on the vanilla open-world model by incorporating a novel weighted variance minimization term in the objective. We extend the idea of enforcing the model to be more certain in its predictions, which is often used in semi-supervised learning, to the open-world setting. We introduce the notion of *cautious confidence*, whereby the model is encouraged to predict more confidently on instance of the seen classes, but not on those of the novel classes. This simple regularization term requires no ground truth labels, and



(a) OpenCon (b) ODAC

Fig. 1: UMAP visualization [19] of feature embeddings for the VisDA dataset by (a) OpenCon, and (b) ODAC.

could be easily adopted even beyond SSL, such as in test-time adaptation. Future work could formulate the confidence differently, such as in terms of entropy, instead of our naive formulation of variance. Furthermore, we utilize DSBN, a simple approach to address the dataset shift. More scrutiny of the domain adaptation literature could be inspiring to effectively align feature distributions and enhance accuracy.

Our preliminary experiments are promising, especially on the seen classes, though the overall performance is still not satisfactory. Nonetheless, our work brings to attention the practical setting of open-world learning under dataset shift, and showcases how existing methods can be extended to develop reliable open-world learning systems.

Acknowledgment

This research was supported by JSPS KAKENHI Grant Number 19H04071, 20H00576, and 23H03460. P. Srey gratefully acknowledges the CN Yang Scholars Programme at the Nanyang Technological University for the financial support.

References

[1] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Generalized category discovery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7492–7501.

[2] K. Cao, M. Brbic, and J. Leskovec, "Open-world semi-supervised learning," *arXiv preprint arXiv:2102.03526*, 2021.

[3] Y. Sun and Y. Li, "Opencon: Open-world contrastive learning," *Transactions of Machine Learning Research*,

[4] M. N. Rizve, N. Kardan, and M. Shah, "Towards realistic semi-supervised learning," in *European Conference on Computer Vision*, Springer, 2022, pp. 437–455.

[5] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," *Advances in neural information processing systems*, vol. 19, 2006.

[6] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*, PMLR, 2015, pp. 97–105.

[7] Y. Ganin, E. Ustinova, H. Ajakan, *et al.*, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[8] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3723–3732.

[9] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Universal domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2720–2729.

[10] Y. Xu, L. Chen, L. Duan, I. W. Tsang, and J. Luo, "Open set domain adaptation with soft unknown-class rejection," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[11] K. Saito, D. Kim, S. Sclaroff, and K. Saenko, "Universal domain adaptation through self supervision," *Advances in neural information processing systems*, vol. 33, pp. 16 282–16 292, 2020.

[12] N. Jean, S. M. Xie, and S. Ermon, "Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[13] A. Chan, A. Alaa, Z. Qian, and M. Van Der Schaar, "Unlabelled data improves bayesian uncertainty calibration under covariate shift," in *International Conference on Machine Learning*, PMLR, 2020, pp. 1392–1402.

[14] W.-G. Chang, T. You, S. Seo, S. Kwak, and B. Han, "Domain-specific batch normalization for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2019, pp. 7354–7362.

[15] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.

[16] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko, *Visda: The visual domain adaptation challenge*, 2017. eprint: arXiv:1710.06924.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8050–8058.

[19] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.