Instance norm improves meta-learning in class-imbalanced land cover classification

Marc Rußwurm
EPFL-ECEO Laboratory
marc.russwurm@epfl.ch

Devis Tuia
EPFL-ECEO Laboratory
devis.tuia@epfl.ch

Abstract

Distribution shift is omnipresent in geographic data, where various climatic and cultural factors lead to different representations across the globe. We aim to adapt to new data distributions with only a few annotated samples using model-agnostic meta-learning, where data sampled from each distribution is seen as an individual task. Transductive batch normalization layers are often employed in meta-learning models, as they reach the highest numerical accuracy on the class-balanced target tasks used as meta-learning benchmarks. In this work, we demonstrate empirically that transductive batch normalization collapses when deployed on a real class-imbalanced land cover classification problem. We propose a solution to replace batch normalization with instance normalization. This modification consistently outperformed all other normalization alternatives across different meta-learning algorithms in our class-imbalanced land cover classification test tasks.

1 Introduction

Distribution shifts are abundant in Earth observation [13]. Local climates shape different regions with varying yearly temperatures and precipitation, leading to specific flora and fauna. Similarly, different societies and cultures shape the appearance of our cities and infrastructure. Categorizing this diversity into a set of distinct classes is the objective of land cover classification, which quantifies the extent of urban areas, forests, and wetlands at a global scale to inform, for instance, climate models. We illustrate this problem in fig. 1 that summarizes this work's main application, problem, and approach. For example, we show the distribution shift in Tanzania and South Africa in fig. 1b: the frequency and presence of some land cover classes and the appearance of samples of the same class substantially differ between these countries, leading to label and covariate shifts, respectively.

Land cover classification models are usually trained on a single combined dataset from different geographic areas. This strategy effectively accounts for the regional diversity by incorporating these possible representations of the Earth surface in a single dataset. However, this big-data approach has several limitations. First, sampling the world uniformly and capturing all variation is not possible, as obtaining accurately annotated land cover images is easier in some geographic areas, e.g., in Europe or US, where these statistics are collected on a governmental level. This introduces systematic biases and disadvantages areas that are underrepresented in the combined training dataset. Second, incorporating accurate representations of all possible appearances of one class in a single model also requires immense model capacity. While building these large foundation models with a large number of parameters is definitely feasible, they are often impractical. They are also inefficient and unnecessarily complex, when we are interested in predicting one geographic area only. Further, using a single global model is also inefficient in label space, as some classes (e.g., savanna and snow) will never co-appear in one geographic area. Instead, we argue that representing globally distributed data as a dataset of tasks reflects more closely the regional variability in global land cover classification problems. Each task contains data samples from one joint distribution of labels and data divided into training and testing partitions of a few examples each. Crucially, the data

Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

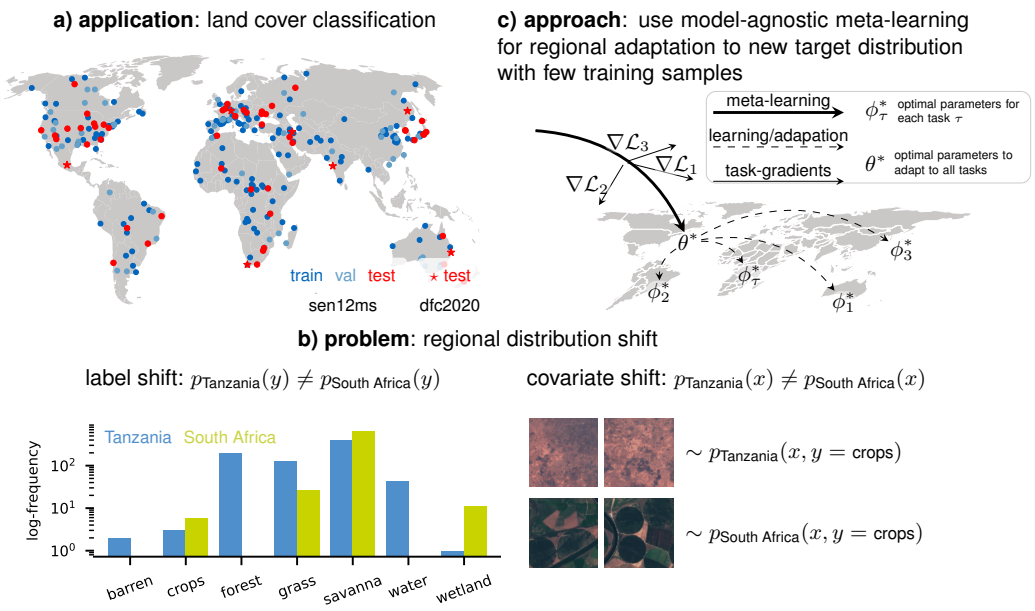

Figure 1: Illustration of the application, problem and approach in this paper.

distribution can vary between tasks that capture each geographic region's regional variability. In this meta-learning framework, a meta-model captures the commonality between different tasks across the globe and is fine-tuned to the specific characteristics of each area, as shown in fig. 1c. Learning from different-but-related problems is the objective of meta-learning: the model-agnostic meta-learning algorithm [2] that we use in this paper trains a model explicitly to learn the characteristics of new unseen data distributions represented by with few data samples.

Most few-shot models trained with model-agnostic meta-learning [2, 15] use batch normalization layers to achieve remarkable numerical results on established benchmark datasets, such as Omniglot [6] or TieredImagenet [9]. The highest numerical accuracy is achieved in a transductive setting when the batch statistics are calculated from the task-dataset at test-time. However, this use of batch normalization makes meta-learning models sensitive to the distribution over the target set, as first identified and discussed by Bronskill and colleagues [1]. They proposed to use specific task normalization layers (tasknorm-I) to address this issue and demonstrated their effectiveness on various benchmark datasets in the MetaDataset [12]. In this work, we propose to replace batch normalization with instance normalization, which yielded more robust results in class-imbalanced and realistic scenarios that we present in the experimental section.

2 Method

We train a ResNet-12 [3] following the implementation of [8] with the model-agnostic meta-learning (MAML) [2] algorithm that optimizes the following objective

$$\underbrace{\min_{\theta} \mathbb{E}_{\tau \sim p(\tau)} \left[L_{\tau}^{\text{test}} \left(\phi_{\tau,K} \left(\theta \right) \right) \right]}_{\text{outer loop/meta-learning}} s.t. \underbrace{\phi_{\tau,k+1} \leftarrow \phi_{\tau,k} - \alpha \nabla L_{\tau}^{\text{train}}}_{\text{inner loop/fine-tuning}} \text{ and } \underbrace{\phi_{\tau,0} = \theta.}_{\text{initialization}} \tag{1}$$

A task-model ϕ_τ is initialized with the parameters of the meta-model θ. In the inner loop, it is fine-tuned with $k \leq K$ steps based on gradients from a loss of training samples ∇L^{train}. The constant α denotes the inner learning rate.

The outer loop updates the meta-model parameters θ to minimize the test loss test loss L_{τ}^{test} over a batch of training tasks $\mathbb{E}_{\tau \sim p(\tau)}$. This yields a meta-model that is explicitly *learned to learn* to adapt to new data distributions with few training samples from the target distribution. We aim to deploy the meta-learning model on downstream target tasks with a different number of classes. We allow this flexibility by training the meta-model as a binary classifier in a one-versus-all setting. To do so, we

first sample k-shot n-way tasks and randomly select one of the four classes as the target class. At test-time, we fine-tune one one-versus-all classifier for each class present in the downstream task and ensemble these binary classifiers for multi-class classification on a variable number of classes.

Normalization layers are required in deep convolutional networks to obtain meaningful solutions. Without normalization, changes in the weights of an early layer cause feature-distribution shifts that cascade and amplify through deeper layers, which inhibits learning. Normalization layers counteract this behavior by normalizing features by their the first two moments that reduce this internal covariate shift [5] within the network. The central difference between normalization layers are the calculation of these moments: in conventional batch normalization [5], these moments are calculated dynamically during training from the entire training batch, learned internally as weights, and then retrieved at test-time. This strategy assumes that data distributions (and their moments) are identical at training and testing time, which not always true in a few-shot meta-learning setting with where data of individual source and target tasks can originate from different distributions. This motivated the use of transductive batch normalization in [2] where the normalization moments are always calculated from the current batch. This, however, makes the model sensitive to the distribution over the target set [1] which does not become apparent class-balanced meta-learning benchmark datasets. In contrary, the dynamic calculation of statistics over all samples within one task (i.e., one complete batch) improves the apparent model accuracy by exploiting knowledge between samples across the task (transductive setting) [1]. In instance normalization [4], the moments are calculated for each instance independently. This non-transductive normalization ensures that the model is independent of the composition of samples in the batch but still reduces the internal covariance shift within the network to train a residual network to a meaningful solution.

3 Datasets

We use two land cover classification datasets that contain annotated satellite images from different globally distributed locations, as shown in fig. 1a.

The *Sentinel-12 Multi-Spectral (Sen12MS)* [11] dataset covers 125 globally distributed geographic regions from different seasons. We use 75 regions for training and 25 for validation and testing. The 128×128px multi-spectral images with 12 spectral bands are classified into 10 simplified International Geosphere Biosphere Program (IGBP) [7] land cover classes. We sample class-balanced few-shot tasks from this dataset where we select four land cover classes within one region and two examples per class from one region resulting in 2-shot 4-way tasks for train and test partitions of 8 images each. One of the four classes is selected randomly as the target class to construct a one-versus-all task. *Data Fusion Contest 2020 (DFC2020)* dataset [10] contains seven regions, follows the same IGBP label scheme, but made use of higher-resolution (segmentation) labels. When modified for image classification, this leads to lower label noise than Sen12MS and makes this dataset best suited for evaluation. We aim to classify all images within DFC2020 regions to create continuous land cover maps. To do so, we split all images of one region into a class-balanced few-shot training set and assigned all remaining images to the class-imbalanced test set. Note that the Sen12MS test tasks are class-balanced and have an identical composition to the training tasks (with images from different regions). These tasks reflect how well the meta-learning model can generalize to images of different regions. However, sampling these tasks requires knowledge of the labels at test time, which we do not have in practice. This is why we will focus our discussion on the performance in the seven DFC2020 tasks, which are heavily class-imbalanced and reflect a more realistic setting.

4 Experiments

In Experiment 1 (table 1a), we first followed standard practices in meta-learning benchmarks and trained the binary ResNet-12 model with batch normalization layers in the transductive batch normalization setting (*transd. BN* hereafter). This model achieved excellent accuracy (85%) when tested on Sen12MS test tasks (see table 1a first row, left column). However, the performance collapsed to 26% (indicated in red) on the realistic DFC2020 tasks (see table 1a, right column). This discrepancy between best performance in a class-balanced setting and failure in realistically imbalanced target tasks is remarkable given that using transductive batch normalization is commonly used in meta-learning benchmarks [1]. Following [1], who first discussed this issue, we find that training MAML in a non-transductive setting with conventional batch normalization (*convent. BN* hereafter) leads to better results on DFC2020 of 60%, but still leads to a slight drop in performance of 1% when applied

metric: **accuracy**	task-datasets	
	Sen12MS	DFC2020
number of tasks	1000	7
task design	idealized	realistic
label distribution	balanced	imbalanced
transd. BN [2]	**0.85**	**0.26**
convent. BN [5]	0.84	0.60
tasknorm-I [1]	0.83	0.59
instancen. (IN) [4]	0.78	**0.80**
groupnorm [16]	0.72	0.54

(a) Experiment 1: testing original MAML [2] on models with different normalization layers.

metric: **accuracy**	task-datasets	
	Sen12MS	DFC2020
number of tasks	1000	7
task design	idealized	realistic
label distribution	balanced	imbalanced
MAML + IN [4]	0.78	**0.80**
FoMAML + IN [2]	0.66	0.77
SpMAML + IN [15]	0.74	0.79
SpFoMAML + IN [15]	0.63	0.74

(b) Experiment 2: varying the meta-learning algorithm with a model with instance normalization (IN) layers.

Table 1: In experiment 1, we find that the commonly used transductive batch normalization fails on realistically class-imbalanced DFC2020 tasks (indicated in red). The proposed instance normalization achieves best results in these tasks also when we vary the training algorithm in experiment 2.

to Sen12MS tasks. Also, replacing batch normalization with the proposed tasknorm-I [1] layers leads to similar results of 83% on Sen12MS and 59% on DFC2020. Crucially, we found that completely replacing batch normalization with instance normalization (*instancen. (IN) hereafter*) lead to best results on the DFC2020 dataset of 80% by a large margin of 20% while being noticeably worse on class-balanced Sen12MS dataset with 78%. Overall, we select instance normalization as best normalization strategy as we are interested in the more realistic DFC2020 setting. For completeness, we also tested models without normalization which did not converge to a meaningful solution with a ResNet model (not shown in the table) and group normalization (*groupnorm*, last row) which was worse to instance normalization in both Sen12MS and DFC2020 tasks.

In the second Experiment 2, we test how well instance normalization translates to other model-agnostic meta-learning variants. To do so, we train a model instance normalization layers with the first-order approximation of MAML (Fo-MAML) and also a recently proposed SparseMAML [15] implementation. In SparseMAML, the model learns during meta-training to selectively fine-tune only a subset of its weights (defined by a sparse mask) to the respective target tasks. We find that the original second-order MAML [2] algorithm still worked best in this setting further supporting the applicability of instance normalization layers on class-imbalanced datasets.

5 Conclusions

In this work, we empirically demonstrated that replacing all batch normalization layers with instance norm [14] significantly outperforms all other (including tasknorm [1] and the recently proposed SparseMAML [15]) meta-learning algorithms in a realistic class-imbalanced setting for land cover classification. In parallel, instance normalization performs sub-optimally in a class-balanced setting, as shown in the Sen12MS column where instance norm (78%) falls behind batch normalization (85% and 84%) and task norm (83%). We believe, these results are roughly consistent with common benchmark datasets [6, 9] where instance normalization is not considered competitive in a class-balanced setting. This may also explain why, to our knowledge, no work has discussed and investigated the role of instance normalization in meta-learning for class-imbalanced test tasks thorougly so far. The limitations of transductive batch normalization have been pointed out by Bronskill et al., (2020) [1] who suggested comparing meta-learning algorithms separately in a transductive and non-transductive setting. However, transductive batch normalizaton still achieves the best numerical results on these benchmarks, which we find misleading given its sensitivity to the distribution over the target set.

Model-agnostic meta-learning is conceptually attractive for many downstream applications where a model can benefit from knowledge of different-but-related sources tasks with few annotated samples per task. Still, we find that the adoption of meta-learning is relatively slow in many fields of application. We believe that this discrepancy between (apparent) good performance on class-balanced benchmark datasets and then collapse in realistic application is a major factor in the perceived brittleness of model-agnostic meta-learning. We hope that the robustness showed in these result will foster new research in meta-learning for real life applications in several fields of science.

References

[1] John Bronskill, Jonathan Gordon, James Requeima, Sebastian Nowozin, and Richard Turner. Tasknorm: Rethinking batch normalization for meta-learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1153–1164. PMLR, 2020.

[2] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1126–1135. PMLR, 2017.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 770–778, 2016.

[4] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017.

[5] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 448–456. PMLR, 2015.

[6] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[7] Thomas Loveland and Alan Belward. The international geosphere biosphere programme data and information system global land cover data set (discover). *Acta Astronautica*, 41(4):681–689, 1997. Developing Business.

[8] Jaehoon Oh, Hyungjun Yoo, ChangHwan Kim, and Se-Young Yun. Boil: Towards representation change for few-shot learning. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[9] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[10] Michael Schmitt, Lloyd Hughes, Pedram Ghamisi, Naoto Yokoya, and Ronny Hänsch. IEEE GRSS Data Fusion Contest, 2020.

[11] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. Sen12ms – a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume IV-2/W7, pages 153–160, 2019.

[12] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*, 2019.

[13] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Recent advances in domain adaptation for the classification of remote sensing data. *IEEE Geosci. Remote Sens. Mag.*, 4(2):41–57, 2016.

[14] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[15] Johannes Von Oswald, Dominic Zhao, Seijin Kobayashi, Simon Schug, Massimo Caccia, Nicolas Zucchet, and João Sacramento. Learning where to learn: Gradient sparsity in meta and continual learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

[16] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.