Introducing the NeurIPS'22 Cross-Domain MetaDL competition

Dustin Carrión-Ojeda [1] **Hong Chen** [2] **Adrian El Baz** [3 4] **Sergio Escalera** [5 6] **Chaoyu Guan** [2] **Isabelle Guyon** [1 5] **Ihsan Ullah** [1] **Xin Wang** [2] **Wenwu Zhu** [2]

Abstract

We introduce a new challenge in the ChaLearn meta-learning series, which has a **special league for NewInML participants (with prizes and certificates)**. The competition is part of the NeurIPS'22 program and the winners will be invited to co-author the analysis paper with the organizers to appear in PMLR. The focus is on "cross-domain" meta-learning, aiming at leveraging experience from previous tasks to solve new tasks efficiently (*i.e.,* with better performance, little training data and/or modest computational resources). While our previous challenge addressed *within-domain* few-shot learning for *N-way k-shot* tasks (*i.e.,* N class classification problems with k training examples), this challenge proposes "any-way" and "any-shot" tasks drawn from various domains (healthcare, ecology, biology, manufacturing, and others), chosen for their humanitarian and societal impact. Code submissions will be blind-tested on CodaLab. The code of the winners will be open-sourced.

1. Introduction

Challenges in machine learning have been instrumental in pushing the state-of-the-art and stimulating participants to tackle new difficult problems. Since 2015, ChaLearn has been organizing challenges in Automated Machine Learning (AutoML) (Guyon et al., 2019) and Automated Deep Learning (AutoDL) (Liu et al., 2021), with the aim of reducing the need of human intervention in the design and implementation of machine learning models, to the greatest possible extent. The results of these challenges motivated us to organize a new ChaLearn challenge series in meta-learning, focusing first on image classification and few-shot learning. This challenge, the **NeurIPS'22 Cross-Domain MetaDL**, is the third edition in the series. **Submissions are open between July 01 and August 31, 2022.** The results will be presented at the NeurIPS'22 conference.

This summary introduces our new challenge and presents baseline results, with the aim of encouraging participation of NewInML community members. Our new design will challenge participants to **generalize across domains** in **different regimes in numbers of ways and shots**, and **compare "de novo" training with the use of pre-trained backbones**. Furthermore, the **code of the winners will be open-sourced** and enable **practical AutoML applications** since the meta-trained learner will be readily usable for few-shot image classification in the 10 domains of the challenge. **We offer extensive tutorial material to lower the barrier of entry. This should be stimulating to newcomers.**

2. Competition design

2.1. Problem setting

The few-shot learning problems are often referred as N-way k-shot problems. In these problems, each task $\mathcal{T}_j = \{\mathcal{D}^{train}_{\mathcal{T}_j}, \mathcal{D}^{test}_{\mathcal{T}_j}\}$ consists of a small training set $\mathcal{D}^{train}_{\mathcal{T}_j}$ and a small test set $\mathcal{D}^{test}_{\mathcal{T}_j}$, referred to as *support* and *query* sets, respectively. The number of ways N denotes the number of classes in a task that represents an image classification problem, the same N classes are present in $\mathcal{D}^{train}_{\mathcal{T}_j}$ and $\mathcal{D}^{test}_{\mathcal{T}_j}$. The number of shots k denotes the number of examples per class in the *support set*. In this challenge, the tasks at meta-test time have a number of classes varying from 2 to 20 ($N \in [2, 20]$), the support set contains 1 to 20 labeled examples per class ($k \in [1, 20]$), and the query set contains 20 unlabeled examples per class, *i.e.*, $|\mathcal{D}^{train}_{\mathcal{T}_j}| = N \times k$, and $|\mathcal{D}^{test}_{\mathcal{T}_j}| = N \times 20$. Moreover, since for this competition the tasks come from the **cross-domain** scenario, the data contained in one task $\mathcal{T}_j$, belongs strictly to one dataset, but different tasks may come from different datasets because the meta-dataset used to carved out the tasks is composed of

[1]LISN/CNRS/INRIA, Université Paris-Saclay, France [2]Department of Computer Science and Technology, Tsinghua University, China [3]MILA - Québec AI Institute, Montréal, QC, Canada [4]NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montréal, Montréal, QC, Canada [5]ChaLearn, USA [6]Computer Vision Center, Universitat de Barcelona, Spain . Correspondence to: Dustin Carrión-Ojeda <dustin.carrion@gmail.com>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

multiple datasets, *i.e.*, $\mathcal{M}_{\mathcal{D}} = \{\mathcal{D}_1, \dots, \mathcal{D}_n\}$. The number of datasets n in the meta-dataset $\mathcal{M}_{\mathcal{D}}$ depends on the phase (see Section 2.3).

2.2. Data

The datasets of this competition belong to the Meta-Album meta-dataset, prepared in conjunction with this competition (Ullah et al., 2022). It consists of 40 re-purposed and novel image datasets from 10 domains: small and large animals, plants and plant diseases, vehicles, human actions, microscopic data, satellite images, industrial textures, and printed characters. For this competition, we selected 30 datasets from the meta-dataset and partitioned them into 3 sets of 10 datasets, one from each domain, used in the various competition phases (Set-0, Set-1, and Set-2). All final test phase datasets are novel to the meta-learning community (not part of past meta-learning benchmarks). Sets 0-2 will be released on OpenML (Vanschoren et al., 2013) after the competition ends.

2.3. Competition protocol

This is an online competition with code submission, *i.e.*, the participants need to provide their solutions as raw Python code that will be executed on our dedicated CodaLab site[1]. To guarantee fairness in the evaluation of the participants, the CodaLab server used in this challenge is equipped with 10 identical computer workers. Each has the following configuration: 4 CPU cores, 1 Tesla T4 GPU, 16GB RAM, 120GB storage.

The competition follows the problem setting described in Section 2.1 and it is composed of 3 phases. During the **Public phase** (Jun 15 - 30, 2022) no submissions can be done; instead, the participants can use the tutorial provided as part of the starting kit (see Section 2.4), and Set-0 to test their solutions before submitting them. Then, during the **Feedback phase** (Jul 01 - Aug 31, 2022), participants can make 2 submissions per day and a maximum of 100 submissions during the whole phase. Each submission is evaluated on 1000 any-way any-shot tasks carved out from Set-1 (100 tasks per dataset), and cannot take more than 5 hours of running time. Lastly, during the **Final phase** (Sep 01 - 30, 2022) the last submission of each participant on the **Feedback phase**, whose performance is above the baseline performance (see Section 2.5), will be evaluated on 6000 any-way any-shot tasks carved out from Set-2 (600 tasks per dataset). Due to the increment of meta-test tasks, the allowed running time will increase to 9 hours.

The submissions must follow our defined API, which was designed to be flexible enough to allow participants to explore any type of meta-learning algorithms. To encourage a diversity of participants and types of submissions, this competition has 5 different leagues. All the details about the API, the leagues, the prizes, and rules can be found in the competition site[1]. Notably, there is a **special league for New-in-ML** for all participants with less than 10 ML publications, none of which were ever accepted to the main track of a major conference.

2.4. Starting kit and additional resources

The starting kit of this competition can be found in our CodaLab site[1]. Since this competition is open to anyone interested in meta-learning including novice people without previous knowledge of meta-learning, our starting kit contains a tutorial with three levels of complexity: beginner, intermediate, and advance. To follow the beginner level there are no prerequisites for the participants, therefore, anyone can follow it. Moreover, to facilitate the accessibility to our tutorial, we provide it in three equivalent formats, *i.e.,* all the formats contain the same information; thus, participants have to follow just **one**: Google Colab, GitHub repository, and zip file. We highly recommend the Google Colab version since with it, the participants do not have to download and/or install anything in their computers.

The tutorial will cover all the details regarding the cross-domain meta-learning problem, the data used in this challenge, the structure for the submissions, the evaluation process, and it provides suggestions for new submissions. By following the tutorial, the participants will be able to create their first submission in less than 10 minutes, and also they will have access to all the baselines described in Section 3. In addition to our starting kit, we highly recommend to check our white paper describing all the details of this competition and the baseline results (Carrión-Ojeda et al., 2022). Additional resources include the lessons learned from our previous NeurIPS 2021 MetaDL challenge (El Baz et al., 2022), and our Meta-Album preprint (Ullah et al., 2022).

2.5. Challenge metrics

Since the meta-test tasks have different configurations in number of ways and shots, this competition uses the balanced classification accuracy (bac) as the evaluation metric, normalized with respect to the number of ways. This metric is defined as follows:

$$\text{Normalized Accuracy} = \frac{bac - bac_{RG}}{1 - bac_{RG}}, \quad (1)$$

where bac also known as macro-averaging recall is defined as:

$$bac = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{correctly classified images of class } i}{\text{total images of class } i}, \quad (2)$$

[1]Competition site: https://codalab.lisn.upsaclay.fr/competitions/3627

and bac_{RG} is accuracy of random guessing, *i.e.*, $1/N$.

The error bars computed for this competition correspond to 95% confidence interval of the mean normalized accuracy at task level computed as follows:

$$CI = \pm t \times \frac{\sigma}{\sqrt{n}}, \tag{3}$$

where t is the *t-value* depending on confidence level and degrees of freedom ($df = n - 1$); σ corresponds to the standard deviation of the normalized accuracy obtained on all meta-test tasks, and n the number of such tasks. The baseline performance that must be surpassed by the participants in the Feedback phase to enter into the Final phase depends on the league. The baseline performance for the Free-style and Meta-learning leagues is 0.587 and 0.361, respectively. These baseline performance were calculated by averaging the normalized accuracy achieved by the best methods (see Section 3.2) in each league over 10 runs varying the random seed of the baseline methods.

3. Baseline results

This section provides the results of six baseline methods carefully selected aiming to have a variety of approaches in terms of training strategy (batch and episodic training), and also in terms of modeling choices (fine-tuning, metric-based, and ensemble). The first one, **Train-from-scratch**, does not perform any meta-training: it directly learns each meta-testing task using only its support set. The second one, **Fine-tuning**, is a simple transfer learning method consisting in pre-training a backbone network with batches of data from the concatenated meta-training datasets, then only fine-tuning the last layer at meta-test time. Three of the remaining baselines are popular meta-learning methods: **Matching Networks** (Vinyals et al., 2016), **Prototypical Networks** (Snell et al., 2017), and **FO-MAML** (Finn et al., 2017). And the last baseline is an adaptation of MetaDelta++ (Chen et al., 2021), which was the winner solution of the previous NeurIPS'21 challenge. All baseline methods but MetaDelta++ use a ResNet-18 backbone (He et al., 2016) with the best-reported hyperparameters by the original authors on 5-way 5-shot miniImageNet. Taking into account that the complete baselines results are available in our white paper describing all the details of this competition and the baseline results (Carrión-Ojeda et al., 2022).

3.1. Experimental setting

Data: We report results for Feedback phase data of the competition (see Section 2.3). Therefore, the 10 datasets of Set-0 were divided into 7 datasets for meta-training, and 3 datasets for meta-validation. This division was randomly made; hence, it was different in each run because of the variation in the random seed.

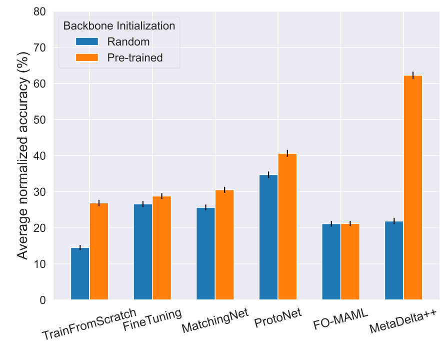

Figure 1: **Comparison of baselines methods using a randomly initialized backbone and a pre-trained backbone.** The barplot show the average normalized accuracy over 3,000 any-way any-shot meta-test tasks (100 tasks per dataset in each run). The corresponding 95% CIs are computed at task level.

Evaluation setting: The meta-learning methods were meta-trained on 30,000 5-ways 10-shots tasks, the Fine-tuning baseline was meta-trained on 30,000 batches of size 16, and the MetaDelta++ baseline was meta-trained during 3.5 hours with batches of size 64. The performance of the meta-trained Learners was validated after every 5,000 meta-training steps on 300 5-ways 5-shots meta-validation tasks except for MetaDelta++ in which case the Learner was validated after every 50 meta-training batches on 50 5-ways 5-shots meta-validation tasks. The query set for every task contained 20 examples per class except for the meta-validation tasks used by MetaDelta++ in which case 5 examples per class were used. The Learner with the best validation performance was evaluated following the protocol of the Feedback phase and using the computational resources described in Section 2.3.

3.2. Results

In Figure 1, we compare the baseline methods by averaging results over all tasks from all datasets. The network backbones are either initialized with random weights or pre-trained weights with ImageNet before meta-training. The figure shows that initializing the backbones with pre-trained weights significantly helps, indicating that perhaps our meta-training set is not large enough or that meta-training time is not sufficient. We hope to see improvements in the Meta-learning league of the challenge regarding using random initialization. Moreover, the winner of the previous challenge (MetaDelta++) performs significantly better than other baselines when using pre-trained backbone, but Prototypical Networks is the best option when no-pretraining is allowed.

4. Conclusion

The NeurIPS'22 Cross-Domain MetaDL competition has lower the barrier of entry by providing an easy-to-follow tutorial with all the information required by a newcomer. Morevoer, the experimental results show that if pre-trained backbones are allowed, MetaDelta++ is the best option among the baselines and, in general, all baselines (except for FO-MAML) benefit from using pre-trained initialization. However, if using pre-trained weights is not allowed, which is the case for some real world applications where no pre-trained backbone is available, Prototypical Networks is the best option within the evaluated methods.

Acknowledgements

We acknowledge support by ChaLearn, ANR AI chair HUMANIA ANR-19-CHIA-0022, TAILOR EU Horizon 2020 grant 952215, and the help of Mike Huisman for baseline method code (exceptMetaDelta++), and of Romain Mussard, Manh Hung Nguyen, and Gabriel Lauzzana as competition beta-testers.Experiments were performed using a Google cloud grant. The authors are in alphabetical order of last name, except the first author.

References

Carrión-Ojeda, D., Chen, H., El Baz, A., Escalera, S., Guan, C., Guyon, I., Ullah, I., Wang, X., and Zhu, W. NeurIPS'22 Cross-Domain MetaDL competition: Design and baseline results. In *Submitted to: Meta-Knowledge Transfer/Communication in Different Systems at ECML PKDD*, 2022.

Chen, Y., Guan, C., Wei, Z., Wang, X., and Zhu, W. Metadelta: A meta-learning system for few-shot image classification. *CoRR*, abs/2102.10744, 2021.

El Baz, A., Ullah, I., Alcobaça, E., Carvalho, A. C. P. L. F., Chen, H., Ferreira, F., Gouk, H., Guan, C., Guyon, I., Hospedales, T., Hu, S., Huisman, M., Hutter, F., Liu, Z., Mohr, F., Öztürk, E., van Rijn, J. N., Sun, H., Wang, X., and Zhu, W. Lessons learned from the NeurIPS 2021 MetaDL challenge: Backbone fine-tuning without episodic meta-learning dominates for few-shot learning image classification. In *Proceedings of the NeurIPS 2021 Competition and Demonstration Track*, volume 176 of *Proceedings of Machine Learning Research*, pp. 80–96. PMLR, 2022.

Finn, C., Abbeel, P., and Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *34th International Conference on Machine Learning*, pp. 1126––1135, 2017.

Guyon, I., Sun-Hosoya, L., Boullé, M., Escalante, H. J., Escalera, S., Liu, Z., Jajetic, D., Ray, B., Saeed, M., Sebag, M., Statnikov, A. R., Tu, W., and Viegas, E. Analysis of the automl challenge series 2015-2018. In *Automated Machine Learning - Methods, Systems, Challenges*, The Springer Series on Challenges in Machine Learning, pp. 177–219. Springer, 2019. doi: 10.1007/978-3-030-05318-5_10. URL https://doi.org/10.1007/978-3-030-05318-5_10.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Liu, Z., Pavao, A., Xu, Z., Escalera, S., Ferreira, F., Guyon, I., Hong, S., Hutter, F., Ji, R., Junior, J. C. S. J., Li, G., Lindauer, M., Luo, Z., Madadi, M., Nierhoff, T., Niu, K., Pan, C., Stoll, D., Treguer, S., Wang, J., Wang, P., Wu, C., Xiong, Y., Zela, A., and Zhang, Y. Winning solutions and post-challenge analyses of the chalearn autodl challenge 2019. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9):3108–3125, 2021. doi: 10.1109/TPAMI.2021.3075372.

Snell, J., Swersky, K., and Zemel, R. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1–13, 2017.

Ullah, I., Carrion, D., Escalera, S., Guyon, I., Huisman, M., Mohr, F., van Rijn, J. N., Sun, H., Vanschoren, J., and Vu, P. A. Meta-album: Multi-domain meta-dataset for few-shot image classification. In *Submitted to: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2022. URL https://meta-album.github.io/.

Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. Openml: networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198. URL http://doi.acm.org/10.1145/2641190.264119.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29:3630–3638, 2016.