NeurIPS'22 Cross-Domain MetaDL competition: Design and baseline results*

Dustin Carrión-Ojeda[1] DUSTIN.CARRION@GMAIL.COM
Hong Chen[2] H-CHEN20@MAILS.TSINGHUA.EDU.CN
Adrian El Baz[3,4] EB.ADRIAN@HOTMAIL.FR
Sergio Escalera[5,6] SERGIO.ESCALERA.GUERRERO@GMAIL.COM
Chaoyu Guan[2] GUANCY19@MAILS.TSINGHUA.EDU.CN
Isabelle Guyon[1,5] GUYON@CHALEARN.ORG
Ihsan Ullah[1] IHSAN2131@GMAIL.COM
Xin Wang[2] XIN_WANG@TSINGHUA.EDU.CN
Wenwu Zhu[2] WWZHU@TSINGHUA.EDU.CN

[1] LISN/CNRS/INRIA, Université Paris-Saclay, France
[2] Department of Computer Science and Technology, Tsinghua University, China
[3] MILA - Québec AI Institute, Montréal, Canada
[4] NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montréal, Canada
[5] ChaLearn, USA
[6] Computer Vision Center, Universitat de Barcelona, Spain

Editors: P. Brazdil, J. N. van Rijn, H. Gouk and F. Mohr

Abstract

We present the design and baseline results for a new challenge in the ChaLearn meta-learning series, accepted at NeurIPS'22, focusing on "cross-domain" meta-learning. Meta-learning aims to leverage experience gained from previous tasks to solve new tasks efficiently (*i.e.,* with better performance, little training data, and/or modest computational resources). While previous challenges in the series focused on *within-domain* few-shot learning problems, with the aim of learning efficiently *N-way k-shot* tasks (*i.e., N* class classification problems with k training examples), this competition challenges the participants to solve "any-way" and "any-shot" problems drawn from various domains (healthcare, ecology, biology, manufacturing, and others), chosen for their humanitarian and societal impact. To that end, we created Meta-Album, a meta-dataset of 40 image classification datasets from 10 domains, from which we carve out tasks with any number of "ways" (within the range 2-20) and any number of "shots" (within the range 1-20). The competition is with code submission, fully blind-tested on the CodaLab challenge platform. The code of the winners will be open-sourced, enabling the deployment of automated machine learning solutions for few-shot image classification across several domains.

Keywords: Image Classification, AutoML, Few-Shot Learning, Meta-Learning, Cross-Domain Meta-Learning.

* The authors are in alphabetical order of last name, except the first author.

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., 2021b), with the aim of reducing the need of human intervention in the design and implementation of machine learning models, to the greatest possible extent. Our challenge series gave rise to the popular auto-sklearn software and outlined the importance of good representations (obtained from pre-trained backbone networks), data augmentation, and meta-learning. These results prompted 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 1 and August 31, 2022.** The results will be presented at the NeurIPS'22 conference.

Traditionally, image classification has been tackled using deep learning methods whose performance relies on the availability of large amounts of data (Phoo and Hariharan, 2021). Recent efforts in meta-learning (Jamal and Qi, 2019) have contributed to making a lot of progress in few-shot learning for image classification problems. Tasks or "episodes" are made of a certain number of classes or "ways" and number of labeled examples per class or "shots". Despite progress made, allowing the community to reach accuracies above 90% in the last ChaLearn meta-learning challenge (El Baz et al., 2022), evaluation protocols have a common drawback: they focus only on within-domain few-shot learning, *i.e.,* even when evaluated on multiple domains (*e.g.,* insect classification, texture classification, satellite images, etc.), models meta-trained on a given domain are meta-tested on the same domain. As documented in the literature, within-domain few-shot learning approaches have poor generalization ability to unrelated domains (Phoo and Hariharan, 2021). Nevertheless, this kind of generalization is crucial since there are scenarios where only one or two examples per class are available (*e.g.,* rare birds or plants), and there is no close domain with enough information to be used as source domain. Therefore, addressing domain variations has become a research area of great interest. Additionally, in most works about few-shot learning, the number of ways and shots is fixed, which is not always the case in real application scenarios.

Currently, the most popular benchmark used for cross-domain meta-learning is Meta-Dataset (Triantafillou et al., 2020). This benchmark tackles the problems mentioned above by including 10 image classification datasets from several application domains in one collection and analyzing the impact of using a variable number of ways and shots. Although it has been widely used to evaluate state-of-the-art methods (Dvornik et al., 2020; Liu et al., 2021a; Triantafillou et al., 2021; Li et al., 2021, 2022), it cannot be used in our competition because it is already well-known by the meta-learning community. Additionally, the datasets in Meta-Dataset have a large variance in the number of classes and examples per class, introducing bias in our competition design.

The main contributions of this paper are the design of a new challenge in the ChaLearn meta-learning series and the presentation of baseline results. 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**.

Note that "de novo" training means that the algorithm needs to learn from scratch without using any previous knowledge such as pre-trained backbones, *i.e.,* the backbones must be initialized randomly. In conjunction with the organization of this challenge, we developed a large meta-dataset called Meta-Album described in a companion paper (Ullah et al., 2022), including 40 datasets belonging to 10 different domains, relevant to "AI for good", such as ecology, medicine, and biology, with the intent of maximizing the economic and societal impact of the challenge. In this competition, 30 of these datasets will be used for meta-training and meta-testing, then released publicly as a **long-lasting benchmark** to further push the state-of-the-art. A single (final) submission will be evaluated during the final challenge phase, using ten datasets previously unused by the meta-learning community. 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.

2. Problem setting

This challenge has two motivational scenarios: (1) Few-shot image classification and (2) Meta-learning from limited amounts of meta-learning data. For the former problem, we target users wishing to create an image classifier from a few pictures of each class (*e.g.,* taken with a smartphone) in a **new domain** (*e.g.,* classify clouds). The challenge winning solution(s) should make this possible for **any-way any-shot** in the range [2-20] "ways" (classes) and [1-20] "shots" (training examples per class). For the latter problem, the solution of the winner should deliver a meta-learning algorithm leveraging knowledge from previous tasks, without relying on pre-trained backbones, applicable to a wider range of applications than image classification (encouraged by the prize distribution, see Appendix D). This section first explains the setting of the previous MetaDL challenge organized for NeurIPS'21 (within-domain few-shot learning). Then, it explains the new variant we developed for NeurIPS'22 (cross-domain any-way any-shot learning). Both competitions are with code submission and the participants must supply code following a designated API (Liu et al., 2019), featuring Python objects (see Appendix C): `MetaLearner`, `Learner`, and `Predictor`. `MetaLearner` uses meta-training data (a dataset of datasets) to create `Learner`; `Learner` in turn uses training examples (images) to return `Predictor`; finally, `Predictor` uses unlabeled test examples to return predicted class labels. The competitions are composed of 2 main phases, a **feedback phase** with many submissions allowed and immediate feedback provided on a leaderboard, and a **final test phase** with only 1 submission tested on new datasets. In both phases, data are not visible to the participants; only the code submitted has access to evaluation data. Ground truth labels of test data are kept secret and are only visible to the scoring program.

2.1. Within-domain few-shot learning

The NeurIPS'21 MetaDL competition focused on "within-domain" few-shot learning image classification in the N-way k-shot setting (El Baz et al., 2021). The `Learner` was meta-tested on many 5-way 5-shot tasks carved out from several multi-class image datasets, each task including $N = 5$ classes drawn at random, with $k = 5$ examples per class in the support (training) set and 20 examples per class in the query (test) set. Half of the classes of

each dataset were reserved for meta-training and the other half for meta-testing. During meta-training, the `MetaLearner` could choose the configuration of data received: (1) batch training with examples of all classes within the domain at hand, (2) episodic training with examples grouped in tasks having a support and a query set. Importantly, meta-learning and meta-testing were performed using classes from the **same dataset**, which we refer to as **within domain meta-learning**. Submissions made by participants were then ranked per dataset, and the final ranking was obtained by averaging such ranks. Five datasets from 5 domains (ecology, bio-medicine, manufacturing, optical character recognition, and remote sensing) were used in the feedback phase, and 5 other fresh datasets from the same domains were used in the final evaluation phase. All datasets had at least 20 classes and 40 images per class (El Baz et al., 2022).

2.1.1. Lessons learned and limitations

The NeurIPS'21 MetaDL competition considered a refined competition protocol developed for a previous MetaDL competition (El Baz et al., 2021), introducing multiple domains, which added additional sophistication in terms of scoring as well as GPU-time budgeting. However, the setting remained relatively simplified since the meta-training and meta-testing were performed within-domain (*i.e.,* non-overlapping classes of the same dataset were used for meta-training and meta-testing) using meta-test tasks with a fixed number of ways and shots. The winners (Chen et al., 2021) obtained over 92% accuracy on all 5 domains in the final phase (with complete blind-testing of their code). Thus, this indicates that we can move to more complicated problems. Following these observations, the Cross-Domain MetaDL challenge intends to mix tasks from multiple domains and present variable numbers of *ways* and *shots*.

Although the NeurIPS'21 MetaDL competition did not constrain participants to use deep-learning, de facto, all participants based their solutions on deep-learning models with convolutions (specifically, either convolutional neural networks or transformer models). Additionally, fine-tuning on meta-training data turned out to be important. However, there are indications that off-the-shelf backbones pre-trained with self-supervised learning on massive datasets might be the most promising approach, essentially making meta-learning unnecessary for image classification problems. Thus, meta-learning should be benchmarked in de novo training conditions to prepare for scenarios (in other domains) in which such backbones are not available.

As reported by several top-ranking teams, meta-learning was possible within domains (in the form of fine-tuning pre-trained backbones), but MAML-style episodic meta-learning did not turn out to be more effective than vanilla pre-training with gradient descent. Based on the embedding generated by the backbones, prototypical classifiers seem more efficient than linear classifiers. Hence, the Cross-Domain MetaDL challenge also allows further probing of the effectiveness of various meta-learning solutions.

2.2. From within-domain to cross-domain any-way any-shot learning

Following the lessons learned from the NeurIPS'21 competition, the new Cross-Domain MetaDL challenge aims to push the complete automation of few-shot learning by demanding

participants to design learning agents capable of producing a trained classifier in the **cross-domain any-way any-shot setting**.

As introduced in Section 2.1, the few-shot learning problems are often referred as N-way k-shots 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 in this competition, the tasks come from the **cross-domain** scenario, the data contained in one task $\mathcal{T}_j$ belongs strictly to one dataset. Nonetheless, different tasks may come from different datasets because the meta-dataset used to carved out the tasks is composed of multiple datasets, *i.e.,* $\mathcal{M}_{\mathcal{D}} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$. The number of datasets n in the meta-dataset $\mathcal{M}_{\mathcal{D}}$ depends on the phase (see Section 3.1).

The proposed setting consists of three stages: meta-training, meta-validation (optional), and meta-testing, which are used for meta-learning, model selection, and evaluation, respectively. During the meta-training stage, the participants can choose to use data in the form of *tasks* $\mathcal{T}_j$ or *batches* which are a collection of sampled examples from a single large dataset resulting of concatenating all datasets of the meta-training dataset, *i.e.,* $\mathcal{D}^{train} = concat(\mathcal{D}_1, \ldots, \mathcal{D}_n)$. Additionally, they can specify their preferred configurations for the selected data format at this stage. The meta-validation stage is optional; therefore, it is up to the participants to use it, but the data for this stage is always in the form of tasks. Nevertheless, the participants can still specify their preferred configurations for the meta-validation tasks. Lastly, during the meta-testing stage, the participants have no control over the data, which always arrives in the form of *any-way any-shot tasks* with $N \in [2, 20]$ and $k \in [1, 20]$. During meta-testing, the labels of the query set are hidden from the participants' codes.

3. Competition design

3.1. 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 or 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. We preprocessed data in a standard format suitable for few-shot learning. The preprocessing pipeline includes image resizing with anti-aliasing filters into a uniform shape of 128x128x3 pixels. 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., 2014) after the competition ends.

3.2. Competition protocol

NeurIPS'22 Cross-Domain MetaDL 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]. Detailed competition rules are found in Appendix A. 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, and 120GB storage.

The competition follows the problem setting described in Section 2.2. It is composed of 3 phases. During the **Public phase** (June 15-30, 2022), no submissions can be made; instead, the participants can use the tutorial provided as part of the starting kit (see Appendix B) and Set-0 to test their solutions on their computers or Google Colab. Then, during the **Feedback phase** (July 1 - August 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) different from the ones used for meta-training (Public data). Additionally, each submission cannot take more than 5 hours of running time. Lastly, during the **Final phase** (September 1-30, 2022), the last submission of each participant on the **Feedback phase**, whose performance is above the baseline performance (see Section 3.3), 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 (see Appendix C), 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, the Cross-Domain MetaDL competition has 5 different leagues. Appendix D details the leagues and prizes. Notably, there is a league to **encourage meta-training from scratch ("de novo" training)** as opposed to using pre-trained backbones.

3.3. Challenge metrics

Since the meta-test tasks have different configurations in the 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 (which is the number of classes in the task). This metric is defined as follows:

$$\text{Normalized Accuracy} = \frac{bac - bac_{RG}}{1 - bac_{RG}}, \tag{1}$$

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

$$bac = \frac{1}{num\_ways} \sum_{i=1}^{num\_ways} \frac{\text{correctly classified examples of class } i}{\text{total examples of class } i}, \tag{2}$$

and bac_{RG} is the accuracy of random guessing, *i.e.,* $\frac{1}{num\_ways}$. Note that by (1), a normalized accuracy of 0 means that the performance of the submission is equivalent to random guessing.

1. CodaLab site for the Cross-Domain MetaDL competition: https://codalab.lisn.upsaclay.fr/competitions/3627

Moreover, the normalized accuracy can be negative, indicating that the submission is worse than random guessing, and the maximum achievable normalized accuracy is 1.

The error bars correspond to 95% confidence intervals 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 is the number of such tasks.

In this competition, CI calculations are only indicative and not used to select winners or declare ties. The baseline performance the participants in the Feedback phase must surpass 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 (see Appendix D for the definition of leagues). These baseline performances were calculated by averaging the normalized accuracy achieved by the best methods (see Section 4.3) in each league over 10 runs varying the random seed of the baseline methods.

To select the winners in the Final phase, all eligible entries are run three times, with various random seeds. The average normalized classification accuracy over all meta-test tasks is computed in each run, and the lowest of the three runs is used for the final ranking. Ties are broken according to the first submission made. Note that the baseline performances quoted in Section 4.3 are obtained by averaging the performance over multiple runs to reduce variance, while the final evaluation of participants is made based on the worst performance over three runs.

4. Baseline results

In this section, we present experiments to evaluate the difficulty of the new challenge setting. We run several baseline methods to evaluate whether: (1) "Cross-domain meta-learning any-way any-shot" (new setting) is significantly more complicated than "Within domain 5-way 5-shot" (old setting); (2) Baseline methods perform significantly better when using a backbone pre-trained on ImageNet rather than meta-training (or training) "from scratch"; (3) the choice of datasets is appropriate to separate method performances.

4.1. Baseline methods

This competition provides six baseline methods as part of its "starting kit". The first one, **Train-from-scratch**, does not perform any meta-training; instead, it directly learns each meta-testing task using only its support set. The second one, **Fine-tuning**, is a simple transfer learning method consisting of pre-training a backbone network with batches of data from the concatenated meta-training datasets and 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). Furthermore, the last baseline is an adaptation of MetaDelta++ (Chen et al., 2021), which corresponds to the solution of the winners of the NeurIPS'21 challenge. All the baseline methods were carefully selected, aiming to have

a variety of approaches in terms of training strategy (batch and episodic training) and modeling choices (fine-tuning, metric-based, and ensemble). A detailed description of each method is presented in Appendix E. 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 (see Appendix E). For all baselines, the backbone can be either initialized with random weights or weights pre-trained on ImageNet (as in Meta-learning league and Free-style leagues, respectively, see Appendix D for league definition).

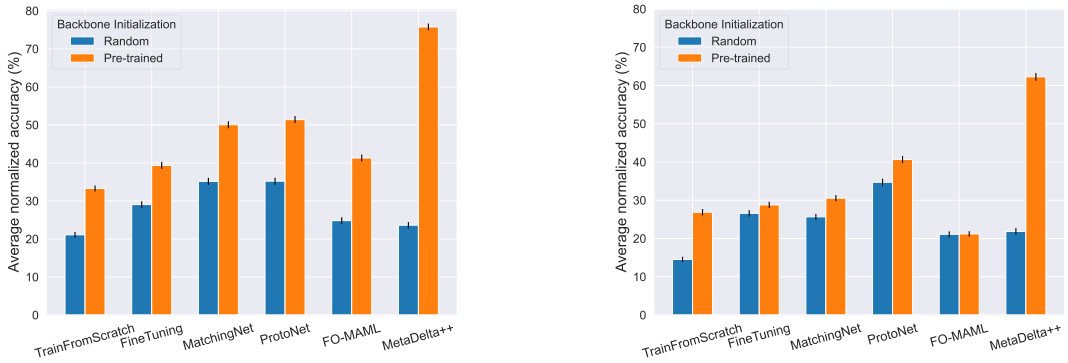
4.2. Experimental setting

Our experiments aim to compare and contrast the protocol of the NeurIPS'21 challenge (within-domain MetaDL) with that of the NeurIPS'22 challenge (cross-domain MetaDL).

Data: We report results for Feedback phase data of the Cross-Domain MetaDL challenge. Accordingly (see Section 3.2), the meta-training and meta-testing datasets correspond to Meta-Album Set-0 and Set-1, respectively. The 10 datasets of Set-0 were divided into 7 for meta-training and 3 for meta-validation. This division was randomly made; hence, it was different in each run because of the random seed variation. For the within-domain protocol, only Set-1 was used. In this case, each dataset of Set-1 was divided into meta-training, meta-validation, and meta-testing sets with non-overlapping classes using 70%, 15%, and 15% of the available classes, respectively.

Cross-Domain setting (NeurIPS'22). The meta-learning methods were meta-trained on 30,000 5-way 10-shot 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 Learners produced during the meta-training phase was validated after every 5,000 meta-training tasks (or batches in the case of the Fine-tuning method) on 300 5-way 5-shot tasks drawn from the meta-validation split except for MetaDelta++, in which case, the Learner was validated after every 50 meta-training batches on 50 5-way 5-shot tasks drawn from the meta-validation split. The query set for every task contained 20 examples per class except for the meta-validation tasks used by MetaDelta++, which contained 5 examples per class. The Learner with the best validation performance was evaluated following the protocol of the Feedback phase described in Section 3.2.

Within-Domain setting (NeurIPS'21). We evaluated the same baseline methods in the same way as for the Cross-Domain setting but with the protocol of the NeurIPS'21 MetaDL competition. However, since in the cross-domain setting the meta-training and meta-validation sets were composed of 7 and 3 datasets, respectively, and in the within-domain setting, 1 dataset is divided into meta-training, meta-validation, and meta-testing; we adapt the number of meta-training and meta-validation iterations to have comparable results between these two protocols. Thus, the meta-learning methods were meta-trained on 4,290 5-way 10-shot tasks, the Fine-tuning baseline was pre-trained on 4,290 batches of size 16, and the MetaDelta++ baseline was pre-trained for 30 minutes with batches of size 64. The performance of the Learners produced during the meta-training phase was validated after every 750 meta-training tasks (or batches in the case of the Fine-tuning method) on 100 5-way 5-shot tasks drawn from the meta-validation split except for MetaDelta++, in which case the Learner was validated after every 50 meta-training batches on 50 5-way 5-shot tasks



(a) Within-Domain (b) Cross-Domain

Figure 1: **Comparison of "within-domain" and "cross-domain" few-shot learning using a randomly initialized backbone and a pre-trained backbone.** Both barplots show the average normalized accuracy over 3,000 meta-test tasks (100 tasks per dataset in each run). In addition, the meta-test task configuration on the left barplot is 5-way 5-shot while on the right is any-way any-shot. The corresponding 95% CIs are computed at task level.

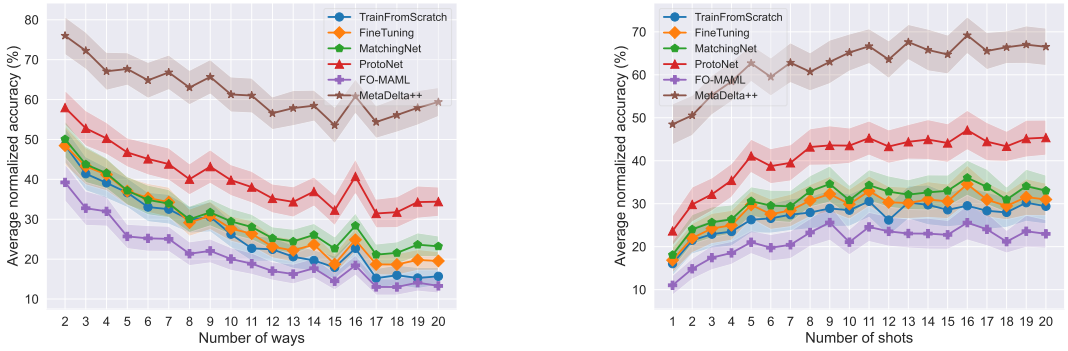
drawn from the meta-validation split. To resemble the NeurIPS'21, during meta-testing, the configuration for the tasks was 5-way 5-shot.

Computational resources. All experiments are carried out with the same resources as used in the Cross-Domain MetaDL competition (see Section 3.2).

4.3. Results

We aggregated results in various manners to compare the settings of the NeurIPS'21 and the NeurIPS'22 challenges with respect to (1) within-domain *vs.* cross-domain and (2) pre-trained *vs.* randomly initialized backbone.

Method comparison. In Figure 1, we compare baseline methods by averaging results over all tasks from all datasets. Before meta-training, the backbone networks are initialized with random or pre-trained weights on ImageNet. The figure shows that initializing the backbones with pre-trained weights helps significantly, indicating that perhaps our meta-training set is not large enough or that the meta-training time is insufficient. 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 a pre-trained backbone. However, Prototypical Networks is the best option when no pre-training is allowed. Additionally, we see that the new cross-domain setting is more complicated than the within-domain setting. In Figure 2, we study the influence of the number of ways and shots on the method performance in the new "cross-domain" setting. To that end, we averaged results over tasks with the same configuration (number of ways and shots) from all datasets and plotted the



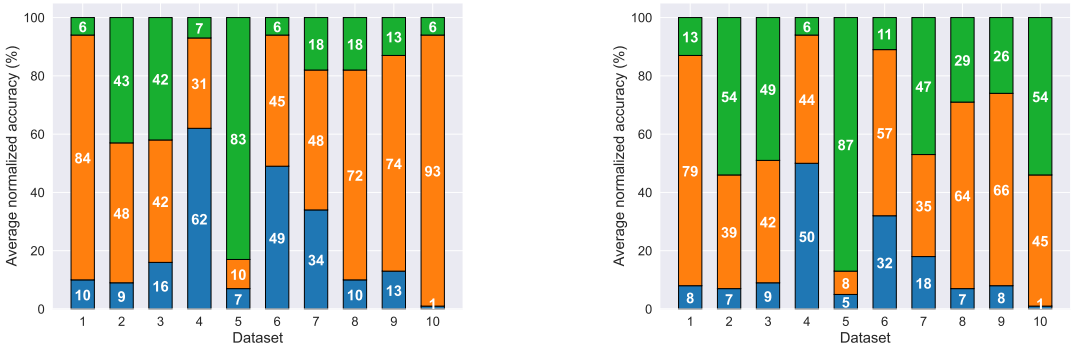
(a) Influence of the number of ways

(b) Influence of the number of shots

Figure 2: **Comparison of the influence of the number of ways and shots on the performance in the "cross-domain" setting using a pre-trained backbone.** We plot the average normalized accuracy achieved by the baselines using pre-trained weights. The corresponding 95% CIs are computed at task level.

normalized accuracy. Notably, the curves do not cross, indicating that **the ranking of methods is not influenced by the number of ways and shots**. We show only results using pre-trained backbone networks because the curves obtained with randomly initialized weights are qualitatively similar (only worse, and ordered differently, as in Figure 1). As expected, performances degrade with the number of ways and increase with the number of shots. Interestingly, the most significant increment occurs up to 5 shots. Appendix F contains the detailed results for all figures presented in this section.

Dataset comparison. In Figure 3, we averaged performances per dataset and reported only the results of the worst baseline (Train-from-scratch without pre-training) and the best baseline (MetaDelta++ with pre-training, previous challenge winners). These performances allow us to evaluate the intrinsic difficulty of the datasets (difference between the maximum achievable performance and the performance of the best baseline method – green bar) and the modeling difficulty (difference between the best and worst baseline methods – orange bar). As can be seen, the datasets show a range of difficulty, from dataset 4, which seems relatively easy, even to the worst baseline, to dataset 5, which is challenging even for the best baseline. Most datasets (except 5) have a reasonably large orange bar, indicating that the performance of methods spread over an extensive range, which is desirable in a challenge to separate methods. Dataset 10 is an interesting case: the best method performed well in the within-domain setting, but its performance dropped significantly in the cross-domain setting. We find that this domain does not resemble others; hence this is not so surprising that meta-learning within the domain should be more favorable. Generally, performances drop when we move to the new cross-domain setting; thus, the participants of the new challenge have some margin for improvement.



(a) Within-Domain (b) Cross-Domain

Figure 3: **Difficulty comparison of feed-back data** for "within-domain" and "cross-domain" few-shot learning, with a randomly initialized or pre-trained backbone. The top of the blue bar indicates the worst baseline performance (Train-from-scratch without pre-training). The top of the orange bar indicates the best baseline performance (MetaDelta++ with pre-training). The top of the green bar indicates the maximum achievable performance. The larger the green bar, the larger the *intrinsic difficulty*. The larger the orange bar, the larger the *modeling difficulty*. The average normalized accuracy was computed over 300 meta-test tasks (100 tasks per dataset in each run). Left: 5-way 5-shot; Right: any-way any-shot.

5. Conclusion and further work

We evaluated several baselines covering a variety of approaches to tackle few-shot learning problems to compare the protocols of NeurIPS'21 and NeurIPS'22 challenge settings. The experimental results show that the new proposed any-way any-shot cross-domain setting is more challenging than the previously studied 5-way 5-shot within-domain setting. This increment in problem complexity will allow us to encourage the participants to aim at finding methods capable of learning from multiple domains and generalize to all those domains in a more realistic test environment. Additionally, our findings show that if pre-trained backbones are allowed, MetaDelta++ is the best option among the baselines. 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. Moreover, our experiments allowed us to estimate the difficulty level of each dataset used in the Feedback phase of the new Cross-Domain MetaDL competition. The observed modeling difficulty is a good motivation for this competition since there is room for improvement, which is the expected outcome of this challenge. Finally, these results show that, due to the differences among domains, the difficulty of some datasets increases significantly in the new setting compared to the previous one.

While this competition studies cross-domain meta-generalization across 10 domains, it does not challenge participants to meta-generalize out of these domains since meta-test data includes new datasets from these exact 10 domains. We plan to organize a "domain independent" sequel, in which datasets from new domains not seen during meta-training will be used for meta-testing.

Acknowledgments

We acknowledge support from ChaLearn, ANR AI chair HUMANIA ANR-19-CHIA-0022, TAILOR, an ICT48 network funded by EU Horizon 2020 program GA 952215, and the help of Mike Huisman to create the baselines code (except MetaDelta++), and of Romain Mussard, Manh Hung Nguyen, and Gabriel Lauzzana who worked as competition beta-testers. Experiments were performed using a Google cloud grant.

References

Yudong Chen, Chaoyu Guan, Zhikun Wei, Xin Wang, and Wenwu Zhu. MetaDelta: A Meta-Learning System for Few-shot Image Classification. In I Guyon, J. N. van Rijn, S. Treguer, and J. Vanschoren, editors, Proceedings of the AAAI Workshop on Meta-Learning and MetaDL Challenge, volume 140 of Proceedings of Machine Learning Research, pages 17–28. PMLR, 2021. URL https://proceedings.mlr.press/v140/chen21a.html.

Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Selecting Relevant Features from a Universal Representation for Few-shot Classification. arXiv preprint, 2020. URL https://arxiv.org/abs/2003.09338.

Adrian El Baz, Isabelle Guyon, Zhengying Liu, Jan N. van Rijn, Sebastien Treguer, and Joaquin Vanschoren. Advances in MetaDL: AAAI 2021 challenge and workshop. In I Guyon, J. N. van Rijn, S. Treguer, and J. Vanschoren, editors, Proceedings of the AAAI Workshop on Meta-Learning and MetaDL Challenge, volume 140 of Proceedings of Machine Learning Research, pages 1–16. PMLR, 2021. URL https://proceedings.mlr.press/v140/el-baz21a.html.

Adrian El Baz, Ihsan Ullah, Edesio Alcobaça, André C. P. L. F. Carvalho, Hong Chen, Fabio Ferreira, Henry Gouk, Chaoyu Guan, Isabelle Guyon, Timothy Hospedales, Shell Hu, Mike Huisman, Frank Hutter, Zhengying Liu, Felix Mohr, Ekrem Öztürk, Jan N. van Rijn, Haozhe Sun, Xin Wang, and Wenwu Zhu. Lessons learned from the NeurIPS 2021 MetaDL challenge: Backbone fine-tuning without episodic meta-learning dominates for few-shot learning image classification. In D. Kiela, M. Ciccone, and B. Caputo, editors, Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track, volume 176 of Proceedings of Machine Learning Research, pages 80–96. PMLR, 2022. URL https://proceedings.mlr.press/v176/el-baz22a.html.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In D. Precup and Y. W. Teh, editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of

Machine Learning Research, pages 1126—1135. PMLR, 2017. URL https://proceedings.mlr.press/v70/finn17a.html.

Isabelle Guyon, Lisheng Sun-Hosoya, Marc Boullé, Hugo Jair Escalante, Sergio Escalera, Zhengying Liu, Damir Jajetic, Bisakha Ray, Mehreen Saeed, Michèle Sebag, Alexander R. Statnikov, Wei-Wei Tu, and Evelyne Viegas. Analysis of the AutoML Challenge Series 2015–2018. In F. Hutter, L. Kotthoff, and J. Vanschoren, editors, Automated Machine Learning - Methods, Systems, Challenges, pages 177–219. Springer International Publishing, 2019. doi: 10.1007/978-3-030-05318-5_10.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Muhammad Abdullah Jamal and Guo-Jun Qi. Task Agnostic Meta-Learning for Few-Shot Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11711–11719, 2019. doi: 10.1109/CVPR.2019.01199.

Wei-Hong Li, Xialei Liu, and Hakan Bilen. Universal Representation Learning from Multiple Domains for Few-shot Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 9506——9515, 2021. doi: 10.1109/ICCV48922.2021.00939.

Wei-Hong Li, Xialei Liu, and Hakan Bilen. Cross-domain Few-shot Learning with Task-specific Adapters. arXiv preprint, 2022. URL https://arxiv.org/abs/2107.00358.

Lu Liu, William L. Hamilton, Guodong Long, Jing Jiang, and Hugo Larochelle. A Universal Representation Transformer Layer for Few-Shot Image Classification. In Proceedings of the 9th International Conference on Learning Representations (ICLR), 2021a. URL https://openreview.net/forum?id=04cII6MumYV.

Zhengying Liu, Zhen Xu, Meysam Madadi, Julio Jacques Junior, Sergio Escalera, Shangeth Rajaa, and Isabelle Guyon. Overview and unifying conceptualization of Automated Machine Learning. In Proceedings of the Automating Data Science workshop at ECML PKDD, 2019.

Zhengying Liu, Adrien Pavao, Zhen Xu, Sergio Escalera, Fabio Ferreira, Isabelle Guyon, Sirui Hong, Frank Hutter, Rongrong Ji, Julio C. S. Jacques Junior, Ge Li, Marius Lindauer, Zhipeng Luo, Meysam Madadi, Thomas Nierhoff, Kangning Niu, Chunguang Pan, Danny Stoll, Sebastien Treguer, Jin Wang, Peng Wang, Chenglin Wu, Youcheng Xiong, Arbër Zela, and Yang Zhang. Winning Solutions and Post-Challenge Analyses of the ChaLearn AutoDL Challenge 2019. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(9):3108–3125, 2021b. doi: 10.1109/TPAMI.2021.3075372.

Cheng Perng Phoo and Bharath Hariharan. Self-training For Few-shot Transfer Across Extreme Task Differences. In Proceedings of the 9th International Conference on Learning Representations (ICLR), 2021. URL https://openreview.net/forum?id=O3Y56aqpChA.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical Networks for Few-shot Learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples. In Proceedings of the 8th International Conference on Learning Representations (ICLR), 2020. URL https://openreview.net/forum?id=rkgAGAVKPr.

Eleni Triantafillou, Hugo Larochelle, Richard Zemel, and Vincent Dumoulin. Learning a Universal Template for Few-shot Dataset Generalization. In M. Meila and T. Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 10424–10433. PMLR, 2021. URL https://proceedings.mlr.press/v139/triantafillou21a.html.

Ihsan Ullah, Dustin Carrion, Sergio Escalera, Isabelle Guyon, Mike Huisman, Felix Mohr, Jan N. van Rijn, Haozhe Sun, Joaquin Vanschoren, and Phan Anh Vu. 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/.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: networked science in machine learning. ACM SIGKDD Explorations Newsletter, 15(2):49–60, 2014. doi: 10.1145/2641190.2641198.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf.