HAL
open science

Lessons learned from the NeurIPS 2021 MetaDL challenge: Backbone fine-tuning without episodic meta-learning dominates for few-shot learning image classification

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, et al.

To cite this version:

Adrian El Baz, Ihsan Ullah, Edesio Alcobaça, André C. P. L. F. Carvalho, Hong Chen, et al.. Lessons learned from the NeurIPS 2021 MetaDL challenge: Backbone fine-tuning without episodic meta-learning dominates for few-shot learning image classification. NeurIPS 2021 Competition and Demonstration Track, Dec 2021, On-line, United States. hal-03688638

HAL Id: hal-03688638

https://hal.science/hal-03688638v1

Submitted on 7 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Lessons learned from the NeurIPS 2021 MetaDL challenge: Backbone fine-tuning without episodic meta-learning dominates for few-shot learning image classification

Adrian El Baz[*] EB.ADRIAN8@GMAIL.COM
Ihsan Ullah IHSAN2131@GMAIL.COM
Edesio Alcobaça E.ALCOBACA@GMAIL.COM
André C. P. L. F. Carvalho ANDRE@ICMC.USP.BR
Hong Chen H-CHEN20@MAILS.TSINGHUA.EDU.CN
Fabio Ferreira FERREIRA@CS.UNI-FREIBURG.DE
Henry Gouk HENRY.GOUK@ED.AC.UK
Chaoyu Guan GUANCY19@MAILS.TSINGHUA.EDU.CN
Isabelle Guyon GUYON@CHALEARN.ORG
Timothy Hospedales T.HOSPEDALES@ED.AC.UK
Shell Hu SHELL.HU@SAMSUNG.COM
Mike Huisman M.HUISMAN@LIACS.LEIDENUNIV.NL
Frank Hutter FH@CS.UNI-FREIBURG.DE
Zhengying Liu ZHENGYING.LIU@INRIA.FR
Felix Mohr FELIX.MOHR@UNISABANA.EDU.CO
Ekrem Öztürk OZTURK@INFORMATIK.UNI-FREIBURG.DE
Jan N. van Rijn J.N.VAN.RIJN@LIACS.LEIDENUNIV.NL
Haozhe Sun HAOZHE.SUN@UNIVERSITE-PARIS-SACLAY.FR
Xin Wang XIN_WANG@TSINGHUA.EDU.CN
Wenwu Zhu WWZHU@TSINGHUA.EDU.CN

Editor: Editor's name

Abstract

Although deep neural networks are capable of achieving performance superior to humans on various tasks, they are notorious for requiring large amounts of data and computing resources, restricting their success to domains where such resources are available. Meta-learning methods can address this problem by transferring knowledge from related tasks, thus reducing the amount of data and computing resources needed to learn new tasks. We organize the MetaDL competition series, which provide opportunities for research groups all over the world to create and experimentally assess new meta-(deep)learning solutions for real problems. In this paper, authored collaboratively between the competition organizers and the top-ranked participants, we describe the design of the competition, the datasets, the best experimental results, as well as the top-ranked methods in the **NeurIPS 2021 challenge**, which attracted 15 active teams who made it to the final phase (by outperforming the baseline), making over 100 code submissions during the feedback phase. The solutions of the top participants have been **open-sourced**. The lessons learned include that learning good representations is essential for effective transfer learning.

Keywords: Automated Machine Learning, meta-learning, competition

[*] The two first authors are principal challenge organizer and dataset preparer; the other authors are in alphabetical order.

© 2022 A. El Baz *et al.*

1. Introduction

Automated Machine Learning (AutoML) has made great progress in the past years, driven by the increasing need for solving machine learning and data science tasks efficiently with limited human effort. A variety of approaches have been proposed, including Autosklearn (using Bayesian optimization, Feurer et al. (2015)), MLPlan (using tree Search, Mohr et al. (2018)), TPOT (using genetic algorithms, Olson and Moore (2019)), and learning curves (Mohr and van Rijn, 2022).

Machine learning challenges have been instrumental in benchmarking AutoML methods (Guyon et al., 2019), and bringing to the community state-of-the-art, open-source solutions, such as auto-sklearn (Feurer et al., 2015). Recently, AutoML challenges have started addressing deep learning problems (Liu et al., 2020, 2021) and it has become clear that meta-learning is one central aspect that deserves more attention. Meta-learning (Brazdil et al., 2022) aims at 'learning to learn' more effectively and transferring expertise from task to task, to improve performance, cut down training times and the need for human expertise, and/or reduce the number of training examples needed.

Advances in meta-learning are beneficial for a wide range of scientific domains, in particular when obtaining a large number of examples of any given task is costly. Such problems are often encountered, e.g., in medical image analysis, diagnosis of rare diseases, design of new materials and analysis of questionnaires. Other areas in which meta-learning is particularly critical include multi-class classification problems in which some classes are particularly rare, to the extent that only one or two examples may be available, e.g., rare plant or animal species. Such applications of meta-learning, focusing on rare, expensive, or difficult data to collect, have obvious economic and societal impact. Furthermore, work on meta-learning will contribute to advancing methods, which reduce the need for human expertise in machine learning and democratize the use of machine learning (by open-sourcing automated methods). In this challenge, we have aimed to maximize the societal impact of our effort by assembling datasets from a variety of practical domains, with direct relevance to 'AI for good', including medicine, ecology, biology, and pharmacology.

Our contributions are the following. We have developed a new benchmark for meta-learning, consisting of 10 new datasets from 5 practical domains, which we aim to make publicly available. The challenge has been entered by 15 teams,[1] who made a total of more than 100 submissions. The top-ranked method, MetaDelta++ (Chen et al., 2021), was able to achieve an accuracy score of more than 0.92 on all of the 5 hidden datasets.

2. Competition Setup

In this section, we describe the framework, the datasets and the implementation of the challenge.

2.1. Framework

The competition follows the protocol of El Baz et al. (2021), aiming at testing models on **few-shot learning problems**, in the 'N-way k-shot setting'. The main differences in this

1. See: https://autodl.lri.fr/competitions/210, this excludes test submissions made by various organizers

new competition, are that more and harder datasets are provided, the number of shots (examples per class) is increased from 1 to 5, while the number of ways (classes per task) remains constant at 5.

The competition spans two phases: (1) a **feedback phase** (in which participants iteratively develop and submit their methods and get immediate performance feedback on 5 meta-datasets hidden on the platform by submitting their code), and (2) a **final phase** (in which the last method of each participant submitted in the feedback phase is evaluated on 5 fresh final meta-datasets). The feedback phase lasted for 2 months, after which the final phase was run to determine the final ranking of the participants, using the final submission of the participants, in a fully blind test (no feedback).

This competition focuses on few-shot image classification tasks. We use the N-way, k-shot classification setting, commonly used in the meta-learning literature (Finn et al., 2017; Snell et al., 2017; Huisman et al., 2021b). The setup includes two stages: meta-training and meta-test. Given a (meta-)dataset, two disjoint class pools are created, one class pool is used for meta-training while the other one is used for meta-test. Each stage contains (possibly overlapping) tasks/episodes, each of which contains exactly N classes (called "ways") and k examples per class (called "shots"). A meta-training set includes examples of such tasks (pairs of training and test sets; both labelled). A meta-test set is then presented to the meta-trained algorithms, including labelled training sets and unlabeled test sets. Since the meta-test classes were not seen during training, this is a form of out-of-distribution evaluation (Setlur et al., 2021).

Each submitted script follows a specific API (as outlined by El Baz et al. (2021)), which we defined as challenge organizers. This API is designed to be flexible enough to be used to describe any meta-learning procedure. It defines 3 main classes which methods need to override to completely define a meta-learning algorithm. Its design relies on the definition of the different algorithms' implementation levels that have been identified by Liu et al. (2019). The 3 main classes are the following (as described by El Baz et al. (2021)):

- **Meta-learner**: has a `meta_fit()` method that encapsulates the meta-training procedure. Using the previously defined notation, it essentially processes the meta-dataset and captures the reusable information across meta-training tasks. It takes the meta-train set as an argument and outputs a **learner**.

- **Learner**: has a `fit()` method that encapsulates the training procedure (e.g., the adaptation phase). It takes a train set (examples and labels) as an argument along with the associated information from the meta-learning procedure to output a **predictor**.

- **Predictor**: has a `predict()` method that predicts the labels of test examples. It takes a test set (unlabeled examples) as an argument and returns the predictions. These can be evaluated by the competition software.

2.2. Datasets

We collected 10 datasets from 5 domains (two per domain). The meta-dataset of each phase consists of 5 datasets, one from each domain. This ensures that there is some resemblance

between the datasets that were public during the feedback phase and the datasets that the systems were evaluated on.

The datasets used in this competition come from five domains: ecology, bio-medicine, manufacturing, optical character recognition (OCR) and remote sensing, each having two datasets. The datasets are image datasets with at least 20 classes and 40 images per class. Sample images from each datasets are shown in Figure 1 (in Appendix A in the supplement).

All datasets, except OCR, are preprocessed, i.e., cropped, resized with an anti-aliasing filter into 128x128 size, and in some cases, some background padding was added to make the images fit in a square. The OCR datasets are generated directly in the required size by the OmniPrint software (Sun et al., 2021). Details about the domain, number of classes/categories and number of images in each dataset along with the phase of the competition in which the datasets have been used is shown in Table 1 (in Appendix A in the supplement). These datasets (and others in preparation for our next challenge) will be released as part of a benchmark suite called Meta-Album, which will be made publicly available.

2.3. Implementation

The challenge is implemented on the CodaLab platform, which allows competitions with code submission and the implementation of flexible protocols. All submissions from the feedback phase were automatically run on CodaLab so that all participants were allocated an equal amount of computing resources on the same hardware. Once the feedback phase was completed, we did manually run the last valid submission of each participant on the meta-datasets associated with the final phase, on a dedicated CodaLab instance.

Each submission would run for a maximum of 10 hours (2 hours maximum per meta-dataset). The number of submissions was loosely constrained during the feedback phase (5 per day, 100 in total). However, in the final test phase, a single submission per team was allowed, thus preventing teams from obtaining feedback on their performance on the private test datasets.

In the feedback phase, scores on the 5 feedback phase datasets were publicly visible on the leaderboard, but the feedback phase datasets themselves (and their identity) were not accessible to the participants. The final evaluation was carried out in similar conditions on the final phase datasets. The final datasets and their identity also remained inaccessible to the participants.

NeurIPS 2021 MetaDL competition was built upon the infrastructure of the previous AAAI 2021 MetaDL competition. We provided participants with the same starting package that we used for the AAAI 2021 competition.[2] It contains baseline methods such as MAML (Finn et al., 2017), Prototypical networks (Snell et al., 2017) and a naive baseline which accumulates all data from meta-training and trains a neural network on it.

3. Top ranked participants

All participants were invited to fill out fact sheets. Top-ranked teams summarized theirs below and were invited to co-author this paper.

2. https://github.com/ebadrian/metadl/tree/master/starting_kit

3.1. MetaDelta++

The authors of MetaDelta++ (displayed as team 'ForeverYong' in the ranking of Table 1) won the challenge. They enhanced their previous MetaDelta solution, which won the AAAI 2021 MetaDL-mini competition (Chen et al., 2021). It is composed of three base meta-learners with a pre-trained backbone at multi-scale input and different training strategies. Specifically, for each domain (since all domains are meta-trained and meta-tested separately) the parameters of the pre-trained backbone are adapted for each meta-learner to the specific meta-training subset (including samples from a subset of the classes in the given meta-dataset, the other classes being used for meta-testing). To that end, the authors remove the original last layer and substitute it with a 3-layer MLP classifier, then fine-tune the backbone by freezing the parameters of the shallower layers (layers close to the input) to preserve the general knowledge and prevent overfitting to the small meta-training dataset. Furthermore, hand-crafted data augmentation (like rotation) is designed to help the fine-tuning process. After obtaining a fine-tuned backbone, for each meta-learner, the three-layer MLP is removed and replaced by a prototype-based classification method during meta-testing (inference stage), modelled as an optimal transport problem (Hu et al., 2021), based on the features extracted by the backbone. Finally, the three meta-learners are 'auto-ensembled' to stabilize the performance of the whole meta-learning system, by which we mean that the best backbone is selected by (meta-)cross-validation. They carefully monitor the time budget with a well-designed controller.

Ablation studies, conducted by the authors, show that the pre-trained backbones and the specially designed fine-tuning methods are of great significance for few-shot learning, while data augmentation and optimal transport can also help but not so critically.

The authors made the solution available on GitHub.[3]

3.2. Edinburgh-Samsung: Self-supervised transfer learning

Following recent debate about simple embedding-learning *vs.* meta-learning (Chen et al., 2019; Tian et al., 2020; Zhang et al., 2021), this team (displayed as team 'henrygouk' in the ranking of Table 1) eschewed meta-learning altogether and focused exclusively on leveraging larger external pre-training datasets and neural architectures. In particular, due to a previous observation showing self-supervised features often obtain better performance in transfer learning settings (Ericsson et al., 2021), they ultimately exploited the powerful vision transformer architecture trained with self-supervision by DINO (Caron et al., 2021) on ImageNet1K. The use of this substantial external dataset allowed fitting the larger ViT model. Their entry simply extracted DINO/ViT features and trained a well regularised logistic regression classifier during the meta-test step.

The simplicity and comparatively high performance of this solution demonstrate that large scale pre-training and modern architectures may often be the easiest way to achieve high-performance few-shot learning in practice. A refined version of this approach is under review (Hu et al., 2022). The code for this entry is available on GitHub.[4]

3. https://github.com/Frozenmad/MetaDelta
4. https://github.com/henrygouk/neurips-metadl-2021

3.3. Meta-Padawan

The Meta-Padawan (displayed as team 'padawan' in the ranking of Table 1) solution learns from previous pre-trained models, combining features descriptors previously learned to achieve fast model generalization using few images. The hypothesis was that it is possible to rapidly generalize new deep neural network models by combining different features descriptors learned from previous models with simple features extracted from images themselves. To assess this hypothesis, Meta-Padawan designed the 2SoF (Two Set of Features) method, which combines features from pre-trained models and images. Initially, 2SoF does data augmentation using training images, generating 10 new images for each image. This uses a rotation range of 20, zoom range of 0.15, width/height shift of 0.2, the sheer intensity of 0.15, and horizontal flip. Afterwards, 2SoF extracts features from each augmented image using Principal Component Analysis (PCA), capturing 95% of the variance. These features complement the features extracted from the last convolutional layer from pre-trained models InceptionResNetV2 and VGG12 (Simonyan and Zisserman, 2015; Szegedy et al., 2017). The combination of the PCA-based features with the features extracted by the last convolutional layers of deep networks creates a new feature space. Finally, the proposed method fits a Logistic Regression classification model with L2 regularization to this new feature space. No systematic ablation studies were conducted, but the authors confirm that both types of features contributed to the good performance of their method.

The code of Meta-Padawan was made publicly available on GitHub.[5]

3.4. Meta-ZAP

The authors of Meta-ZAP (Meta Zero-shot AutoML from Pre-trained Models, displayed as team 'ekremtzr' in the ranking of Table 1) adapted their method from the AutoDL competition (Liu et al., 2021), in which they meta-learn a model that selects the proper meta-learning pipeline, i.e., a pre-trained feature extractor and the hyperparameters of the meta-learning pipeline, based on the properties of a given dataset (e.g., the number of images and classes) in a zero-shot setting (without any model validation). For learning this model selector, they create a meta-dataset that consists of classification accuracies (measured as in the competition setting) of many meta-learning pipelines across datasets and learn it with the algorithm selection approach AutoFolio (Lindauer et al., 2015).

Contrary to their previous work in the AutoDL competition (Liu et al., 2021), they adapt their approach to the MetaDL setting by using a different configuration space consisting of seven hyperparameters to search for a meta-model and by fixing hyperparameters on fine-tuning (Saikia et al., 2020). Moreover, they generate 21 datasets via ICGen (Stoll, 2020), a dataset-level augmentation tool, from several TensorFlow Datasets (Abadi et al., 2015; Google, 2021) (TFDS) datasets and also include eight datasets from meta-dataset (Triantafillou et al., 2020). They utilize Task2Vec (Achille et al., 2019) embeddings as dataset features. Additionally, they populate the source and support examples by applying image augmentations via TrivialAugment (Müller and Hutter, 2021).

The Meta-ZAP solution was open-sourced on GitHub.[6]

5. https://github.com/ealcobaca/meta-padawan
6. https://github.com/ekremozturk/ZAP-few-shot

		Meta-dataset 1		**Meta-dataset 2**		**Meta-dataset 3**		**Meta-dataset 4**		**Meta-dataset 5**	
Team	**Average rank**	Accuracy	Rank	Accuracy	Rank	Accuracy	Rank	Accuracy	Rank	Accuracy	Rank
ForeverYong	3.80	0.983	2	0.943	**1**	0.990	**1**	0.921	**1**	- (0.939*)	14 (1*)
henrygouk	4.00	0.979	4	0.770	2	0.419	9	0.803	4	0.875	**1**
padawan	4.00	0.981	3	0.714	4	0.488	7	0.856	3	0.865	3
pikachu	4.20	0.985	**1**	0.611	6	0.355	10	0.904	2	0.873	2
ekremtzr	4.60	0.937	8	0.732	3	0.921	2	0.762	6	0.817	4
sheling343	5.20	0.977	5	0.659	5	0.515	6	0.778	5	0.783	5
paisiasach	6.20	0.949	6	0.412	8	0.684	4	0.698	7	0.732	6
BucketHead	7.20	0.946	7	0.411	9	0.786	3	0.660	8	0.639	9
lucia	7.80	0.842	10	0.467	7	0.552	5	0.612	9	0.698	8
ericlhan	8.80	0.851	9	0.382	10	0.458	8	0.611	10	0.719	7
perathem	11.20	0.749	12	0.334	11	0.208	12	0.411	11	0.538	10
brunosez	12.00	0.752	11	0.330	12	0.206	14	0.409	12	0.516	11
vermashreth	13.20	0.453	14	0.285	13	0.235	11	0.307	14	0.304	14
Vilupa	13.60	0.422	15	0.247	14	0.207	13	0.316	13	0.339	13
mrm	14.00	0.563	13	0.244	15	0.20	15	0.261	15	0.414	12

Table 1: Final phase results of the NeurIPS 2021 MetaDL competition. For each meta-dataset, the result displayed is the worst out of 3 runs. Again, for each meta-dataset, each algorithm is evaluated on 600 episodes with a 5-way 5-shot configuration. Classes and associated images are drawn from the associated meta-test set. The Accuracy columns are the average accuracy over the 600 episodes of the corresponding meta-dataset. *ForeverYong 5th meta-dataset result is displayed for completeness when the minor bug encountered is fixed.

4. Results

In this section, we overview the results of the final phase and describe the method of how the individual scores are aggregated to a final ranking.

4.1. Final phase overview

During the final phase, the last valid feedback phase submission from participants was considered and re-run from scratch on 5 new meta-datasets. As in the feedback phase, every submission is ranked on each meta-dataset independently and the final score of the submission is the average of these ranks. A submission from the feedback phase is considered valid if it beats the determined baseline using a popular few-shot learning method: Prototypical Networks (Snell et al., 2017). A few participants had multiple submissions that beat the baseline, but only their last feedback phase submission was considered. Ultimately we had 15 participants entering the final phase.

4.2. Evaluation protocol and results

The 5 new meta-datasets in the final phase had similar image domains to the ones used during the feedback phase. To increase reproducibility, and avoid participants winning by chance, we ran each submission 3 times, using different random seeds. For each meta-dataset, we considered only the lowest accuracy among these 3 runs. The participants' results are displayed in Table 1.

We evaluated several metrics to rank scores of each submission across meta-datasets: average rank, relative difference and the Copland method. Regardless of the metric used we

obtained the same top-4 results. The choice between these metrics is tightly linked to the goal of the problem which usually is either to have a 'generalist' or 'specialist' algorithm (Pavao et al., 2021). The competition aims to foster algorithms that are capable of quickly dealing with few-shot image classification problems within a single image domain. Hence, we preferred the average rank metric which emphasized generalist behaviour.

MetaDelta++ (team indicated as 'ForeverYong') finished in first place with an improvement of the algorithm that got them the first place in the first edition of MetaDL at AAAI 2021 (El Baz et al., 2021). Most notably, they did reach first place even though their algorithm encountered an error during the evaluation of the 5th meta-dataset, which resulted in the worst rank for the associated meta-dataset rank. The minor error was mainly due to a hardcoded number of classes to draw for episode generation, which worked well for the feedback phase but did not for the final phase. The next version of MetaDL will explicitly disclose final phase metadata to the participants to avoid these unfortunate errors. Also, there was a tie between the second and the third place and our competition rules state that in the occurrence of such an event, the algorithm that was submitted first is considered better. Therefore, Edinburgh-Samsung (displayed as 'henrygouk') finished in the second position, while Meta-Padawan (displayed as 'padawan') finished in the third position.

5. Discussion

In designing a good meta-learning system, several decision must be made:

1. whether to use a **filter** method to perform model selection and hyper-parameter selection (referred to as zero-shot learning by team Meta-ZAP) or a **wrapper** method (with meta-cross-validation) like MetaDelta++, and whether **meta-features** (extracted or provided) should be used to help model selection;

2. whether to use a **pre-trained backbone**, and if so, whether to **fine-tune** its weights and how (with or without the **provided meta-datasets** and with or without **data augmentation**;

3. whether to use other types of **feature extraction/embeddings** instead or in conjunction with features extracted from a pre-trained backbone;

4. which type of **classifier** to use after feature extraction, amenable to few-shot learning (e.g., simple linear classifier or example-based methods);

5. whether and how to use model **ensembles**;

6. how to efficiently manage the **time budget**.

Regarding filter *vs.* wrapper, in this challenge, the wrapper setting used by MetaDelta++ seems to have been very effective. It is more computationally expensive than filter methods, but the winning team made efficient use of the computational resources and therefore could afford it. Possibly a smart combination of filter and wrapper methods would be advisable, e.g., using filter methods as initialization or prior in a smart search strategy inspired by Bayesian optimization. This was not explored by the participants. Some participants attempted to compute dataset meta-features and noted that the organizers did not provide

such meta-features, such as application domain, the scale of the image, etc. In particular, Meta-Padawan suggested providing a wider variety of image subtopics and pre-trained models to enhance the initial meta-knowledge.

Regarding the use of pre-trained backbones, a complete consensus emerged: they are essential. All top-ranking participants used convolutional backbones or transformer models. The winners MetaDelta++ managed to efficiently fine-tune the backbones they chose, but they warned about the danger of overfitting the meta-training data. To avoid that, they froze the layers closest to the input and used data augmentation. Interestingly, the second-place participants did not fine-tune their backbone. It was trained on ImageNet with self-supervised learning. Massive pre-training with large unlabeled datasets may be the solution of the future for image classification. However, this strategy relies on the fact that massive out-of-domain data is available. The value of small amounts of in-domain data vs large amounts of out-of-domain data (or how to combine both) remains to be seen. In future challenges, we intend to have a dedicated track for comparing with de-novo training methods, which can only rely on the data provided.

Another consensus is that 'episodic' training (like MAML) is not necessary. Backbones are fine-tuned with regular gradient descent and classifiers are trained with standard classification algorithms. Several teams used a simple linear classifier, trained e.g., with logistic regression. The winners MetaDelta++ favoured a prototype-based method using optimal transport to compute a discrepancy metric rather than Euclidean distance. This finding is corroborated by other studies which demonstrate the benefit of access to global class labels (Wang et al., 2021a,b) and show that classical episodic training can lead to suboptimal performance (Laenen and Bertinetto, 2021).

Finally, the optimization of usage of computational resources played an important role in winning, as reported by the participants. However, it is difficult to assess since it is entangled with algorithm implementation. In particular, the competitors of the Meta-Padawan team shared that efficient time management and code were critical for them. They reported that processing costs could be significantly reduced by limiting data augmentation or avoiding unnecessary code loops.

From an organizational perspective, there were several pitfalls in our competition protocol that we will correct in the future. The first thing that stands out from the results is that the MetaDelta++ solution (Team 'ForeverYong' in Table 1) performs best on all public feedback datasets and on four out of five private test datasets, but *did not successfully run on the fifth dataset*. Analyzing the code, we discovered that this was due to a special case that was not anticipated. In the future, we will provide more meta-data to facilitate anticipating such cases or a variety of datasets in the feedback phase covering all scenarios encountered in the final phase. There is also room for improvement in our scoring method. Although the ranking of top participants was stable concerning the various possible methods of combining the 5 scores (each corresponding to a domain), it is known that the average rank method we used is influenced by the removal or addition of participants. When a team submits a solution that crashes on a single dataset, as a consequence it will be ranked last. The severity of this penalty depends now on the number of participants; when there are only few participants this might not be that severe, whereas when there are many participants, this can be quite severe. One possible remedy could be to introduce a maximum rank that participants can achieve so that one dataset does not extremely in-

fluence the score. We might also conduct bootstrap experiments to study the stability of ranking and report distributions rather than single results (see, e.g., Turner et al. (2021)).

6. Conclusion and further work

We organized MetaDL (Meta-Deep-Learning), a competition to benchmark state-of-the-art meta-learning techniques and to further improve the field, focusing on few-shot image classification. The MetaDelta++ system (Chen et al., 2021), based on pre-trained backbone networks, performed best on most of these datasets. Other well-performing solutions are Edinburgh-Samsung (Hu et al., 2022), Meta-Padawan and Meta-ZAP, which are all described in this paper. All winning solutions are open-sourced.

This competition is part of a well-established competition series, consisting of (among others) the AutoML and AutoDL competition series. Although it does not specifically 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). Fine-tuning on meta-training data turned out to be important, though there are indications that off-the-shelf backbones pre-trained with self-supervised learning on massive datasets might become the way of the future, essentially making meta-learning unnecessary for image classification problems. Thus, meta-learning should be benchmarked in *de novo* training conditions, in the future, to prepare for scenarios (in other domains) in which such backbones are not available.

In this competition, we have refined the 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, our setting remained relatively simplified, in that images were small (128x128), meta-training and testing were performed within sub-domains (e.g., insect classification, texture classification, OCR, etc.), with sub-tasks involving only 5 classes (ways) relatively well separated (not from the same super-category) and with 5-shots. Indeed, the winners obtained over 92% accuracy on all 5 domains in the final phase (with complete blind-testing of their code). Thus, we must move on to harder problems. In an upcoming challenge, we intend to mix tasks from multiple domains and present variable numbers of 'ways' and 'shots'. The participants also expressed the desire that we would organize an even more challenging competition, in the spirit of AutoDL (Liu et al., 2021), with meta-datasets stemming not only from different domains but different modalities (speech, image, video, text, tabular data, etc.) We might do this next.

For this competition, we have assembled 10 new datasets, from 5 practical domains. From each domain, one dataset was used in a public feedback phase, and the other dataset was used in a private test phase. As reported by several top-ranking teams, meta-learning was possible within sub-domains (in the form of fine-tuning pre-trained backbone networks), 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. To further probe the effectiveness of various meta-learning solutions, we are preparing a larger cross-domain meta-learning challenge. We intend to re-design our API to facilitate ablation studies by modularizing the meta-learning workflow.

Acknowledgements

This paper is authored collaboratively by both the organizers and top-ranked participants of the competition. All participants that ranked in the top-5, and also filled in the fact-sheet detailing their method, were invited and co-authored the paper. Funding and support have been received by ANR Chair of Artificial Intelligence HUMANIA ANR-19-CHIA-0022, TAILOR (a project funded by EU Horizon 2020 research and innovation program under GA No. 952215) and ChaLearn, computing cloud units were donated by Microsoft and Google, and prizes were donated by 4Paradigm. We want to thank everyone that contributed to the creation of datasets: Jennifer (Yuxuan) He, Benjia Zhou, Professor Yui Man Lui, and Phan Anh Vu. We also thank Sébastien Treguer and Adrien Pavao for helpful discussions.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/.

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C. Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6429–6438. IEEE, 2019.

Nihar Bendre, Hugo Terashima Marín, and Peyman Najafirad. Learning from few samples: A survey. *arXiv preprint arXiv:2007.15484*, 2020.

Pavel Brazdil, Jan N. van Rijn, Carlos Soares, and Joaquin Vanschoren. *Metalearning: Applications to Automated Machine Learning and Data Mining*. Springer, 2nd edition, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, ICLR'19, 2019.

Yudong Chen, Chaoyu Guan, Zhikun Wei, Xin Wang, and Wenwu Zhu. Metadelta: A meta-learning system for few-shot image classification. *CoRR*, abs/2102.10744, 2021.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and*

Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, pages 3606–3613. IEEE Computer Society, 2014.

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 *AAAI Workshop on Meta-Learning and MetaDL Challenge*, volume 140 of *Proceedings of Machine Learning Research*, pages 1–16. PMLR, 2021.

Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5414–5423, 2021.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems 28*, pages 2962–2970, 2015.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

Mario Fritz, E. Hayman, B. Caputo, and J. Eklundh. The kth-tips database, 2004. URL https://www.csc.kth.se/cvap/databases/kth-tips/index.html.

G. Geetharamani and J. Arun Pandian. Identification of plant leaf diseases using a nine-layer deep convolutional neural network. *Computers & Electrical Engineering*, 76:323–338, 2019.

Google. TensorFlow Datasets, a collection of ready-to-use datasets. https://www.tensorflow.org/datasets, 2021.

Isabelle Guyon, Lisheng Sun-Hosoya, Marc Boullé, Hugo Jair Escalante, Sergio Escalera, Zhengying Liu, Damir Jajetic, Bisakha Ray, Mehreen Saeed, Michèle Sebag, Alexander Statnikov, Wei-Wei Tu, and Evelyne Viegas. *Analysis of the AutoML Challenge Series 2015–2018*, pages 177–219. Springer International Publishing, Cham, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Emily F. Brownlee Heidi M. Sosik, Emily E. Peacock. Annotated plankton images - data set for developing and evaluating classification methods., 2015. URL https://hdl.handle.net/10.1575/1912/7341.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2021.

Shell Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a

difference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

Yuqing Hu, Vincent Gripon, and Stéphane Pateux. Leveraging the feature distribution in transfer-based few-shot learning. In *Artificial Neural Networks and Machine Learning - ICANN 2021*, volume 12892 of *Lecture Notes in Computer Science*, pages 487–499. Springer, 2021.

Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7304–7308. IEEE, 2013.

David P. Hughes and Marcel Salathé. An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing. *CoRR*, abs/1511.08060, 2015.

Mike Huisman, Aske Plaat, and Jan N van Rijn. Stateless neural meta-learning using second-order gradients. *arXiv preprint arXiv:2104.10527*, 2021a.

Mike Huisman, Jan N Van Rijn, and Aske Plaat. A survey of deep meta-learning. *Artificial Intelligence Review*, 54(6):4483–4541, 2021b.

Norbert Jankowski, Wlodzislaw Duch, and Krzysztof Grabczewski, editors. *Meta-Learning in Computational Intelligence*, volume 358 of *Studies in Computational Intelligence*. Springer, 2011.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, NIPS'12, pages 1097–1105, 2012.

Gustaf Kylberg. The kylberg texture dataset v. 1.0. Technical Report 35, Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, Uppsala, Sweden, 2011.

Steinar Laenen and Luca Bertinetto. On episodes, prototypical networks, and few-shot learning. In *Advances in Neural Information Processing Systems*, 2021.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, 2005.

Haifeng Li, Xin Dou, Chao Tao, Zhixiang Wu, Jie Chen, Jian Peng, Min Deng, and Ling Zhao. Rsi-cb: A large-scale remote sensing image classification benchmark using crowd-sourced data. *Sensors*, 20(6):1594, 2020.

Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-SGD: Learning to Learn Quickly for Few-Shot Learning. *arXiv preprint arXiv:1707.09835*, 2017.

Marius Lindauer, Holger H. Hoos, Frank Hutter, and Torsten Schaub. AutoFolio: an automatically configured algorithm selector. *Journal of Artificial Intelligence Research*, 53(1):745–778, 2015.

Marius Lindauer, Jan N van Rijn, and Lars Kotthoff. Open algorithm selection challenge 2017: Setup and scenarios. In *Open Algorithm Selection Challenge 2017*, volume 79 of *PMLR*, pages 1–7, 2017.

Marius Lindauer, Jan N van Rijn, and Lars Kotthoff. The algorithm selection competitions 2015 and 2017. *Artificial Intelligence*, 272:86–100, 2019.

Zhengying Liu, Zhen Xu, Meysam Madadi, Julio Jacques Junior, Sergio Escalera, Shangeth Rajaa, and Isabelle Guyon. Overview and unifying conceptualization of Automated Machine Learning. *Automating Data Science workshop @ ECML PKDD 2019*, page 8, September 2019.

Zhengying Liu, Zhen Xu, Shangeth Rajaa, Meysam Madadi, Julio C. S. Jacques Junior, Sergio Escalera, Adrien Pavao, Sebastien Treguer, Wei-Wei Tu, and Isabelle Guyon. Towards automated deep learning: Analysis of the autodl challenge series 2019. In *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, pages 242–252. PMLR, 2020.

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, 2021.

Jiang Lu, Pinghua Gong, Jieping Ye, and Changshui Zhang. Learning from very few samples: A survey. *arXiv preprint arXiv:2009.02653*, 2020.

P. Mallikarjuna, Alireza Tavakoli Targhi, Mario Fritz, E. Hayman, B. Caputo, and J. Eklundh. The kth-tips 2 database, 2006. URL https://www.csc.kth.se/cvap/databases/kth-tips/index.html.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *International Conference on Learning Representations*, ICLR'18, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Felix Mohr and Jan N. van Rijn. Learning curves for decision making in supervised machine learning - A survey. *CoRR*, abs/2201.12150, 2022.

Felix Mohr, Marcel Wever, and Eyke Hüllermeier. Ml-plan: Automated machine learning via hierarchical planning. *Machine Learning*, 107(8-10):1495–1515, 2018.

Samuel G. Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 774–782, October 2021.

Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *Proceedings of the 34th International Conference on Machine Learning*, ICML'17, pages 2554–2563. JLMR.org, 2017.

Devang K Naik and Richard J Mammone. Meta-neural networks that learn by learning. In *International Joint Conference on Neural Networks*, pages 437–442. IEEE, 1992.

Alex Nichol, Joshua Achiam, and John Schulman. On First-Order Meta-Learning Algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Randal S. Olson and Jason H. Moore. TPOT: A tree-based pipeline optimization tool for automating machine learning. In Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren, editors, *Automated Machine Learning - Methods, Systems, Challenges*, The Springer Series on Challenges in Machine Learning, pages 151–160. Springer, 2019.

Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.

Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Adrien Pavao, Isabelle Guyon, Nachar Stéphane, Fabrice Lebeau, Martin Ghienne, Ludovic Platon, Tristan Barbagelata, Pierre Escamilla, Sana Mzali, Meng Liao, et al. Aircraft numerical "twin": A time series regression competition. In *20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 441–448. IEEE, 2021.

Sachin Ravi and Hugo Larochelle. Optimization as a Model for Few-Shot Learning. In *International Conference on Learning Representations*, ICLR'17, 2017.

Mengye Ren, Sachin Ravi, Eleni Triantafillou, Jake Snell, Kevin Swersky, Josh B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *International Conference on Learning Representations*, ICLR'18, 2018.

Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, ICLR'19, 2019.

Tonmoy Saikia, Thomas Brox, and Cordelia Schmid. Optimized generic feature learning for few-shot classification across domains. *CoRR*, abs/2001.07926, 2020.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with Memory-augmented Neural Networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, ICML'16, pages 1842–1850, 2016.

Jürgen Schmidhuber. Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook. Master's thesis, Technische Universität München, 1987.

Hortense Serret, Nicolas Deguines, Yikweon Jang, Grégoire Lois, and Romain Julliard. Data quality and participant engagement in citizen science: comparing two approaches for monitoring pollinators in france and south korea. *Citizen Science: Theory and Practice*, 4(1):22, 2019.

Amrith Setlur, Oscar Li, and Virginia Smith. Two sides of meta-learning evaluation: In vs. out of distribution. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems 30*, NIPS'17, pages 4077–4087. Curran Associates Inc., 2017.

Danny Stoll. Icgen, 2020. URL https://github.com/automl/ICGen.

Haozhe Sun, Wei-Wei Tu, and Isabelle M Guyon. Omniprint: A configurable printed character synthesizer. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, NIPS'21, 2021.

Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *European Conference on Computer Vision*, pages 206–222. Springer, 2016.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to Compare: Relation Network for Few-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208. IEEE, 2018.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer, 2018.

Sebastian Thrun. Lifelong Learning Algorithms. In *Learning to learn*, pages 181–209. Springer, 1998.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020.

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 *International Conference on Learning Representations*, 2020.

Ryan Turner, David Eriksson, Michael McCourt, Juha Kiili, Eero Laaksonen, Zhen Xu, and Isabelle Guyon. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 3–26. PMLR, 2021.

Joaquin Vanschoren. Meta-Learning: A Survey. *arXiv preprint arXiv:1810.03548*, 2018.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems 29*, NIPS'16, pages 3637–3645, 2016.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Haoxiang Wang, Han Zhao, and Bo Li. Bridging multi-task learning and meta-learning: Towards efficient training and effective adaptation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 10991–11002. PMLR, 2021a.

Ruohan Wang, massimiliano pontil, and Carlo Ciliberto. The role of global labels in few-shot classification and how to infer them. In *Advances in Neural Information Processing Systems*, 2021b.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

X. Zhang, D. Meng, H. Gouk, and T. Hospedales. Shallow bayesian meta learning for real-world few-shot recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.