MetaDL: Few Shot Learning Competition with Novel Datasets from Practical Domains[*]

Adrian El Baz Isabelle Guyon Zhengying Liu Jan N. van Rijn
Haozhe Sun Sebastien Treguer Ihsan Ullah Joaquin Vanschoren
Phan Anh Vu
metadl@chalearn.org

April 1, 2021

Abstract

Meta-learning is an important machine learning paradigm leveraging experience from previous tasks to make better predictions on the task at hand. The proposed challenge focuses on supervised learning, and more particularly "few shot learning" classification settings, aiming at learning a good model from very few examples, typically 1 to 5 per class. We will include various datasets from healthcare, ecology, biology, and chemistry. By using datasets from these practical domains, we aim to maximize the humanitarian and societal impact. The competition will consist of three phases: a public phase, a feedback phase, and a final phase. The last two phases will be run with code submissions, fully bind-tested on the Codalab challenge platform. A single (final) submission will be evaluated during the final phase, using five fresh datasets, previously unknown to the meta-learning community. Together with the public datasets and feed-back datasets, the datasets, formatted in a uniform manner for this competition, will constitute a long-lasting valuable resource for research in this field. The scientific and technical questions addressed by this challenge include scalability, robustness to domain diversity, influence of pre-training as a form of transfer learning, and effectiveness of episodic meta-learning (from many "small" tasks).

Keywords

Deep Learning, Automated Machine Learning, Meta-Learning, Few Shot Learning.

Competition type

Regular.

[*] The authors are in alphabetical order of last name.

1 Competition description

1.1 Background and impact

Background: Automating Machine Learning (AutoML) has made great strides in the past few years, driven by the increasing needs for solving machine learning and data science in a timely manner with limited human effort. Machine learning challenges have been instrumental in benchmarking AutoML methods [5], and bringing to the community state-of-the-art, open source solutions, such as auto-sklearn. Recently, AutoML challenges have started addressing Deep Learning problems [13] and it has become clear that **meta-learning** is one central aspect that deserves more attention. By meta-learning, we mean "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. The increase in interest of the NeurIPS community in meta-learning is rapidly rising as evidenced by the number of accepted paper: $\sim$ 50 in 2020, compared to $\sim$ 30 in 2019, $\sim$ 20 in 2018, and only 4 in 2017.

Impact: The domains, which benefit from meta-learning, cover the full range of classical machine learning applications, with particular emphasis in domains in which obtaining a large number of samples of any given task is costly. Such problems are often encountered in medical image analysis, for example. Other areas in which meta-learning is particularly critical include multi-class classification problems in which some classes are particularly depleted, 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 economical and societal impact. Furthermore, work on meta-learning will contribute to advancing methods, which **reduce the need for human expertise in machine learning**, and are instrumental in democratizing the use of machine learning (by open-sourcing automated methods). We aim 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**. Application examples in such domains should be inspiring to **deploy automated solutions powered by meta-learning in developing countries**, where acquiring large amounts of data and/or obtaining human expertise to develop learning machines may be difficult.

Novel contributions: While the benefits of having a meta-learning challenge are clear, designing one is far from trivial, particularly because of the heavy computational requirements, and the difficulty of obtaining enough training and test data (since in meta-learning problems, one example is one dataset). This challenge proposes **innovative protocols for evaluating meta-learning with code submission**, and will make available to the community **novel datasets, never used before as a benchmark in this field**, and larger in size and domain diversity. Because of the emphasis on domains in which Deep Learning excels, such as computer vision, we coined the name **MetaDL (for Meta-Deep-Learning)**, although the participants will be free to use other techniques.

Scientific motivations: The motivations for advancing research in MetaDL are many, including reducing the need for large training sets, reducing the need for human expertise, cutting down on hyper-parameter tuning, and generally making learning more efficient and getting better generalization, by capitalizing on experience gained by learning on other similar tasks. Indeed, the performance of many Machine Learning algorithms, and in particular Deep Learning, depends highly upon the quality and quantity of available data, and (hyper)-parameter settings. This has motivated recent progress on hyperparameter optimization, with methods including random search, Bayesian optimization [17, 18, 10], evolutionary optimization, and bandit-based methods [11]. Meta-Learning is a way to address both issues. Simple, but effective approaches reported recently include pre-training models on similar datasets [19, 7, 2, 6]. This way, a good algorithm or hyperparameters can be pre-determined or several model parameters can be transferred to the new dataset. As such, higher performance can be achieved with the same amount of data, or similar performance with less data (*few shot learning*). The field of few shot learning is versatile, with many emerging techniques, see, e.g., MAML [3], Prototypical Networks [16], Latent Embedding Optimization [15], PT+MAP [20]. Our interest has been drawn to this particular setting of meta-learning, both because of its fundamental interest and because it is amenable to being benchmarked more easily since each task includes only few examples.

Although benchmarks in meta-learning exist[1], many techniques have been evaluated extensively, and code is often made publicly available, we identify three forms of bias:

Benchmark suites Techniques are often compared against each other on a small set of standard benchmarks, such as Omniglot, mini-ImageNet and the CUB birds dataset. Data scientists will get a better understanding of these specific benchmark suites, and as such it is questionable whether new techniques are generally applicable or will work well only on the current set of benchmark suites.

Default hyperparameters Due to computational demands, the experimental protocol is often designed with efficiency in mind, neglecting proper neural architecture search and hyperparameter optimization. As such, it is questionable whether a new contribution performs better due to novelty, or due to better fine-tuning of the hyperparameters or neural architecture.

Baselines in each paper introducing a new contribution, the authors make an educated assessment what would be the best baselines, and are responsible for selecting the right hyperparameters for also the baseline methods.

[1] For examples of benchmarks, see `https://paperswithcode.com/task/meta-learning`

Open competitions and challenges are a great way to address these issues and avoid the "inventor-evaluator bias". On one hand, the competition organizers introduce new datasets challenging the participants beyond their zone of comfort; on the other hand the participants have all freedom to provide suitable models, search strategies, architectures, hyperparameters, etc. Finally, all participants are given equal server-side resources, making the competition fair for participants from institutes that have less computational resources.

Scientific questions addressed in the challenge: Based on previous competitions we organized, including a first small-scale version of MetaDL that served to prototype our competition protocol[2], we identified the following **research scientific and technical questions**, which will be addressed in the challenge:

1. **Scalability:** We will be challenging participants to address supervised learning tasks of practical interest in domains such as medical imaging, pharmacology, ecology, and printed text recognition in the wild, as opposed to toy problems such as CIFAR-100 and Omniglot.

2. **Universality:** While we will format all datasets in a common tensor-based format, the modalities addressed will be diverse and include tabular data, images, and sequences. Because of domain diversity, the type of data distribution will vary a lot.

3. **Influence of pre-training:** A poor-man's way of performing meta-learning is to reuse Deep Learning models trained on massive datasets such as ImageNet, rip out the last layer and perform shallow learning, using the features of the pre-trained model. While this method may be effective, we want to decouple its effect from further algorithmic advances. Hence we will have two tracks, one with and one without pre-training outside of the platform.

4. **Identification of key "ingredients":** A variety of ingredients (other than pre-training) may play a key role in the success of proposed methods. This includes data management (e.g. episodic learning), hyper-parameter selection, model evaluation, and ensembling solutions. To facilitate conducting systematic experiments, we will propose an API making use of a generic workflow inspired by the wining strategies in our proposed starting kit. While this may bias solution, this will facilitate our post-challenge experiments to conduct ablation studies and A/B testing [1] and functional ANOVA [8] .

5. **Domain adaptation:** We intend to test knowledge transfer between domain as well as as within domain. To that end, post-challenge experiments will include training on all domains but one and testing on the remaining one.

[2] See `https://metalearning.chalearn.org/`.

Table 1: ChaLearn Competition Series

Conference	Challenge	Description
ICML & NeurIPS 2016-18	AutoML	Automating complete ML pipeline
WAIC 2019	AutoNLP	Natural Language Processing
ECML PKDD 2019	AutoCV	Computer Vision
ACML 2019	AutoWeakly	Weakly Supervised Learning
WSDM 2019	AutoSeries	Time Series
NeurIPS 2019	AutoDL	Misc. domains
KDD cup 2020	AutoGraph	Classification of Graph Data
InterSpeech 2020	AutoSpeech	Speech Recognition
AAAI 2021	2020 MetaDL-mini	Few shot learning, trial run

Participation and reach out to under-represented communities: The challenge will be implemented on the Codalab platform, which allows competitions with code submission and the implementation of flexible protocols. This will also ensure attracting participants, who are regularly entering Codalab competitions. Based on past challenges and workshops we organized in AutoML, AutoDL, and meta-learning, we anticipate that at least **100 participants** will enter the challenge, including some of the most skilled teams presently working on the topic, who participated in our recent AAAI 2021 workshop[3]. We maintain a mailing list of past events we organized to reach out to them. The rising interest of the NeurIPS community of meta-learning should also attract people new to this problem, who want to learn about it. To attract participants from various under-represented communities in machine learning challenges, we intend to organize one or several **on-line bootcamps, with introduction webinars and hand-on working sessions**. Form experience we have in organizing such events, we intend to reach out to the communities of NeurIPS diversity and inclusion groups, and other groups with which we already have ties, such as NewInML, Data Science Africa, and Women in Data Science, as well as include participants recruited from contacts we have in India (Birla Institute of Technology and Science, Pilani), South America (INAOE, Mexico) and Thailand (KKU and AIT).

1.2 Novelty

The competition is part of an established series of competitions. Table 1 lists several of those. There is overlap between the organizing committee of these past competitions and the present proposal.

The proposed challenge would be our second meta-learning challenge organized around the concept of few shot learning, following 2020 MetaDL-mini, which was a “trial run”, to test the design, the interface, evaluation mechanisms, and implementation of the en-

[3] `https://metalearning.chalearn.org/`

vironment. We ran it in conjunction with a workshop accepted for AAAI 2021 (see `https://metalearning.chalearn.org/`).

The 2020 MetaDL-mini challenge included only three datasets, 1 public dataset, 1 feedback dataset and 1 hidden dataset for final evaluation, upon which the participants were ranked. All datasets were already known to the public, but were obfuscated to avoid that they would be easily recognized. Furthermore, the challenge was run with code submission in the feed-back and final phases (blind testing on the challenge platform, without revealing the datasets to participants), to foster real automated machine learning. The datasets were limited to small images (28x28 or 32x32 pixels), covering hundreds of classes. This first challenge was a good means of ironing out our protocol, but we use only three well-known "toy" datasets, which presented important limitations:

1. The dimension and simplicity of images is unrealistic compared to practical applications.

2. By evaluating on just a single dataset in the final phase, the strengths and weaknesses of various techniques were not extensively tested.

3. Expert users could unintentionally or willingly capitalize on previous exposure to the datasets, which were used in the field as benchmarks.

In the proposed NeurIPS 2021 MetaDL competition, we build upon the infrastructure of the previous 2020 challenge, but as our main contribution we pledge to provide 5 fresh datasets, for a fair evaluation of few shot learning in the final phase. In addition, we will use in the feed-back phase 5 known datasets, but not previously used in meta-learning benchmarks, and format over 10 known meta-learning benchmark datasets as "public data" for use with our starting kit. After the challenge, these datasets will be made available for the community to further benchmark new techniques.

Other that that, we are only aware of another previous challenge in on-shot-learning for gesture recognition in videos [4] that some of us co-organized in 2011-2012[4], targeting a different community of computer vision specialists.

1.3 Data

We obtained access to a large number of new datasets, currently unavailable to the community. One of our main contributions in organizing the proposed NeurIPS 2021 MetaDL challenge will be to format these datasets for use in a few shot learning setting. As in previous challenges we organized, the competition will span three phases: (1) a **public phase** (with no submission on the platform, but the release of a starting kit and "public datasets" including those of the 2020 MetaDl-mini challenge), (2) a **feedback phase** (in which participants develop their methods and get immediate performance feed-back on 5

[4] https://gesture.chalearn.org/2011-one-shot-learning

Table 2: **Datasets** that we envision to introduce, and for which we have data and labels. Some of the datasets are still under investigation, and not all meta-data could be filled in. The five first are the "freshest" and will be used for the final test phase; the five last ones will be used in the final phase. The datasets are roughly paired to feature similar domains in the feed-back and final phase.

FINAL TEST PHASE		
	Task	Description
1. Ecology	Insect Classification	257,056 observations, hierarchical classes, 128x128 images
2. Medicine	Skin disease classif.	$\sim 20,000$ images; ~ 3000 classes; various sources and sizes
3. Pharmacology	Toxic activity	Tabular data, from molecules
4. Biology	fMRI classification	Various sources; various size images
5. OCR	Character Recognition	Printed char., synthetic data, 689 classes, $\geq$ 32x32 images
FEED-BACK PHASE		
Domain	Task	Description
1. Ecology	Plankton Classification	$\sim$ 10,000 observations, $\pm$100 classes, various size images
2. Medicine	Skin disease classif.	$\sim$ 6000 images; $\sim$ 198 classes, various sources and sizes
3. Pharmacology	Chemical Activity	QSAR features from unstructured data
4. Biology	fMRI classification	Various sources; various size images
5. OCR	Character Recognition	50 alphabets; 1623 characters; 50 writers (Omniglot)

datasets hidden on the platform, by submitting their code), and (3) a **final phase** (in which the last method of each participant submitted in the previous phase is evaluated on 5 fresh final datasets).

We will define pairs of datasets relatively similar in terms of task and characterization, and provide one of each in the feedback and final test phase. We will keep those that are really fresh and novel for the final test phase and allow ourselves to recycle datasets previously used in machine learning for the feed-back phase, as long as they were not previously used by the meta-learning community as a benchmark.

In order to exploit the full potential of our data, we will make use of two types of datasets:

- Images: each example will be presented to the participants as tensors (width, height, color channels). Participants can apply relatively large convolutional neural networks on these.
- Unstructured data: each example will be presented to the participants as feature vectors. Participants cannot assume apriori dependencies between certain features (these needs to be learned), and as such convolutional neural networks are discouraged.

The participants will be informed on the type of dataset modality and format such that an adequate architecture can be chosen.

Table 2 shows datasets, that we have access to. Many of these datasets have not been used in the machine learning community, several of those are publicly available, but have not been used by the meta-learning community (e.g., Plankton). The OmniGlot dataset is the only dataset that is commonly used in the meta-learning community; as such this dataset will be used in the feedback phase, as it will have not influence on the final score. For all datasets, we are able to obtain the license to use the data and publish it after the challenge on OpenML.

We envision to use Plankton, a Skin disease dataset, an FMRI dataset, QSAR and Omniglot in the feedback phase. We envision to use insects, another Skin disease dataset, another FMRI dataset, Toxicology and Omniprint in the final phase.

Figure 1 shows some examples of the insects dataset. The data owner is the French 'National Museum of Natural History' in Paris. As the internal dataset has a taxonomy (hierarchical set) of classes, the main task in composing the dataset is to determine a trade-off between the quantity of classes and the distinction classes. We expect that this dataset will be widely adopted by the community after the challenge.

Other datasets under consideration also come from the practical domains (skin desease, QSAR and predictive toxicology). They are part of the master or PhD these of our students and they are actively working on preparing them.

The only dataset we intend to use having an artificial component is the OmniPrint dataset, where the task is to recognize synthetic printed characters, imitating printed characters "in the wild", with various fonts, styles, perspectives, colors, and backgrounds. Each task would be to train/test on a new alphabet, with one example from one font for training and examples from other fonts for testing, including variations in style, perspective, color, and background. This dataset being under development, it is currently unknown to the machine learning community. As such, we will use it in the test phase. We will pair it in the feed-back phase with a known handwritten character recognition dataset (Omniglot).

1.4 Tasks and application scenarios

All problems will be applied on real-world datasets in a few-shot classification setting (see, e.g., [3, 16, 7]), and details on the task format in the next section. Typically, a dataset is broken up into mini-tasks, each of which includes examples (called "shots") of a few classes (called "ways"). A meta-training set includes examples of such tasks (pairs of training and test sets; both labeled). A meta-test set is them presented to the meta-trained algorithms, including labeled training sets and unlabeled test sets. Additionally, a meta-validation set may be provided for hyper-parameter tuning (see next section).

This setting emulates the situation in which data practitioners want to use machine learning (often deep learning) methods, but do not have access to ample amounts of data, in each given task. Yet as a whole, considering all mini-tasks, there should be enough data to train even a large model. In such scenarios, practitioners can resort to meta-learning i.e., leverage data from similar data sources, to maximize the potential of the machine learning method.

We have made efforts in assembling realistic tasks from practical domains, i.e., biological

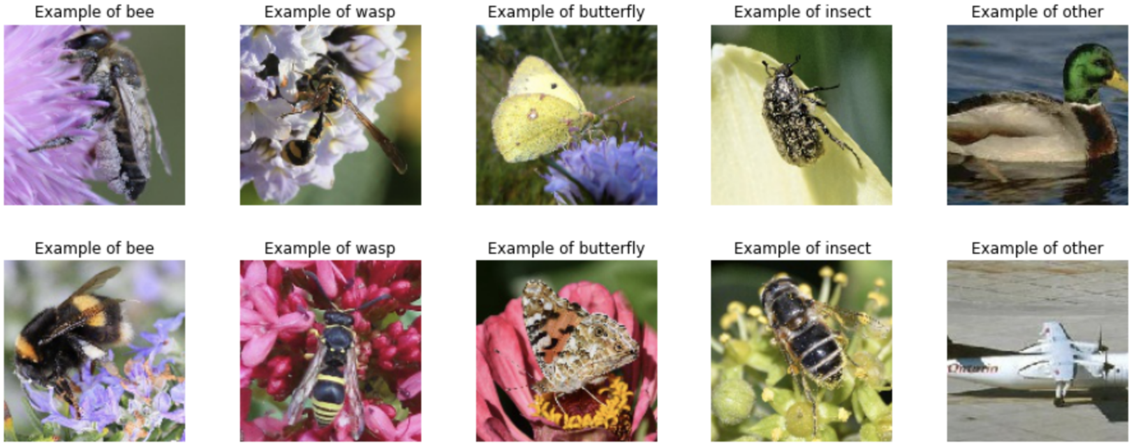

Figure 1: Example images of the Insect Classification dataset.

domain, chemical domain and the medical domain, as described in the previous section. This provides us with realistic scenarios. We are in contact with the respective data owners to dimension the tasks in a realistic way.

For instance, the insect classification problem comes from a crowd-sourcing effort in France of collecting images of insects in the wilderness, some of which are very rare (many classes, or "ways", few examples per class or "shots"). Volunteer amateurs shoot pictures and assign tentative labels. The goal of the project is to assist them in identifying insects and determining whether those are rare species worth further investigation. Many insects of rare species have no more than a few images labeled by experts in the database. We are in touch with the data owners to determinate which level of the insect classification hierarchy we should use to make the outcome of this research most practically useful. If we are successful, we envision that meta-learning solutions to this problems will be embedded in a smart-phone app. The entire methodology combining crowd-sourcing and meta-learning could be deployed in other countries to monitor the health of the insect biosphere.

Another real life scenarios we are using comes from toxicology in bio-medicine. Here, each mini-task does not consist in classifying a subset of the categories of a bigger problem, but each subset corresponds to data collected on the same problem but in different conditions (samples collected in a different hospital or analyses in a different lab, hence prone to different artifacts or biases.

1.5 Evaluation Metric: N-way, k-shot classification

We will use the common N-way, k-shot classification setting (see Figure 2), use recently in the meta-learning literature. This setting includes three stages—meta-train, meta-

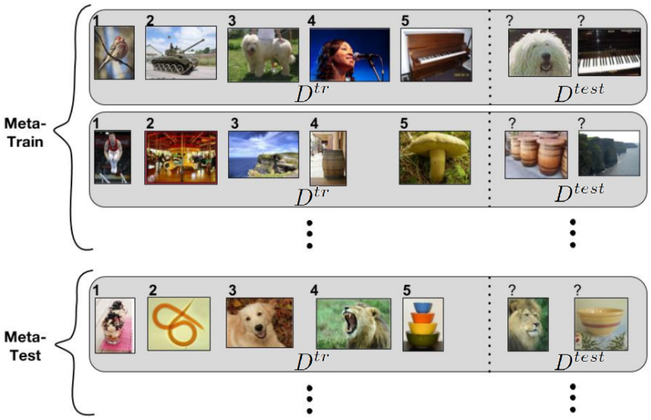

Figure 2: Illustration of N-way, k-shot classification, where $N = 5$, and $k = 1$. This means that we have only one example of each of 5 classes for learning. The test set includes a number of "query" examples, which are labeled in the meta-training set, but unlabeled in the meta-test set. Meta-validation tasks are not displayed. At utilization time, we would get new tasks similar to the Meta-test situation, for which we should be able to make predictions on queries for a new set of classes, after having seen only one example per class. Adapted from [14].

validation, and meta-test—which are used for meta-learning, meta-learner hyperparameter optimization, and evaluation, respectively. Each stage has a corresponding set of disjoint labels. In a given stage, elementary *tasks/episodes* $\mathcal{T}_j = (D^{tr}_{\mathcal{T}_j}, D^{test}_{\mathcal{T}_j})$ are obtained by sampling examples $(\boldsymbol{x}_i, y_i)$ from that task.

This setting assumes that are are carving many tasks out of a mother dataset $\mathcal{D}$ (e.g. one of the 10 datasets of Table 2). For simplification, the construction of tasks is guided by the N-way, k-shot principle, which states that every training data set $D^{tr}_{\mathcal{T}_j}$ should contain exactly N classes and k examples per class, implying that $|D^{tr}_{\mathcal{T}_j}| = N \cdot k$. Furthermore, the true labels of examples in the test set $D^{test}_{\mathcal{T}_j}$ must be present in the train set $D^{tr}_{\mathcal{T}_j}$ of a given task $\mathcal{T}_j$. $D^{tr}_{\mathcal{T}j}$ acts as a support set, literally supporting classification decisions on the query set $D^{test}_{\mathcal{T}_j}$.

The meta-learning objective in the training phase is to minimize the loss function of the model predictions on the query sets, conditioned on the support sets. As such, for a given task $\mathcal{T}_j$, the model 'sees' the support set, and extracts information from the support set to guide its predictions on the query set. By applying this procedure to different episodes/tasks $\mathcal{T}_j$, the model will slowly accumulate meta-knowledge, which can ultimately speed up learning on new tasks. After seeing a pre-defined number of meta-train episodes, the algorithm is presented with the meta-test dataset. This consists of various episodes, where in each episode the algorithm is presented with the train instances of D^{tr} (along with corresponding labels), and are required to make predictions on test instances D^{test}. Participants are evaluated on their **accuracy score on the meta-test dataset**.

N-way, k-shot classification is often performed for small values of k (since we want our models to learn new concepts quickly, i.e., from few examples). In that case, one can refer to it as few shot learning.

For the final phase, we will run the solution of the participants several times, with various divisions of meta-train and meta-test. As this will be done on our servers, the participants will not be able to exploit the earlier obtained knowledge. We will rank the participants based on the highest average score.

1.6 Baselines, code, and material provided

We provide participants with the same starting-package[5] that we used for the trial version (2020 MetaDL-mini). It contains the following relevant code. Note that the baselines are implementations of the skeleton interface, and can be used as example for the participant how to structure their submission.

- Skeleton interface: The code with place-holder methods that the participant should fill in, to make a submission.
- Baseline: A naive approach, that accumulates all data from meta-train and trains a neural network on it. This network is then applied on meta-test.
- Baseline: MAML [3]
- Baseline: Prototypical networks [16]

Additionally, we will provide the participants with datasets in the public phase (that they can use to experiment on their own computer). They will be able to submit provided sample submissions to the feedback phase for practice purposes and them start improving with their own algorithm. The number of submissions will be loosely constrained on the feedback phase (5 per day, 100 in total). However, in the final test phase, a single submission per team will be allowed. In the feed-back phase, scores on the 5 feed-back mother datasets $\mathcal{D}$ will be publicly visible on a leaderboard, but the feed-back datasets themselves will not be accessible to the participants (blind testing of their code). The final evaluation will be carried out in similar conditions on the final test set, but a unique final submission of the participants will be evaluated. The final datasets will also remain inaccessible to the participants (blind testing of their code).

[5] `https://github.com/ebadrian/metadl/tree/master/starting_kit`

1.7 Tutorial and documentation

We provide the participants with the following textual documentation.[6] This documentation features the following information:

- Practical information how to install and setup the environment
- Detailed information about the evaluation process
- Information on how to make a submission
- Troubleshooting (which we aim to extend if we encounter reoccurring problems)

More specifically, we provide the participants with the following coding tutorial:[7]

- Exploring simple properties of the data.
- The coding interface and the functions to implement.
- How to run the method on the datasets.

Since this document and the tutorial are the evolution of the starting kit that was used for the trial run (2020 MetaDL-mini), we are confident that it contains ample information to get the participants on their way. We also have a paper in preparation on the analysis of the 2020 MetaDL-mini challenge, which we will make available for inspiration.

2 Organizational aspects

2.1 Protocol

The competition will consist of three phases:

- **Public Phase**: The public phase starts at the same time as when we start promoting the competition. During the public phase, participants have access to the starting kit, a public (meta-)dataset on which the algorithm can be tested, and a sandbox mimicking our competition environment.
- **Feedback Phase:** This phase is the main mode of the competition. We will make available five datasets within the competition environment, and when participants upload their code, it will be evaluated on these datasets. The participants will get feedback on how their submission performed on these datasets, and can compare it against the submissions of other participants. The participants are encouraged to extensively make use of the feedback phase by trying out many variations of their algorithm. This phase will last 2 months.

[6] `https://github.com/ebadrian/metadl/blob/master/starting_kit/README.md`

[7] `https://github.com/ebadrian/metadl/blob/master/starting_kit/tutorial.ipynb`

- **Final Phase:** Once the feedback phase ends, the last submitted model of each of the participants will be evaluated on the five novel datasets in the final phase. Although performing well on the feedback phase is a good indication, the only performance that will matter for the final ranking is obtained in the final phase.

The participants have the option to run their algorithms locally on the datasets from the public phase. During the feedback phase, the participants have the option to upload their solution to our competition environment. We will run all submissions during the feedback phase and final phase on our own compute resources. The participants will receive in the order of 12 GPU-hours per submission. This way, all participants will be allocated an equal amount of compute resources on the same hardware.

Protocol testing: We will use the established platform CodaLab. Since we have team members of the Codalab platform as co-applicants (Isabelle Guyon, Zhengying Liu) of the competition, we will be able to address CodaLab bugs and issues efficiently. In order to test our competition protocol, we ran a trial version of the competition late 2020 (see Section 1.2). During these trial runs, we were able to optimize the provided resources (e.g., starting kit) and competition protocol.

Novelty: The main novelty in this challenge will be the introduction of ten new datasets for few shot learning, which are significantly larger in input dimension, number of classes, and number of examples, compared to the 2020 MetaDL-mini challenge. These new datasets will allow us to test the **scalability** of algorithms and there **universality** to tackle problems from various input domains, two of the important questions we want to address in this competition. Additionally, we are presently improving the sample code provided in the starting kit to make it more **modular**, and provide a **clear and simple API**. This API will serarate well functionalities such as data management (episodic learning), hyper-parameter selection, model evaluation, and ensembling solutions. This should facilitate performing systematic experiments in post-challenge studies. To that end, we will organize a post-challenge "coopetition", in which challenge participants, who wish to contribute to our analysis paper, will have to produce part of the results of the systematic study. One aspect of particular interest is the **influence of pre-training** externally to the challenge platform, e.g. on massive datasets such as Imagenet. We are contemplating the possibility of having two separate tracks one with pretrained models and another without. Another possibility would be to relegate this comparison to post challenge analyses and include in the API means of re-initializing the predictive model. A final idea is to test the capability of the submitted solutions to learn from mini-tasks coming from heterogenoeus domains, and generalize to a new one – a form of **domain adaptation**. To that end, we will train on all meta-training sets from all-domains-but-one and test on the remaining one – a form of meta cross-validation.

The applicability of the challenge results largely depends on the realism of the tasks carved out of the datasets we chose. We work closely together with the dataset owners to ensure that we use these datasets in realistic scenarios (see Section 1.4).

2.2 Rules

Draft of the rules:

- **General Terms**: This challenge is governed by the General ChaLearn Contest Rule Terms, the Codalab Terms and Conditions, and the specific rules set forth.

- **Announcements**: To receive announcements and be informed of any change in rules, the participants must provide a valid email.

- **Conditions of participation**: Participation requires complying with the rules of the challenge. Prize eligibility is restricted by US government export regulations, see the General ChaLearn Contest Rule Terms. The organizers, sponsors, their students, close family members (parents, sibling, spouse or children) and household members, as well as any person having had access to the truth values or to any information about the data or the challenge design giving him (or her) an unfair advantage, are excluded from participation. A disqualified person may submit one or several entries in the challenge and request to have them evaluated, provided that they notify the organizers of their conflict of interest. If a disqualified person submits an entry, this entry will not be part of the final ranking and does not qualify for prizes. The participants should be aware that ChaLearn and the organizers reserve the right to evaluate for scientific purposes any entry made in the challenge, whether or not it qualifies for prizes.

- **Dissemination**: The challenge is part of the official selection of the NeurIPS 2021 conference. There will be publication opportunities for competition reports co-authored by organizers and participants.

- **Registration**: The participants must register to Codalab and provide a valid email address. Teams must register only once and provide a group email, which is forwarded to all team members. Teams or solo participants registering multiple times to gain an advantage in the competition may be disqualified.

- **Anonymity**: The participants who do not present their results at the conference can elect to remain anonymous by using a pseudonym. Their results will be published on the leaderboard under that pseudonym, and their real name will remain confidential. However, the participants must disclose their real identity to the organizers to claim any prize they might win. See our privacy policy for details.

- **Submission method**: The results must be submitted through this CodaLab competition site. The number of submissions per day and maximum total computational time are restrained and subject to change, according to the number of participants. Using multiple accounts to increase the number of submissions in NOT permitted. In case of problem, send email to metalearningchallenge@googlegroups.com. The entries must be formatted as specified on the Instructions page.

- **Reproducibility**: The participant should make efforts to guarantee the reproducibility of their method (for example by fixing all random seeds involved). In the Final Phase, all submissions will be run three times, and the worst performance will be used for final ranking.

- **Prizes**: The three top ranking participants in the Final phase blind testing may qualify for prizes. The last valid submission in Feedback Phase will be automatically submitted to the Final Phase for final evaluation. The participant must fill out a fact sheet (TBA) briefly describing their methods. There is no other publication requirement. The winners will be required to make their code publicly available under an OSI-approved license such as, for instance, Apache 2.0, MIT or BSD-like license, if they accept their prize, within a week of the deadline for submitting the final results. Entries exceeding the time budget will not qualify for prizes. In case of a tie, the prize will go to the participant who submitted his/her entry first. Non winners or entrants who decline their prize retain all their rights on their entries and are not obliged to publicly release their code.

Discussion: The competition centers around the research questions described in Section 1.1. The competition should serve as a community-wide benchmark, giving insights in how state-of-the-art techniques perform in newly assembled datasets from the practical domain.

The rules have been designed with the criteria of *inclusiveness to all participants* and *openness of results* in mind.

We aim to achieve inclusiveness to all participants by allowing participants to enter anonymously, and allowing them cycles of computation (for the feedback phase and final phase) on our compute resources. This way, participants that do not have ample compute resources will not be limited by this and have a fair chance to win the challenge.

Additionally, we envision to host one or several online hackathons. By means of the hackathons, we ensure that participants are aware of the meta-learning techniques and will enable them to understand the techniques and the structure of the starting kit. Having such hackathons will also implicitly encourage them to participate to the challenge. To ensure participation to those hackathons, we will reach out to several communities including NeurIPS diversity and inclusion groups, and other groups with which we already have ties, such as NewInML, Data Science Africa, and Women in Data Science. We will also recruit from several institutions that our organizers are collaborating with (e.g., Asian Institute of Technology, Khon Kean University, Indian Birla Institute of Technology and Science, Pilani, South American INAOE, Mexico, and Thai KKU and AIT.

We aim to achieve openness of results by requiring all participants to upload their code base, and afterwards fill in a fact sheet about the used methods. This allows us to do a post-challenge analyses about which method generally works better, and which components of the methods perform well. The new API of our starting kit will allow us to take this one step further by organizing a post-challenge “coopetition” in which the participants

Table 3: Envisioned schedule

Date	Phase	Description
February 2020 - September 2020	Preparation	Challenge design, various hackathons
October 2020 - December 2020	Preparation	Trial run of protocol
January 2021 - March 2021	Preparation	Reflection on trial run, collection of new datasets
April 2021 - July 2021	Preparation	Setting up environment for new challenge
August 2021	Public Phase	Start of public phase, publicity
September 2021 - October 2021	**Feedback Phase**	**Main mode of competition**
November 2021	Final Phase	Evaluating performance on hidden datasets
December 2021	NeurIPS 2021	Conference

who wish will be engaged in post-challenge systematic experiments in preparation of a collaborative paper. We already have experience with working on such a collaborative paper in the analysis of the NeurIPS 2019 AutoDL challenge.

Cheating Prevention: Since the challenge is ran on our personal compute cluster, with datasets that are currently not available to the community, we have taken reasonable steps to minimize the probability of cheating. We will pro-actively reach out to participants that have a suspicious submission pattern (submissions of various usernames from the same IP-addresses).

2.3 Schedule and readiness

Table 3 shows the envisioned schedule of the competition. We invested already quite some time into the preparation, as such we are well on schedule for running the challenge later this year. We started with the challenge design early 2020, and as such all infrastructure (implementation of the challenge platform and protocol, starting kit, documentation) is already in an advanced stage. We are finalizing improvements to the API of sample submissions in the starting kit.

The bulk of the remaining work is to format the new datasets. We have four master and PhD students who will be diligently working on this in the next few months, starting from the feedback phase data, then the final test phase data. Several datasets were already used in part internal projects and we have some familiarity with them (although they were not used yet for meta-learning).

The most important part of the challenge is the feedback phase, in which participants can iteratively make submissions and get feedback on the performance. According to the NeurIPS guidelines, we scheduled this to be finished during October. It is currently scheduled to start from September onward, to not collide with traditional holiday months.

After the final deadline of the feedback phase (October 2021), we will use the remaining time to run the submissions on the datasets from the final phase. The winner will be announced during NeurIPS 2021. After NeurIPS 2021 we will run the post-challenge analyses and write our collaborative paper.

2.4 Competition promotion

We will use the following channels of promotion:

- Participants that entered our trial run MetaDL 2020
- Mailing lists on the topic of Meta-learning and AutoML, the AutoML Slack group
- Displacement on the frontpage of OpenML (120,000 unique visitors yearly)
- In-network advertisement, such as personal twitter accounts and personal messages to interested colleagues
- Organization of hackathons/bootcamps.

3 Resources

3.1 Organizing team

- **Adrian El Baz** is a freelance research engineer for Chalearn. He previously did his end-of-studies internship under the supervision of Isabelle Guyon and Zhengying Liu as part of his master program MVA at Université Paris Saclay. He has expertise in Meta-Learning and co-created the AAAI 2021 MetaDL challenge. He will be updating the starting kit.
- **Isabelle Guyon** is Full Professor of Data Science and Machine Learning at Université Paris-Saclay, head master of the CS Artifical Intelligence master program, and researcher at INRIA. She is also founder and president of ChaLearn, a non-profit dedicated to organizing challenges in Machine Learning and community lead on the development of the competition platform CodaLab. She was co-program chair of NeurIPS 2016 and co-general chair of NeurIPS 2017, and now serving on the board of NeurIPS. She is an AMIA and an ELLIS fellow and action editor at JMLR, CiML springer series editor, and BBVA award recipient. She will be advising with challenge design and mentoring her students Hoazhe Sun, Ihsan Ullah, and Phan Anh Vu, which will be formatting several datasets.

- **Zhengying Liu** is a PhD student at Université Paris-Saclay, under the supervision of Isabelle Guyon. His research interests lie in AutoML, deep learning and artificial intelligence in general including logic and automatic mathematical reasoning. He is one of the organizers of AutoDL challenges and has organized corresponding workshops at ECMLPKDD 2019 and NeurIPS 2019. He will be helping Adrian El Baz in preparing the starting kit.

- **Jan N. van Rijn** is Assistant Professor at Leiden University (NL), and has expertise in Automated Machine Learning. He is one of the co-founders of OpenML, an open source experiment database for Machine Learning and meta-learning research, and one of the authors of the 2nd edition of the book Metalearning (to appear in 2020). Moreover, he has organized the Algorithm Selection Competition 2015–2017 [12]. He and his students will be working on the pharmacology datasets.

- **Sebastien Treguer** is an independant researcher and a deep learning instructor at Humancoders with a background in signal and image processing. He seats at the board of directors of Chalearn. He has been involved in the organization of AutoML and AutoDL challenges, and associated workshops at NeurIPS. His research interests lie in developing approaches able to learn and combine various level of abstraction from different modalities and building bridges with neuroscience. He will be working on the FMRI datasets.

- **Joaquin Vanschoren** is Assistant Professor at Technical University Eindhoven (NL). He is co-author of the book 'Automated Machine Learning: methods, systems and challenges' [9] and is therefore a leading person from the Automated Machine Learning and meta-learning community. Moreover, he is one of the co-founders of OpenML and one of the organizers of the running workshop series on Meta-Learning at NeurIPS. He will be advising on the challenge design and contributing to post-challenge analyses.

- **Haozhe Sun**, **Ihsan Ullah**, and **Phan Anh Vu** are students of Isabelle Guyon. The OCR dataset, medical datasets, and ecology datasets are material they are using in the preparation of their theses, for which they have ample experience. They are in contact with the data original owners.

3.2 Resources provided by organizers, including prizes

We aim to provide the following resources:

- **Competition infrastructure:** We will use CodaLab as the main platform for the competition. With Isabelle Guyon and Zhengying Liu, we have two persons who are directly involved with the platform, and will be able to make infrastructural changes and bug fixes if necessary.

- **Non-commercial datasets:** Our main contribution will be to provide 10 non-commercial datasets, most of these will be new to the community.
- **Computing Resources:** Through our sponsors[8], we should have access the computing resources necessary to run all online phases of the competition (feedback phase, final phase) on a homogeneous cluster. Concretely speaking, we expect to have access to 100,000 credits available for computation, which is the equivalent of approximately 91575 GPU hours on a Tesla M60 GPU. Additionally, we have access to our institutional computation clusters[9].
- **Prize money:** We prefer to invest our resources in the quality of the competition, as winning a NeurIPS competition should be plenty of incentive to participate in the competition. However, we have access to some funding to be able to award at least 1000 USD in prize money, spread over the top-three finishers. We are committed to further extend this, if sponsors present themselves.
- **Travel award:** Since NeurIPS 2021 will be a virtual only conference, we will be able to award the top-three finishing teams with a sponsored NeurIPS conference ticket.

3.3 Support requested

We will take the main responsibility for the publicity of our competition, ensuring plenty of participants from the NeurIPS community. To support us in this publicity, we count on the NeurIPS organization for the following matters (all of which are already suggested by the call for competitions):

- Displayment of the competition on the NeurIPS 2021 website.
- Referring participants to the various competitions that are organized.
- A time slot during the program, among which we can announce the winner and discuss the setup.

References

[1] A. Biedenkapp, M. Lindauer, K. Eggensperger, C. Fawcett, H. H. Hoos, and F. Hutter. Efficient parameter importance analysis via ablation with surrogates. In *Proc. of AAAI 2017*, pages 773–779. AAAI Press, 2017.

[2] P. Brazdil, C. Giraud-Carrier, C. Soares, and R. Vilalta. *Metalearning: Applications to Data Mining.* Springer Publishing Company, Incorporated, 1 edition, 2008.

[8] We had this year grants from Microsoft Azure and Google cloud to run our challenges, which we are optimistic will be renewed, if this challenge is accepted.

[9] `https://www.universiteitleiden.nl/en/research/research-facilities/alice-leiden-computer-cluster`

[3] 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.

[4] Isabelle Guyon, V. Athitsos, P. Jangyodsuk, H. J. Escalante, and B. Hamner. Results and analysis of the chalearn gesture challenge 2012. In *Revised Selected and Invited Papers of the International Workshop on Advances in Depth Image Analysis and Applications - Volume 7854*, page 186–204, Berlin, Heidelberg, 2012. Springer-Verlag.

[5] 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.

[6] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

[7] Mike Huisman, Jan N van Rijn, and Aske Plaat. A survey of deep meta-learning. *arXiv preprint arXiv:2010.03522*, 2020.

[8] F. Hutter, H. H. Hoos, and K. Leyton-Brown. An efficient approach for assessing hyperparameter importance. In *Proc. of ICML 2014*, pages 754–762, 2014.

[9] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.

[10] A. Klein, S. Falkner, S. Bartels, P. Hennig, and F. Hutter. Fast Bayesian optimization of machine learning hyperparameters on large datasets. In *Proc. of AISTATS 2017*, volume 54, pages 528–536. PMLR, 2017.

[11] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: Bandit-Based Configuration Evaluation for Hyperparameter Optimization. In *Proc. of ICLR 2017*, 2017.

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

[13] 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 Hugo Jair Escalante and Raia Hadsell, editors, *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, volume 123 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 08–14 Dec 2020.

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

[15] 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'18, 2018.

[16] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.

[17] J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems 25*, pages 2951–2959. ACM, 2012.

[18] K. Swersky, J. Snoek, and R. Adams. Multi-task Bayesian optimization. In *Advances in Neural Information Processing Systems 26*, pages 2004–2012, 2013.

[19] Joaquin Vanschoren. Meta-learning. In Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren, editors, *Automatic Machine Learning: Methods, Systems, Challenges*, chapter 2, pages 39–68. Springer, 2019.

[20] Stéphane Pateux Yuqing Hu, Vincent Gripon. Leveraging the Feature Distribution in Transfer-based Few-Shot Learning. *arXiv arXiv:2006.03806v3*, 2020.