Towards Robust Episodic Meta-Learning

Beyza Ermis[1] **Giovanni Zappella**[1] **Cédric Archambeau**[1]

[1]Amazon Web Services, Berlin, Germany

Abstract

Meta-learning learns across historical tasks with the goal to discover a representation from which it is easy to adapt to unseen tasks. Episodic meta-learning attempts to simulate a realistic setting by generating a set of small artificial tasks from a larger set of training tasks for meta-training and proceeds in a similar fashion for meta-testing. However, this (meta-)learning paradigm has recently been shown to be brittle, suggesting that the inductive bias encoded in the learned representations is inadequate. In this work we propose to compose episodes to robustify meta-learning in the few-shot setting in order to learn more efficiently and to generalize better to new tasks. We make use of active learning scoring rules to select the data to be included in the episodes. We assume that the meta-learner is given new tasks at random, but the data associated to the tasks can be selected from a larger pool of unlabeled data, and investigate where active learning can boost the performance of episodic meta-learning. We show that instead of selecting samples at random, it is better to select samples in an active manner especially in settings with out-of-distribution and class-imbalanced tasks. We evaluate our method with Prototypical Networks, foMAML and protoMAML, reporting significant improvements on public benchmarks.

1 INTRODUCTION

Meta-learning attempts to generalize well on unseen tasks by using only a modest amount of observations. The aim is, both, to learn a representation and adapt it to unseen related tasks in regimes where only a few data points are available [Finn et al., 2017]. The recent literature refers to this setting as "few-shot learning" and it is considered as the primary setting to evaluate meta-learning algorithms [Santoro et al., 2016, Finn et al., 2017, 2018, Ren et al., 2018a], even if meta-models can be defined over a broader variety of learning problems [e.g., Bertinetto et al., 2019].

When the set of tasks is heterogeneous, meta-learning algorithms that leverage only a few data points can be brittle: recent results suggest existing approaches fail to adapt to new tasks, even though they are designed to do so [Zou and Feng, 2019, Iwata and Kumagai, 2020]. In practice, having a relatively small number of labeled data per task originates in the cost of acquiring the labels. This issue is exacerbated when the number of tasks increases. In this work we show that using a guided strategy to select the data composing the tasks can significantly improve the results on the tasks at hand by ensuring, both, a better representation and a better adaptation for the meta-learned model. This procedure also improve the robustness of meta-learning algorithms in presence of out-of-distribution tasks.

Recent studies in few-shot meta-learning order the data and the tasks to achieve faster convergence [Harwood et al., 2017, Weinshall et al., 2018, Sun et al., 2019], re-weight outliers and poorly labeled data to improve the robustness [Ren et al., 2018b, Killamsetty et al., 2020, Mazumder et al., 2021] or consider an active learning policy to label the data across related tasks with the use of reinforcement learning [Bachman et al., 2017, Konyushkova et al., 2017, Pang et al., 2018]. However, actively composing good episodes for training a meta-learner has received modest attention so far. Active learning (AL) [Cohn et al., 1996] improves the data-efficiency in supervised learning by minimizing the amount of data that needs to be labeled. Instead of a priori collecting and labelling a large dataset, which often comes at a significant expense, labels are only acquired iteratively for the data that are deemed most informative. To address the matter of composing better episodes for few-shot learning, we propose to leverage small-budget AL.

We show empirically that choosing more informative examples can greatly improve the model performance in the few-shot setting, where labeled datasets are small. We use

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

these techniques for generating the context and the query sets at meta-training and meta-testing time (in different combinations). Our goal is to investigate if we can improve meta-training by actively drawing query and context sets during meta-training and to evaluate the impact of actively drawing the context set on the speed of adaptation during meta-test. The best results are obtained with a combination of procedures performed at meta-train and meta-test time, but we also show that the results improve even if the selection is only performed for the adaptation at meta-test time. A robust meta-learner, once trained, should be able to take new and different data from a test task, extract task-specific knowledge from the context set, and generalize well on the query set. The meta-learner adapt more easily to the new tasks when informative data are included in the context set. Active composition of the context set is useful especially when the train and test tasks are diverse.

We apply our methodology to Prototypical Networks [Snell et al., 2017], foMAML [Finn et al., 2017], and protoMAML [Triantafillou et al., 2019], showing that carefully selecting the data in episodic meta-learning can provide a significant benefit for, both, metric-based and optimization-based meta-learning. We run extensive experiments on in- and out-of-distribution settings and evaluate robustness to class imbalance. We theoretically analyze the computational cost of our method that has linear complexity in selecting each data point for episode composition and the empirical results confirm that analysis. The experiments show that our approach for context and/or query set construction consistently improves the accuracy of few-shot classification compared to uniformly random chosen sets.

2 EPISODIC META-LEARNING

Our goal is to learn new tasks from small amounts of data as effectively as possible. The paradigm of choice is meta-learning [see e.g., Sachin and Hugo, 2017, Finn et al., 2017], where a meta-model is optimized on a set of episodes. Each episode is composed of a *context data set*, used for model-specialization and mimicking the small datasets used for adaptation at (meta-)test time, and a *query data set*, used to compute and optimize a training loss given the specialized model. An episode τ is called N-way and K-shot when it has a context set $\mathscr{C}_\tau = \{C_{\tau,n}\}_{n=1}^{N}$ with $C_{\tau,n} = \{(x_i, y_i) \sim \mathscr{D}, y_i = n\}_{i=1}^{K}$. Hence, the episodes are composed of K labeled examples for each of the N classes. An episode has also a query set, which is generated as a fixed-shot set with examples from the classes in the context set; N-way, K-shot query sets are denoted by $\mathscr{Q}_\tau = \{Q_{\tau,n}\}_{n=1}^{N}$ with $Q_{\tau,n} = \{(x'_i, y'_i) \sim \mathscr{D} \setminus \mathscr{C}_\tau, y_i = n\}_{i=1}^{K}$.

The above procedure commonly used in the meta-learning literature artificially creates balanced classes, which is rarely verified in practice. Hence, we also consider a more realistic setting where N-way episodes are composed of a total of $T = NK$ data points, but we do not enforce exactly K data points per class. This procedure results in episodes that are more diverse and possibly unbalanced.

3 META-LEARNERS

There are two main approaches to meta-learning: *metric-based* and *optimization-based*. We choose to experiment with one reference method from each family, namely prototypical networks [Snell et al., 2017] and first-order model-agnostic meta-learning (foMAML) [Finn et al., 2017]. We also consider the recently proposed protoMAML [Triantafillou et al., 2019], which can be viewed as a hybrid between prototypical networks and MAML.

3.1 PROTOTYPICAL NETWORKS

Prototypical networks [Snell et al., 2017] is a metric-based meta-learning approach for few-shot classification. A neural network is trained as an embedding function $f_\theta : \mathbb{R}^d \to Z$ mapping from input space to a latent space Z where points of the same class tend to cluster. The embedding function f_θ is a neural network with parameters θ and it is used to compute a prototype for each class, by averaging the embeddings of all points in the support belonging to that class:

$$c_{\tau,n} = \frac{1}{|S_{\tau,n}|} \sum_{i \in S_{\tau,n}} f_\theta(x_i), \tag{1}$$

where $c_{\tau,n}$ is the prototype for the n^{th} class and $S_{\tau,n}$ is the set of indices of data points in the context set $\mathscr{C}_\tau$ with class n. Once prototypes of all classes are obtained, query points are also embedded in the same space, and then classified based on their distances to the prototypes via a *softmax* function. For a new data point $(\tilde{x}, \tilde{y})$ with an embedding $f_\theta(\tilde{x})$, the probability of belonging to class n is computed as follows:

$$P_\theta(\tilde{y} = n \mid \tilde{x}) = \frac{\exp\{-\text{dist}(f_\theta(\tilde{x}), c_{\tau,n})\}}{\sum_{n'=1}^{N} \exp\{-\text{dist}(f_\theta(\tilde{x}), c_{\tau,n'})\}}, \tag{2}$$

where $\text{dist}(\cdot,\cdot)$ is the Euclidean distance.

Meta-training of prototypical networks consists in minimizing the negative log-likelihood computed over the query set $\mathscr{Q}_\tau$, i.e. $\mathscr{L}(\theta; \mathscr{Q}_\tau) = -\sum_i \log P_\theta(y'_i \mid x'_i)$, with respect to θ. Episodes are formed by randomly selecting a subset of classes from the training set, then choosing a subset of the data within each class to act as the context set and a subset of the remainder to serve as query points. During meta-testing, θ is fixed. Both context and query sets are generated in a similar fashion. Query points are now only used to estimate the model performance.

3.2 MAML

Model-agnostic meta-learning (MAML) is an optimization-based approach to meta-learning. The goal of MAML is to

learn a model initialization that is well-suited for few shot learning. Meta-training aims to enable rapid adaptation by learning a suitable representation, such that a good model can be obtained with a few gradient steps at meta-testing.

At each meta-training iteration (i.e., for each episode τ), MAML performs an *inner update* with small number of t gradient steps to compute the adapted model θ_τ:

$$\theta_r = \theta_{r-1} - \alpha \nabla_{\theta_{r-1}} \mathscr{L}(\theta_{r-1}; \mathscr{C}_\tau), \tag{3}$$

where α is a step size and $r \in \{1, \ldots, t\}$. The resulting model $\theta_\tau = \theta_t$, which is a function of $\theta = \theta_0$, is then used to compute the loss on the query set $\mathscr{L}(\theta_\tau(\theta); \mathscr{Q}_\tau)$. The meta-loss is obtained by summing over all episodes and optimized via an *outer update* (or meta-update) via gradient descent with respect to θ. At meta-testing time, we obtain an adapted model by taking t inner update steps starting from θ given $\mathscr{C}_{\tau'}$. The adapted model $\theta_{\tau'}$ is evaluated on the query set $\mathscr{Q}_{\tau'}$.

The optimization of the meta-loss requires differentiating through the t gradient descent updates wrt θ. This involves second order derivatives that are challenging to compute for high-dimensional θ. [Finn et al., 2017] propose a first order approximation of MAML, called foMAML, which omits the second order derivatives and shown to perform well in practice.

3.3 PROTOMAML

ProtoMAML is a hybrid between prototypical networks and MAML proposed by [Triantafillou et al., 2019]. It endows prototypical networks with an adaptation mechanism similar to (fo)MAML and was shown to outperform foMAML by a significant margin on Meta-Dataset.

When considering the Euclidean distance, prototypical networks can be re-parameterized as a linear classifier applied to a learned representation f_θ. ProtoMAML computes the class prototypes $c_{\tau,n}$ from the context set $\mathscr{C}_\tau$ to form the corresponding linear classifier, which is then optimized wrt θ in the same manner as (fo)MAML. We refer to [Triantafillou et al., 2019] for details.

4 ACTIVE EPISODIC META-LEARNING

To robustify episodic training in the few-shot setting we propose to use active learning scoring rules to construct episodes. The approach is agnostic to the meta-learner and, as shown in the experiments, can boost the performance of, both, metric- and optimization-based approaches.

Active Learning (AL) is a learning paradigm that attempts to reduce the cost of human intensive labeling processes. The promise of AL is to use a trained classifier to decide which unlabeled data are best to label in order to improve the classifier the most. The majority of popular approaches are based on heuristics, such as choosing the data whose label the model is most uncertain about, choosing the data whose addition will cause the model to be least uncertain when predicting the label of other data, or choosing the data that is most "different" compared to other unlabeled data according to some similarity function [Seung et al., 1992, Joshi et al., 2009, Sener and Savarese, 2017].

One of the most popular AL strategies is *uncertainty sampling*. This simple, yet effective approach selects the data the classifier is the most uncertain how to label [Lewis and Gale, 1994], which is captured by the entropy $\mathbb{H}(\cdot)$ of the classifier:

$$x_* = \underset{x \in \mathscr{X}_u}{\arg\max}\ \mathbb{H}\left(P_\theta(y|x)\right), \tag{4}$$

where $\mathscr{X}_u$ is the set of unlabeled data samples, $P_\theta(y|x)$ is the likelihood function, that is, the class probability produced by classifier θ at x. For binary classification, applying (4) corresponds to picking x such that $P_\theta(y=1|x)$ is closest to 0.5.

We employ active learning in our episodic training procedure to create "better" episodes helping the meta-learner to learn quickly. In our procedure, an episode τ is formed by first randomly selecting a *candidate set* $\mathscr{E}_\tau = \{E_{\tau,n}\}_{n=1}^N$ where $E_{\tau,n} = \{(x_i, y_i) \sim \mathscr{D}\}_{i=1}^M$ and M is the size of the candidate set. In a real-world scenario the candidate set is made of the unlabeled data that the a user is providing to a machine learning system. If the candidates in $\mathscr{E}_\tau$ have unknown targets, we need an oracle $\mathscr{O}_\tau$ such that $\mathscr{O}_\tau(x) = y$, $\forall x$. This is just an abstraction hiding information manually provided by the labelers. The data points in the candidate set will be labeled sequentially in $r_\tau > 0$ rounds. The context set $\mathscr{C}_\tau$ and/or a query set $\mathscr{Q}_\tau$ are then constructed by performing r_τ queries on the candidate set $\mathscr{E}_\tau$. The queried points are selected by a policy, which in our experiments is uniform random sampling or a scoring based on the uncertainty sampling, but our methodology is applicable with any policy, whether pre-defined or learned offline (e.g., via RL).

To accomplish robust and quick meta-model adaptation, we replace the random generation of the context set $\mathscr{C}_\tau$ by a scoring based generation procedure. We consider N-way, T-size (where $T \leq M$) context sets $\mathscr{C}_\tau \subset \mathscr{E}_\tau$. Therefore, we use $r_\tau = T$ rounds of data point selection. The meta-model is updated after each round in order to select the most informative examples and improve the performance of episodic training. Algorithm 1 summarizes the procedure.

Model adaptation is very important during meta-testing, especially when the the training tasks are distinct from the testing tasks. Indeed, selecting the informative context samples to use can substantially accelerate the adaptation at meta-test. However, selecting the context sets at meta-test might not be always possible. We consider the scenarios where we are allowed to select the context sets at meta-test time. For each meta-learner defined in Section 3, the context

Algorithm 1: Active context set selection for a train episode τ. N_T is the total number of classes in the training set, $N \leq N_T$ is the number of classes p er episode, M is the number of examples in the candidate set, T is the number of examples chosen by score function.

Input : $\mathscr{D}, N_T, N, T, M$
Output : Context set $\mathscr{C}_\tau$

$\mathscr{E}_\tau \leftarrow$ RANDOMSAMPLE$(\mathscr{D}, M, N)$. [1] ▷ Select candidate examples

$\mathscr{C}_\tau \leftarrow \emptyset$ ▷ Initialize the context set
for $i \leftarrow 1$ ***to*** T **do**
  foreach $(x_i, y_i) \in \mathscr{E}_\tau$ **do**
    Compute the uncertainty score with the current classifier and using Equation 4
  end
  Select sample (x_i) from $\mathscr{E}_\tau$ based on the score, ask oracle $\mathscr{O}_\tau$ for labeling (if the sample has unknown target), and add it to $\mathscr{C}_\tau$
  $\mathscr{E}_\tau \leftarrow \mathscr{E}_\tau \setminus (x_i)$
  $\mathscr{C}_\tau \leftarrow \mathscr{C}_\tau \cup (x_i, y_i)$
end

sets are used in the same way as in the meta-training. Hence, we apply the score based selection procedure defined in Algorithm 1 to compose context sets in meta-test.

To learn a better inductive bias encoded in the meta model f_θ, we propose to generate query sets $\mathscr{Q}_\tau$ via scoring based selection during meta-training. Note that using uncertainty scores for query set selection during meta-testing would likely result in a conservative estimation of the meta-model performance as more difficult examples would be used to evaluate it. We compose N-way T-size query sets for each episode τ where $\mathscr{Q}_\tau \subset \mathscr{E}_\tau \setminus \mathscr{C}_\tau$. Unlike $\mathscr{C}_\tau$, the scoring based selection of $\mathscr{Q}_\tau$ is done without updating the meta-model after the selection of each candidate because those labels are not revealed to the learning algorithm. Instead, we select the top-T samples based on the uncertainty scores. Diversity can be ensured by re-sampling $\mathscr{E}_\tau$ in each round [see "59" trick in Smola and Schölkopf, 2000].

5 RELATED WORK

A significant amount of work has been done to improve the performance of neural networks by the mean of intelligent training strategies. Curriculum learning was proposed by Bengio et al. [2009] and is popular for multi-task learning [Pentina et al., 2015, Sarafianos et al., 2017, Weinshall et al., 2018]. In these works, it is shown that it is preferable to introduce an ordering in the tasks, typically from simpler to more complex, to accelerate convergence and possibly improve the generalization. Shrivastava *et al.* [Shrivastava et al., 2016] proposed hard sample mining, which was later used to train with more confusing data achieving higher robustness and a better performance [Canévet and Fleuret, 2016, Harwood et al., 2017]. [Sun et al., 2019] introduce a method to schedule hard tasks in meta-training batches. Our work is different as we are not attempting to order tasks, but propose a method to better compose (small) tasks that can be observed in a random order.

Another related area of research is the robust meta-learning, which attempts to learn out-of-distribution tasks and with noisy data/labels. In this context, a number of recent works introduce robust few-shot learning approaches that reweight the data during meta-learning [Ren et al., 2018b, Lu et al., 2020, Killamsetty et al., 2020, Mazumder et al., 2021]. In particular, [Yoon et al., 2018] proposes a Bayesian method that is robust to overfitting and evaluated their method in active learning in addition to image classification and reinforcement learning. These methods use all data points in tasks and learns to assign weights to these data points. Likewise, our method is able to deal with out-of-distribution tasks, but by choosing informative data points to compose (small) tasks, which result in robuster and data-efficient meta-learners.

Combinations of meta-learning and active learning have also been considered with the goal to learn an AL strategy. Most approaches make use of reinforcement learning to learn an active learning policy that selects the best set of unlabeled items to label across related learning problems [Bachman et al., 2017, Woodward and Finn, 2017, Konyushkova et al., 2017, Pang et al., 2018]. The idea is to define the meta-learner over AL strategies, which requires that, both, the underlying model and the notion of informativeness to be adapted to new target tasks. Despite their complexity, it has been shown that the performance on AL tasks can be improved via meta-learning [Bachman et al., 2017, Konyushkova et al., 2017, Pang et al., 2018, Ravi and Larochelle, 2018, Requeima et al., 2019]. The aim of our work is different since we employ an active learning strategy to improve meta-learning performance and we do not use meta-learning to learn an active learning strategy.

The most relevant previous work is that of [Al-Shedivat et al., 2021]. The authors focus on a subset of the methods that we consider: it proposes an algorithm for active meta-learning, which gradually acquires labels for selected support points at meta-training time. In our work, we use AL for generating the context, as well as the query sets at meta-training and meta-testing time in different combinations. At high level, we can say that using AL for the context set corresponds to using active learning for the better model-specialization and adaptation while using active learning for the query set corresponds to using active learning for learning better representations. To the best of our knowledge we are the first to provide an interpretation of its different

[1] *RANDOMSAMPLE(S, B, K)* denotes a set of *B* elements chosen randomly from set *S* from *K* classes, without replacement.

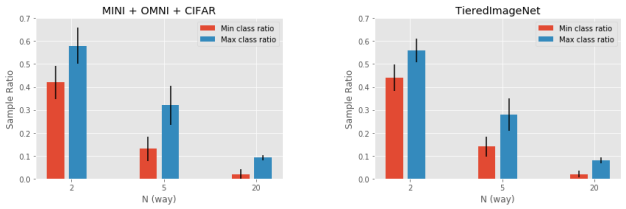

Figure 1: Average minimum and maximum class ratios of candidate sets for TieredImageNet and MINI+OMNI+CIFAR.

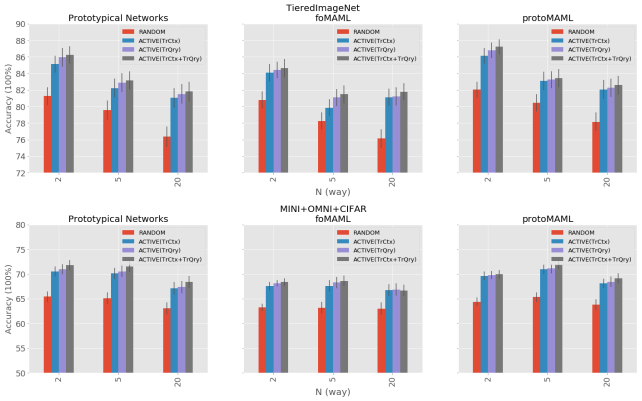

Figure 2: Comparison of random selection (RANDOM) and active selection (ACTIVE) during meta-train for the composition of the query set (TrQry), context set (TrCtx) and their combination (TrCtx+TrQry) for varying N when the episode sizes T are $N \times 10$ on TieredImageNet and MINI+OMNI+CIFAR.

uses.

6 EXPERIMENTS

In this section, we empirically validate our approach and quantify the improvements brought to some widely used meta-learning algorithms. Our experimental evaluation intends to answer the following key questions: Can our approach boost the performance of meta-learning algorithms by helping them learn a better representation? Can our approach accelerate model adaptation by selecting more informative context samples? Does combining the previous two strategies improve the performance further?

Experimental Setup. In order to answer the questions stated above, we create an experimental test-bed matching the standard meta-learning setting. We evaluate several few-shot learning scenarios similar to [Sachin and Hugo, 2017, Snell et al., 2017, Ravi and Larochelle, 2018]. Specifically, we test all the algorithms in a fixed N-way where $N = \{2,5,20\}$ setting. The episode size is $T = N \times K$ for $K = \{2,5,10,20\}$ and the candidate size are fixed to $M = N \times 50$. In order to measure the predictive performance of the methods, we ran each experiment 50 times with different seeds and the results of each run is computed by averaging the classification accuracy over randomly generated episodes selected from the test set. In the following, we report mean and standard deviation computed over the 50 repetitions.

Architecture and Optimization Setup. Following [Snell et al., 2017], we used the same network architecture in all the experiments for learning the embedding. The neural network f_θ is composed of four convolutional blocks where each block comprises a 64-filter 3×3 convolution, batch normalization layer, a ReLU nonlinearity and a 2×2 max-pooling layer. The choice of this architecture was made to keep the experiment as close as possible to previous results and show that our approach does not need significant changes in the setup of the meta-learning algorithms. All of our models were trained via SGD optimizer with a learning rate of 0.05.

Algorithms. The meta-learning algorithms, namely Prototypical Networks, foMAML and protoMAML are run by using the episodic training method described in Section 4. These algorithms were selected as they are canonical representatives of, both, metric-based and optimization-based meta-learning. Note that active meta-learning is by no means tied to any of these methods. To quantify the improvement obtained by creating episodes with active learning (later called ACTIVE), we compare to baselines where the episodes are selected uniformly at random (later called RANDOM). We also employ a variant of RANDOM called RANDOMx2, which gets twice as many data points as random, selected in the same way.

Datasets. We consider several widely-used, well-known datasets: CIFAR100 [Krizhevsky et al., 2009], miniImageNet [Sachin and Hugo, 2017], Omniglot [Lake et al., 2011] and TieredImageNet [Ren et al., 2018a]. CIFAR100 and miniImageNet consist of 100 classes with 600 images per class and Omniglot consists of 50 alphabets. CIFAR100 and miniImageNet are split into separate sets of 64 classes for training, 16 classes for validation, and 20 classes for testing. Omniglot is split into separate sets of 30 classes for training, 10 classes for validation, and 10 classes for testing alphabets. Like miniImagenet, TieredImageNet is a subset of ILSVRC-12 [Russakovsky et al., 2015], with 779165 images representing 608 classes that are hierarchically grouped into 34 categories in total. These are split into 20 training (351 classes), 6 validation (97 classes) and 8 testing (160 classes) categories to ensure that all of the training classes are sufficiently distinct from the testing classes, unlike miniImageNet. This represents a more realistic few-shot learning scenario since in general we cannot assume that test classes will be similar to those seen in training. We note that for all the data sets, the number of data points per class is roughly the same.

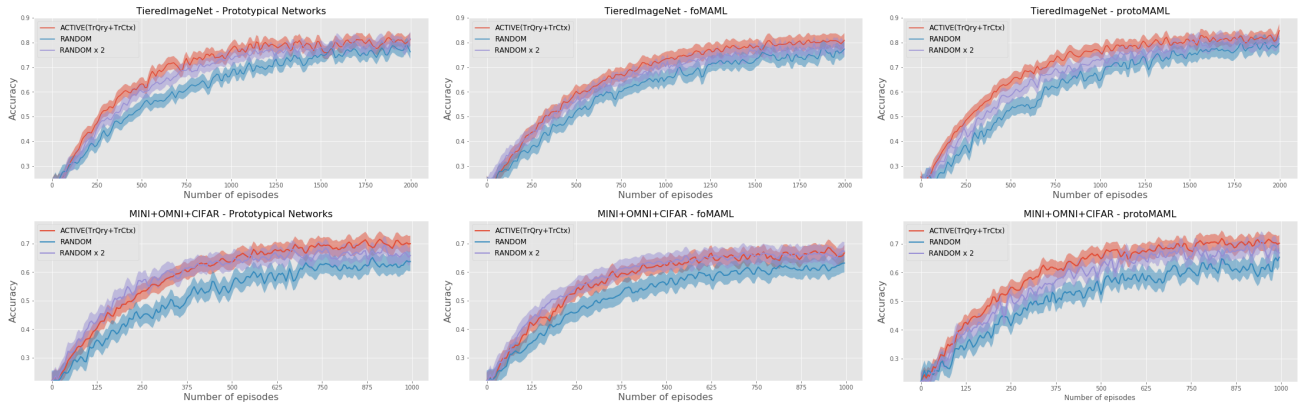

Figure 3: Comparison of random selection (RANDOM) and active selection (ACTIVE) during meta-train for the composition of the query set and context set (TrCtx+TrQry) for $N = 5$ and $T = 50$ on TieredImageNet and MINI+OMNI+CIFAR. (RANDOM x 2) stands for randomly generated episodes when $T = 100$.

For all the experiments, we use disjoint episodes where each episode contains new data points. For every comparison, we ran three different version of the experiment. In the first one, we generated the episodes only from CIFAR100 (results are presented in Appendix A.2 where we use 200 training and 100 test episodes. In the second one, tasks from TieredImageNet are generated as 2000 for meta-training and 1000 for meta-testing. In the last one, we generated tasks from CIFAR100, Omniglot and miniImageNet for both train and test (later called MINI+OMNI+CIFAR). To test the robustness of our approach to the setting with *out-of-distribution* meta-testing tasks (which is addressed in Section 6.2), we generated an additional setting that we call MINI+OMNI+CIFAR (excluded) that only uses tasks from miniImageNet at meta-training and uses tasks from CIFAR100 and Omniglot at meta-test. We use 1000 training and 200 test episodes for these settings.

Sampling Candidate Sets. The create a candidate set, we first sample uniformly at random N class indices from the list of available classes in the dataset. Then, following [Triantafillou et al., 2019] we allow each chosen class to contribute to the candidate set at most 100 of its examples. We multiply this number by a scalar sampled uniformly from the interval (0, 1] to enable the potential generation of "few-shot" episodes. We do enforce, however, that each chosen class has at least one data point in the candidate set, and we select the total candidate set size of M.

Skewness of the episodes. In our experiments we do not employ the stratified sampling approach detailed in Section 2 that is often used in the meta-learning literature for the following reasons: i) it is not realistic to assume to have a perfectly balanced dataset; ii) it is practically impossible to create a perfectly balanced datasets without wasting labels when labels are unknown a priori. To provide a reference point on the unbalancedness of the episode used in the experiments, we compute how the participating classes are distributed over the candidate sets. For each candidate set, we record the number of samples from each participating classes and record the ratio of it to the total number of samples. The average of minimum and maximum class ratios over training and test episodes for $N = \{2, 5, 20\}$ for TieredImageNet and MINI+OMNI+CIFAR setting is reported in Figure 1. For example, the class that has the least samples has 20 samples on average (varies between 2 and 38) while the class that has the most samples has 80 samples on average (varies between 65 and 92) in a candidate set when $N = 20$ for TieredImageNet. In these settings, some of the candidate sets are highly imbalanced especially when $N = 20$. In addition to the results reported in this section, we designed experiments with a balanced N-way K-shot setting and reported the results in Appendix A.4. The results show that the accuracy is higher on the balanced episodes obtained with stratified sampling for both random and active selection, but active selection still outperforms random selection.

6.1 ACTIVE META-TRAINING

In order to understand if we can improve the quality of the learned representations, we run experiments where the query set of each task observed at meta-training is actively composed. Training on more informative data enables the model to achieve higher robustness and better performance [Canévet and Fleuret, 2016, Harwood et al., 2017]. Also intuitively, actively drawing the data points for query set will contain the points on which the model is most uncertain, inducing an higher number of mistakes. This will push most meta-learning algorithms to perform larger updates to their representations. For example, in several meta-learning algorithms, such as variants of MAML and Prototypical Networks, making more mistakes on the query set increases the loss of the current task, which triggers a bigger change during the update of embedding parameters θ.

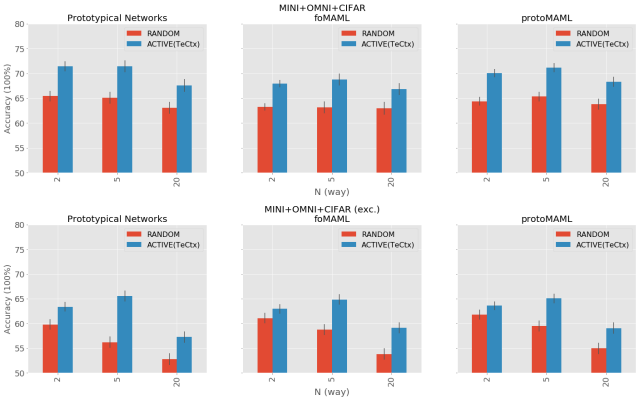

Figure 4: Comparison of random selection (RANDOM) and active selection (ACTIVE) during meta-test for the composition of the context set (TrCtx) for varying N when the episode sizes T are $N \times 10$ on MINI+OMNI+CIFAR and MINI+OMNI+CIFAR (excluded).

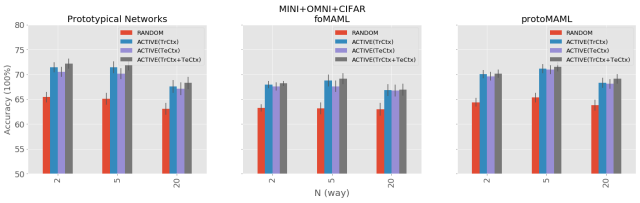

Figure 5: Comparison of random selection (RANDOM) with different combinations of active adaptations (called ACTIVE) at meta-train time (TrCtx) and at meta-test time (TeCtx) for varying N when the episode sizes T are $N \times 10$ on MINI+OMNI+CIFAR.

The results of these experiments for a fixed episode sizes $T = N \times 10$ can be seen in Figure 2. (The results for varying episode sizes are reported in Appendix A.1.) The actively chosen query sets outperforms the random selection for each method and setting. Our results also show that Prototypical Networks and protoMAML benefit more than foMAML. This is consistent with results reported by [Ren et al., 2018a] in different settings and on several datasets. We find that protoMAML remains the best performing method and Prototypical Networks perform almost as well, but both outperform foMAML. These results are not surprising as the performance of AL, and in particular uncertainty sampling, depends on the quality of the classifier.

In this set of experiments, we also test the impact of creating better context sets to see if it can help the algorithms to learn better models. This agrees with the typical usage of active learning in supervised learning. The results reported in Figure 2 show that also in this case active selection helps all the three considered algorithms. The combination of active query and context set construction during meta-training improves the accuracy the most, but active construction of the query set is cheaper and improves the performance more compared to active context set construction. Figure 3 shows the test accuracy by increasing number of training episodes for these selection strategies. The figure shows that for all methods, active selection reaches the same accuracy with random selection by using $\sim 50\%$ less training episodes on TieredImageNet and $\sim 60\%$ less training episodes on MINI+OMNI+CIFAR. The convergence of out approach is even slightly faster than the random selection that creates two times larger episodes.

6.2 ACTIVE META-TESTING

In this set of experiments, we simulate the scenarios where we are allowed to select the context sets at meta-test and show that creating better context sets at test time is useful for the meta-learning algorithms to adapt and perform better on new tasks. This scenario can be possible in machine learning services where a representation is provided, and the users want to leverage this representation to improve the performance on their own task. In such a service, the users can be asked to label "a few" data points, or the service can choose the most informative set of data points among the labeled set of data in the users' tasks.

To test the robustness of our approach to the setting with realistic *out-of-distribution* meta-testing tasks, we generated an additional setting that we call MINI+OMNI+CIFAR (excluded) that only uses tasks from miniImageNet at meta-training and uses tasks from CIFAR100 and Omniglot at meta-test. Figure 4 shows that even in the most complex setup, where we exclude the different datasets in train and test, active context set composition improves the performance of all the three algorithms. The difference between random and active selection is bigger in the setting with out-of-distribution testing tasks. This speaks for the higher ability of the active selection to adapt to the task at hand.

In most of the cases, metric-based meta-learning algorithms get a bigger boost in performance. The main reason of this effect is the better coverage of the space with active selection than with uniformly random selection. For example, we ran an experiment on MINI+OMNI+CIFAR using a single episode and Prototypical Networks with $N = 5$ and $T = 50$. We measured the average Euclidean distance (50 repetitions) between the points in the query set and the closest prototypes. When we apply AL for context set selection, the result is 82.46, when we randomly select the context set, the result is 97.65.

6.3 ACTIVE SELECTION FOR BETTER REPRESENTATIONS *AND* ADAPTATION

Lastly, we would like to verify if the combination of the techniques tested in the previous sections can provide a better overall performance. For this purpose we ran experiments comparing the effect of actively creating query sets at meta-training time, context sets at meta-test time and their combination. In order to see what is the most efficient way

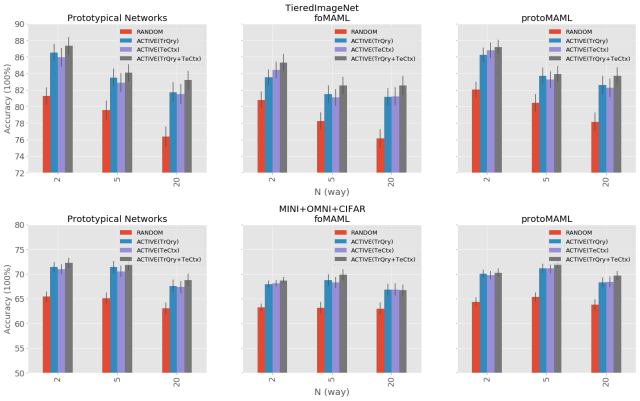

Figure 6: Comparison of random selection (RANDOM) with active selection (called ACTIVE) for the creation of context set at meta-test time (TeCtx), query set at meta-train time (TrQry), and their combination for varying N when the episode sizes T are $N \times 10$.

to combine the techniques, we first compared the efficiency of context selection at train and test.

Is active context set composition more important at meta-training or at meta-testing time? Actively selecting the context set during meta-training and meta-testing has an increased computational cost. Hence, we investigated which of the two procedures gave the biggest advantage and whether the advantage was compounded. We repeated the experiments of the previous section by separating the active composition of the episodes at meta-training and meta-testing time. The results are shown in Figure 5. Once again we see that AL helps meta-learning in both cases. A somewhat surprising result is that the use of AL for adaptation at meta-testing, which is the cheapest and fastest solution, provides a comparable boost in performance to the use of AL for adaptation at meta-training and meta-testing. From a practical point of view, only using AL during meta-testing drastically reduces the cost of employing active episodic learning in scenarios with a large number of training tasks, as it will be discussed in Section 6.4. Finally, note that the usage of AL to create context sets at meta-training time alone has a positive impact, outperforming the random selection too. Additional plots showing that AL can bring advantage when combined with pre-trained models are available in Appendix A.3.

Since, the use of AL for adaptation at meta-testing is more efficient, we combine query selection at train and context selection at test and compare the results in Figure 6. The results show that even if using AL separately at meta-train and meta-test time gives competitive results, the combination of its usage in the two phases gives the best results.

6.4 COST AND COMPLEXITY

Introducing techniques which need to compute additional information, such the uncertainty of the classifier, about different data points comes with an additional computational cost. In particular, there are two different levels of cost associated with the active selection of data points depending on when it is performed. In fact, the update of the estimator (and its uncertainty) is only possible when the context sets are created. When the query set is composed actively, labels are not available and so this operation does not require any update of the estimator.

When composing the context sets, the computational complexity of the operation is $O(TM)$, T is the number of data points selected and M is the size of the candidate set. The values of M are also often small (e.g., $N \times 50$ in our experiments where N is the number of classes) and they do not significantly impact the performance since only a small fraction of the set is selected. Hence, the incurred cost is very limited for the cases we are considering. For example, the best solution observed in our experiments, which combines active query set composition at meta-train and active context set composition at meta-test, for T=50 on average requires 8 minutes and 26 seconds to run end-to-end on an experiment with 1000 training tasks and 200 test tasks. On the same experiment, the uniform random selection requires 6 minutes and 34 seconds, only 112 seconds less. This is the highest difference that we observed in our experiments since we use active selection strategy at both train and test time. We also note that we can make the difference smaller by optimizing the selection and update cycles in the code.

When composing the query sets, the computational complexity of the operation is only $O(M)$. This significantly lowers the complexity derived from the fact that no labels are used in this phase and the classifier is never updated. The selection is performed greedily without any update to the uncertainty values. Scaling up this approach to be able to provide the same kind of advantage on tasks with hundred of thousands or millions of data points could be challenging, but the usage of meta-learning in this setting might be of modest benefit.

7 CONCLUSION AND FUTURE WORK

In this work we provided extensive empirical evidence supporting the use of active episode composition for few-shot classification problems. Our results show that the resulting adaptation is more robust and beneficial even if pre-trained representations are used. Moreover, we show that we can significantly improve the quality of the representation, and thus the performance of the classifiers, by creating query sets on which the meta-learner is more uncertain at meta-train time, an operation which is very cheap to perform as it does not require any model update.

There are at least two aspects that we consider for future

work. The first one is the usage of a batch active learning approach. While sequentially picking the data to label is reasonable on a small scale (e.g., when a user is labelling a few data points for its own task), this could become too demanding when a larger number of data points from different tasks need to be labeled (e.g., the ones in the query set at meta-train time). The use of approaches such as BatchBALD [Kirsch et al., 2019] and BADGE [Ash et al., 2019]) which can request a larger number of data to be labeled in parallel would reduce the operational burden. The second aspect that could further boost the performance is the usage of unlabelled data. Currently, only labelled data are used to train the meta-learner and the information contained in the unlabelled data points is not exploited. The usage of semi-supervised learning (see [Ouali et al., 2020]) in combination with active learning could further boost the efficiency.

References

Maruan Al-Shedivat, Liam Li, Eric Xing, and Ameet Talwalkar. On data efficiency of meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1369–1377. PMLR, 2021.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *CoRR*, abs/1906.03671, 2019. URL `http://arxiv.org/abs/1906.03671`.

Philip Bachman, Alessandro Sordoni, and Adam Trischler. Learning algorithms for active learning. *arXiv preprint arXiv:1708.00088*, 2017.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

Luca Bertinetto, Joao F. Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HyxnZh0ct7`.

Olivier Canévet and François Fleuret. Large scale hard sample mining with monte carlo tree search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5128–5137, 2016.

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

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.

Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, pages 9516–9527, 2018.

Ben Harwood, Vijay Kumar BG, Gustavo Carneiro, Ian Reid, and Tom Drummond. Smart mining for deep metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2821–2829, 2017.

Tomoharu Iwata and Atsutoshi Kumagai. Meta-learning from tasks with heterogeneous attribute spaces. *Advances in Neural Information Processing Systems*, 33, 2020.

Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379. IEEE, 2009.

Krishnateja Killamsetty, Changbin Li, Chen Zhao, Rishabh Iyer, and Feng Chen. A reweighted meta learning framework for robust few shot learning. *arXiv preprint arXiv:2011.06782*, 2020.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *arXiv preprint arXiv:1906.08158*, 2019.

Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Learning active learning from data. In *Advances in Neural Information Processing Systems*, pages 4225–4235, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer, 1994.

Jiang Lu, Sheng Jin, Jian Liang, and Changshui Zhang. Robust few-shot learning for user-provided data. *IEEE transactions on neural networks and learning systems*, 2020.

Pratik Mazumder, Pravendra Singh, and Vinay P Namboodiri. Rnnp: A robust few-shot learning approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2664–2673, 2021.

Yassine Ouali, Céline Hudelot, and Myriam Tami. An overview of deep semi-supervised learning, 2020.

Kunkun Pang, Mingzhi Dong, Yang Wu, and Timothy Hospedales. Meta-learning transferable active learning policies by deep reinforcement learning. *arXiv preprint arXiv:1806.04798*, 2018.

Anastasia Pentina, Viktoriia Sharmanska, and Christoph H Lampert. Curriculum learning of multiple tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5492–5500, 2015.

Sachin Ravi and Hugo Larochelle. Meta-learning for batch mode active learning. 2018.

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

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4334–4343. PMLR, 2018b.

James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. *arXiv preprint arXiv:1906.07697*, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Ravi Sachin and Larochell Hugo. Optimization as a model for few-shot learning. *ICLR*, 2017.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016.

Nikolaos Sarafianos, Theodore Giannakopoulos, Christophoros Nikou, and Ioannis A Kakadiaris. Curriculum learning for multi-task classification of visual attributes. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2608–2615, 2017.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 287–294. ACM, 1992.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

Alex J. Smola and Bernhard Schölkopf. Sparse greedy matrix approximation for machine learning. In *International Conference on Machine Learning 17*, pages 911–918. 2000.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019.

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

Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J Lim. Multimodal model-agnostic meta-learning via task-aware modulation. *arXiv preprint arXiv:1910.13616*, 2019.

Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *International Conference on Machine Learning*, pages 5238–5246. PMLR, 2018.

Mark Woodward and Chelsea Finn. Active one-shot learning. *arXiv preprint arXiv:1702.06559*, 2017.

Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7343–7353, 2018.

Yingtian Zou and Jiashi Feng. Hierarchical meta learning. *arXiv preprint arXiv:1904.09081*, 2019.