Effect of diversity in Meta-Learning

Ramnath Kumar [1,*], Tristan Deleu[1,2], Yoshua Bengio[1,2,3]

Abstract

Few-shot learning aims to learn representations that can tackle novel tasks given a small number of examples. Recent studies show that task distribution plays a vital role in the performance of the model. Conventional wisdom is that task diversity should improve the performance of meta-learning. In this work, we find evidence to the contrary; we study different task distributions on a myriad of models and datasets to evaluate the effect of task diversity on meta-learning algorithms. For this experiment, we train on two datasets - Omniglot and *mini*ImageNet and with three broad classes of meta-learning models - Metric-based (i.e., Protonet, Matching Networks), Optimization-based (i.e., MAML, Reptile, and MetaOptNet), and Bayesian meta-learning models (i.e., CNAPs). Our experiments demonstrate that the effect of task diversity on all these algorithms follows a similar trend, and task diversity does not seem to offer any benefits to the learning of the model. Furthermore, we also demonstrate that even a handful of tasks, repeated over multiple batches, would be sufficient to achieve a performance similar to uniform sampling and draws into question the need for additional tasks to create better models.

1 Introduction

It is widely recognized that humans can learn new concepts based on very little supervision, i.e., with few examples (or "shots"), and generalize these concepts to unseen data as mentioned by [8]. Recent advances in deep learning, on the other hand, have primarily relied on datasets with large amounts of labeled examples, primarily due to overfitting concerns in low data regimes. Although the development of better data augmentation and regularization techniques can alleviate these concerns, many researchers now assume that future breakthroughs in low data regimes will emerge from meta-learning, or "learning to learn." Here, we study the effect of task diversity in the low data regime and its effect on various models. In this meta-learning setting, a model is trained on a handful of labeled examples at a time under the assumption that it will learn how to correctly project examples of different classes and generalize this knowledge to unseen labels at test time.
Although this setting is often used to illustrate the remaining gap between human capabilities and machine learning, we could argue that the domain of meta-learning is still nascent. The domain of task selection has remained virtually unexplored in this setting. Conventional wisdom is that the performance of the model will improve as we train on more diverse tasks. To test this hypothesis to its limits, we define various task samplers which either limit task diversity by selecting a subset of overall tasks or improving task diversity using approaches such as Determinantal Point Processes (DPPs).

Our contributions in this work are as follows:

- We show that, against conventional wisdom, task diversity does not significantly boost performance in meta-learning. Instead, limiting task diversity and repeating the same tasks

[*] Work done during an internship at Mila; Correspondence author ramnathkumar181@gmail.com. [1]Mila, Québec Artificial Intelligence Institute. [2]Université de Montréal. [3]CIFAR, IVADO.

5th Workshop on Meta-Learning at NeurIPS 2021, Sydney, Australia.

over the training phase allows the model to obtain performances similar to models trained on Uniform Sampler without any adverse effects.

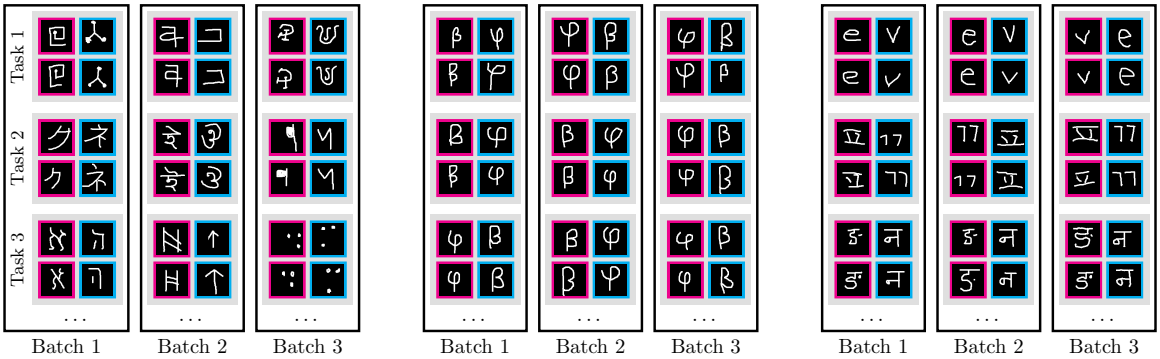
- We also show that increasing task diversity using sophisticated samplers such as DPP or Online Hard Task Mining (OHTM) Samplers do not significantly boost performance. Instead, the dynamic-DPP Sampler harms the model due to the increased task diversity.
- We empirically show that repeating tasks over the training phase can perform similarly to a model trained on the Uniform Sampler, achieving similar performance with only a fragment of data. This key finding questions the need to increase the support set pool to improve the model's performance.

2 Related Works

Meta-learning formulations typically rely on episodic training, wherein an algorithm adapts to a task, given its support set, to minimize the loss incurred on the query set. Meta-learning methods differ in terms of the algorithms they learn, and can be broadly classified under four prominent classes: *Metric-based*, *Model-based*, *Optimization-based* and *Bayesian-based* approaches. *Metric-based methods* such as [5, 22, 20, 21] operate on the core idea similar to nearest neighbors algorithm and kernel density estimation. These methods are also called non-parametric approaches. *Model-based methods* such as [17, 12] depend on a model designed specifically for fast learning, which updates its parameters rapidly with a few training steps, achieved by its internal architecture or controlled by another meta-learner model. Generic deep learning models learn through backpropagation of gradients, which are neither designed to cope with a small number of training samples nor converge within a few optimization steps. To address this, *Optimization-based methods* such as [15, 2, 13] were proposed, which were better suited to learn from a small number of samples. However, all the above approaches are deterministic and are not the most suited for few-shot problems, generally ambiguous. Hence, *Bayesian-based methods* such as [24, 16] were proposed which helped address the above issue.

Although research in meta-learning models has attracted much attention recently, the effect of task diversity is virtually unexplored in the domain of meta-learning. However, task sampling and task diversity have been more extensively studied in other closely related problems such as active learning. Active learning involves selecting unlabeled data items to the label in order to improve an existing classifier best. Although most of the approaches in this domain are based on heuristics, there are few approaches to sample a batch of samples for active learning. [14] proposed an approach to sample a batch of samples using a protonet as the backbone architecture. The model tries to maximize the query set, given support set and unlabeled data. Other works such as [4] proposed a framework named CACTUs, which samples tasks/examples using relatively simple task construction mechanisms such as clustering embeddings. The unsupervised representations learned via these samples lead to a good performance on various downstream human-specified tasks.

Although nascent, a few recent works aim to improve meta-learning by explicitly looking at the task structure and relationships. Among these, [23] proposed an approach to handle the lack of mutual exclusiveness among different tasks through an information-theoretic regularized objective. In addition, several popular meta-learning methods [9, 20], in order to improve the meta-test performance, change the number of ways or shots of the sampled meta-training tasks, thus increasing the complexity and diversity of the tasks. Other works such as [10] proposed an approach to sample classes using class-pair-based sampling and class-based sampling. The Class-pair based Sampler selects pairs of classes that confuse the model the most. The class-based Sampler samples each class independently and does not consider the task's difficulty as a whole. Our OHTM sampler is similar to the Class-pair based Sampler. Other works such as [11] propose to augment the set of possible tasks by augmenting the pre-defined set of classes that generate the tasks with varying degrees of rotated inputs as new classes. Other works such as [18] look at the structure and diversity of tasks specifically through the lens of support set diversity, and show that, surprisingly, reducing diversity (by fixing support set) not only maintains—but in many cases, significantly improves—the performance of meta-learning. This experiment is very similar to our No Diversity Task Sampler if the size of the support set is equal to the number of classes per task. However, in this work, we extend their work on MetaOptNet, Protonet to many other models and a myriad of samplers to better understand task diversity in meta-learning. To the best of our knowledge, we are the first to study the effect of task diversity in meta-learning to this extent.



(a) Uniform sampler (b) No Diversity Task sampler (c) No Diversity Batch sampler

Figure 1: Illustration of (a) the Uniform Sampler, (b) the No Diversity Task Sampler, and (c) the No Diversity Batch Sampler.

3 Background

Here, we review some of the fundamental ideas required to understand our few-shot learning experiments better.

3.1 Episodic few-shot learning

In episodic few-shot learning, an episode is represented as an K-way, N-shot classification problem where N is the number of examples per class and K is the number of unique class labels. During training, the data in each episode is provided as a support set $S = \{(x_{1,1}, y_{1,1}), ..., (x_{N,K}, y_{N,K})\}$ where $x_{i,j} \in \mathbb{R}^D$ is the i-th instance of the j-th class, and $y_j \in \{0, 1\}^K$ is its corresponding one-hot labeling vector. Each episode aims to optimize a function f that classifies new instances provided through a "query" set Q, containing instances of the same class as S. This task is difficult because N is typically very small (e,g, 1 to 10). The classes change every episode. The actual test set used to evaluate a model does not contain classes seen in support sets during training. In the task-distribution view, meta-learning is a general-purpose learning algorithm that can generalize across tasks and ideally enable each new task to be learned better than the last. We can evaluate the performance of ω over a distribution of tasks $p(\mathcal{T})$. Here we loosely define a task to be a dataset and loss function $\mathcal{T} = \{\mathcal{D}, \mathcal{L}\}$. Learning how to learn thus becomes:

$$\min_{\omega} \mathop{\mathbb{E}}_{\tau \sim p(\tau)} \mathcal{L}(\mathcal{D}; \omega) \tag{1}$$

where $\mathcal{L}(\mathcal{D}; \omega)$ measures the performance of a model trained using ω on dataset $\mathcal{D}$ and $p(\tau)$ indicates the task distribution. In this experiment, we extend this setting such that we vary the task diversity in the train split to study the effects on test split, which remains to use uniform or random sampling for tasks.

3.2 Determinantal Point Processes (DPPs)

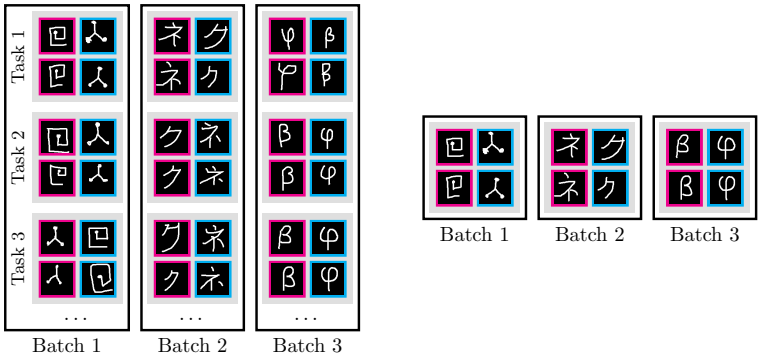
A DPP is a probability distribution over subsets of a ground set $\mathcal{Y}$, where we assume $\mathcal{Y} = \{1, 2, ..., N\}$ and $N = |\mathcal{Y}|$. An L-ensemble defines a DPP using a real, symmetric, and positive-definite matrix $\mathbf{L}$ indexed by the elements of $\mathcal{Y}$. The probability of sampling a subset $Y = A \subseteq \mathcal{Y}$ can be written as:

$$P(Y = A) \propto \det \mathbf{L}_A, \tag{2}$$

where $\mathbf{L}_A := [L_{i,j}]_{i,j \in A}$ is the restriction of $\mathbf{L}$ to the entries indexed by the elements of A. As $\mathbf{L}$ is a positive semi-definite, there exists a $d \times N$ matrix Ψ such that $\mathbf{L} = \Psi^T \Psi$ where $d \leq N$. Using this principle, we define the probability of sampling as:

$$P(Y = A) \propto \det \mathbf{L}_A = \mathit{Vol}^2(\{\Psi_i\}_{i \in A}), \tag{3}$$

where the RHS is the squared volume of the parallelepiped spanned by $\{\Psi_i\}_{i \in A}$. In Eq. 3, Ψ_i is defined as the feature vector of element i, and each element $L_{i,j}$ in $\mathbf{L}$ is the similarity measured by dot



(a) No Diversity Tasks per Batch sampler (b) Single Batch Uniform sampler

Figure 2: Illustration of (a) the No Diversity Task per Batch Sampler, and (b) the Single Batch Uniform Sampler.

products between elements i and j. Hence, we can verify that a DPP places higher probabilities on diverse sets because the more orthogonal the feature vectors are, the larger the volume parallelepiped spanned by the feature vector is. In this work, these feature embeddings represent class embeddings, which are derived using either a pre-trained protonet model or the model being trained as discussed in Sec. 3.3.

In a DPP, the cardinality of a sampled subset, $|A|$, is random in general. A k-DPP is an extension of the DPP proposed in the work of [6], where the cardinality of subsets are fixed as k (i.e., $|A| = k$). In this work, we use k-DPPs as an off-the-shelf implementation to retrieve classes that represent a task used in the meta-learning step.

3.3 Task Sampling

In this work, we experiment with eight distinct task samplers, each offering a different level of task diversity. To demonstrate the task samplers, we use a 2-way classification problem with a meta-batch size of 2 and denote each class with a unique alphabet.

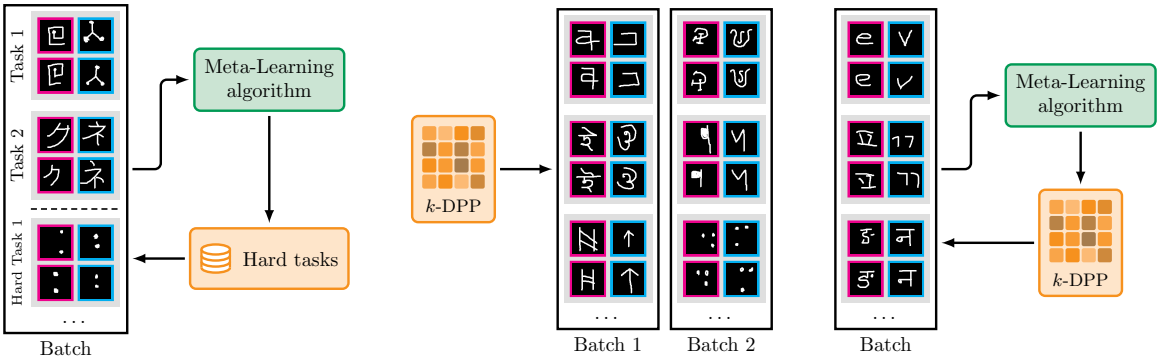
Uniform Sampler This is the most widely used Sampler used in the setting of meta-learning. The Sampler gives equal probability to every task and is intuitively a random sampler. An illustration of this Sampler is shown in Figure 1.

No Diversity Task Sampler In this setting, we uniformly sample one set of the task at the beginning and propagate the same task across all batches and meta-batches. Note that repeating the same class over and over again does not simply repeat the same images/inputs as we episodically retrieve different images for each class. An illustration of this Sampler is shown in Figure 1.

No Diversity Batch Sampler In this setting, we uniformly sample one set of tasks for batch one and propagate the same tasks across all other batches. Furthermore, we shuffle these tasks to enforce that the model does not overfit. An illustration of this Sampler is shown in Figure 1.

No Diversity Tasks per Batch Sampler In this setting, we uniformly sample one set of tasks for a given batch and propagate the same tasks for all meta-batches. We then repeat this same principle for sampling the next batch. Furthermore, we shuffle these tasks to enforce that the model does not overfit. An illustration of this Sampler is shown in Figure 2.

Single Batch Uniform Sampler In this setting, we set the meta-batch size to one. This Sampler is intuitively the same as no diversity task per batch sampler, without the repetition of tasks. This Sampler would be an ideal ablation study for the repetition of tasks in the meta-learning setting. An illustration of this Sampler is shown in Figure 2.



(a) Online Hard Task Mining sampler (b) Static DPP sampler (c) Dynamic DPP sampler

Figure 3: Illustration of (a) Online Hard Task Mining Sampler, (b) the Static DPP Sampler, and (c) the Dynamic DPP Sampler.

Online Hard Task Mining Sampler This setting is inspired by the works of [19] where they proposed OHEM, which yielded significant boosts in detection performance on benchmarks like PASCAL VOC 2007 and 2012. However, to reproduce OHEM for meta-learning, we only apply the OHEM sampler for half the meta-batch size and uniform sampler for the remaining half. This approach would allow us to involve many tasks and not restrict us to only known tasks. Furthermore, to avoid OHEM in the initial stages, we sample tasks with a uniform sampler until the buffer of tasks seen by the model becomes sufficiently big, say 50 in our case. An illustration of this Sampler is shown in Figure 3.

Static DPP Sampler Determinantal Point Processes (DPP) have been used for several machine learning problems such as the works of [7]. They have also been used in other problems such as the active learning settings in the works of [1] and mini-batch sampling problems in the works of [25]. These algorithms have also inspired other works in active learning in the batch mode setting, such as [14]. In this setting, we use DPP as an off-the-shelf implementation to sample tasks based on task embeddings. These task embeddings are generated using our pre-trained protonet model. The DPP instance is used to sample the most diverse tasks based on these embeddings, and used for meta-learning. An illustration of this Sampler is shown in Figure 3.

Dynamic DPP Sampler In this setting, we extend the previous sDPP setting such that the model in training generates the task embeddings. The Sampler is motivated by the intuition that selecting the most diverse tasks for a given model will help learn better. Furthermore, to avoid DPP in the initial stages, we sample tasks with a uniform sampler until the model becomes sufficiently trained, say 500 batches in our case. An illustration of this Sampler is shown in Figure 3.

4 Experiments

The experiment aims to answer the following questions: (a) How does task diversity affect meta-learning? (b) Do sophisticated samplers such as OHEM or DPP offer any significant boost in performance? (c) Are there any rule of thumb or general good practices when it comes to sampling tasks?

To make an exhaustive study on the effect of task diversity in meta-learning, we train on two datasets - Omniglot and *mini*ImageNet. We train three broad classes of meta-learning models on these datasets - Metric-based (i.e., Protonet, Matching Networks), Optimization-based (i.e., MAML, Reptile, and MetaOptNet), and Bayesian meta-learning models (i.e., CNAPs). More details about the datasets which were used in our experiments are discussed in Sec A.1. More details about the models and their hyperparameters are discussed in Sec A.2. We created a pool of 1024 tasks used for testing in our experiments to make an accurate comparison.

The code and implementation of all our experiments are publicly available at `https://github.com/RamnathKumar181/Task-Diversity-meta-learning`.

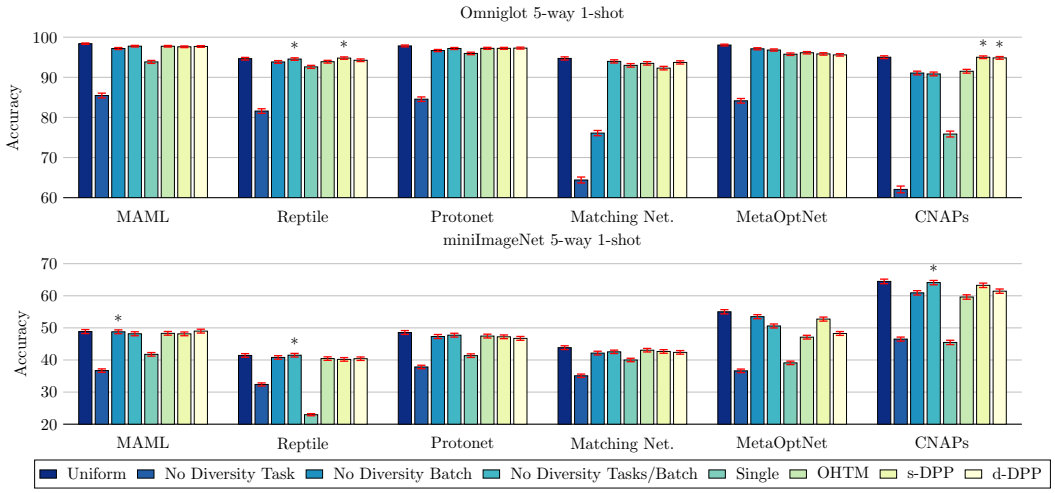


Figure 4: Average accuracy on Omniglot 5-way 1-shot & *mini*ImageNet 5-way 1-shot, with 95% confidence interval. All samplers are poorer than the Uniform Sampler and are statistically significant (with a p-value $p = 0.05$). We use the symbol $*$ to represent the instances where the results are not statistically significant and similar to the performance achieved by Uniform Sampler.

4.1 Results

In this section, we present the results of our experiments. Figure-4 presents the performance of the six models on the Omniglot and *mini*ImageNet under different task samplers in the 5-way 1-shot setting. Table-1 presents the same results with higher precision.

We also reproduce our experiments on the 20-way 1-shot setting on the Omniglot dataset to establish that these trends are shared across different settings. Figure-5 presents our the performance of the models under this setting. Furthermore, the results on the 20-way 1-shot experiments are presented in Table-2 with higher precision.

The above plots are sufficient to convey the crux of our work. We empirically show that task diversity does not lead to any significant boost in the performance of the models. In the subsequent section, we discuss some of the other key findings from our work.

5 Discussion

In this section, we discuss few empirical results from our experiments and shed light on some of the key findings from our research.

Poor performance by NDT Sampler The lowest performance is consistently obtained by the No Diversity Task Sampler, which is reasonable since the model only sees one task throughout its training. What is fascinating is that just one task is sufficient for the model to reach a reasonably decent performance in most cases.

Disparity between Single Batch Uniform and NDTB Sampler Another exciting result is the Disparity between Single Batch Uniform Sampler and No Diversity Tasks per Batch Sampler. As mentioned earlier, the only difference between the two samplers is that tasks are repeated in the latter. However, this repetition seems to offer a great deal of information to the model and allows the model to perform on par with the Uniform Sampler. It might be possible that the Single Batch Uniform Sampler obtains the performance observed by the No Diversity Tasks per Batch Sampler if trained for enough epochs. However, it would be safe to comment that the convergence of the model is significantly faster in the latter. Thus, repeating tasks might help speed up the convergence of the model when we have a fixed and handful amount of data.

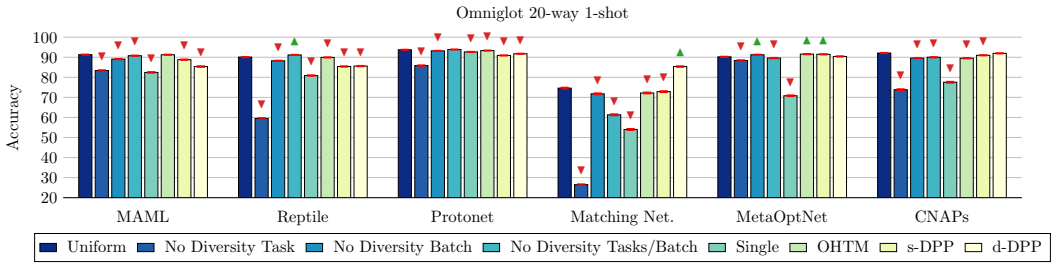


Figure 5: Average accuracy on Omniglot 20-way 1-shot, with a 95% confidence interval. We denote all samplers that are worse than the Uniform Sampler and are statistically significant (with a p-value $p = 0.05$) with ▼, and those that are significantly better than the Uniform Sampler with ▲.

Disparity between s-DPP and d-DPP Sampler We also note that s-DPP and d-DPP samplers do not offer any boost in performance when compared to the regular Uniform Sampler. Furthermore, there seems to be a significant disparity between these two samplers. We believe that d-DPP, which computes the most diverse tasks at regular intervals, harms the model with the diverse tasks since we observe that the model's performance degrades over epochs, especially for methods such as matching networks, protonet, and reptile. For example, consider the scenario where the model is trained on tasks involving dogs and tractors. This task is relatively easy to learn and would not require the model to fine-tune a great deal. However, during test time, suppose our task involves classifying cats and dogs; this would be a problem since the model has not learned the intricacies of the two classes. Thus, diversity seems to do more harm than good in this case. The best example of this is observed by Matching Networks in Omniglot 5-way 1-shot setting as shown in Figure 6, where each instance of diverse sampling harms the model significantly.

OHTM Sampler offers no significant performance boost The OHTM Sampler is quite sophisticated since it regularly samples diverse tasks, as well as selects the most challenging tasks to improve the model. It is needless to say; the model requires more computational power and time than the Uniform Sampler. However, the OHTM Sampler offers no significant boost in performance when compared to the Uniform Sampler.

Comparison between NDTB, NDB, and Uniform Sampler From our experiments, we also notice that the No Diversity Tasks per Batch Sampler and No Diversity Batch Sampler are pretty similar to the Uniform Sampler in terms of performance. This would suggest that the model trained on only a data fragment can perform similarly to a model trained on the Uniform Sampler.

Abnormal run of matching networks d-DPP (20-way 1-shot) In our run on the matching networks with the d-DPP Sampler under the 20-way 1-shot setting, we ran across a peculiar error. The prototypes generated by the matching networks were sometimes not fit to be used by the d-DPP Sampler to sample 20 unique classes. The reason being that the rank of the matrix generated using the embeddings was lower than the required number of classes per task (i.e., 20). To create a workaround for this sole experiment, we chose to sample 5 diverse classes at a time and append them to create the task. We hypothesize that the prototypes created by matching networks are unsuitable for downstream tasks and warrant further research regarding this behavior.

Peculiar behavior with MetaOptNet model Compared to all other models, MetaOptNet seems to be immune to the effects of task diversity to a great extent. The convergence of the model seems to follow a general pattern and achieve similar performance across task distributions except for the Single Batch Uniform Sampler and No Diversity Task sampler. Furthermore, we do not observe the expected pattern of d-DPP Sampler, where the performance drops upon mining diverse tasks. We present the convergence graph of the MetaOptNet model on Omniglot 5-way 1-shot run in Figure-7 with an added smoothing factor of 1.

General Trend From our experiments, we notice that there are generally two classes of samplers: High Performing Samplers and Low Performing Samplers. The High Performing Samplers include No Diversity Batch, No Diversity Tasks per Batch, Uniform, OHTM, and s-DPP Sampler. The Low

Performing Samplers include No Diversity Task, Single Batch Uniform, and d-DPP Sampler. This trend is shared across all datasets and models. There are some perturbations in ranking within the two classes, but the High Performing Samplers tend to perform better than the Low Performing Samplers.

6 Conclusion

In this paper, we have studied the effect of task diversity in meta-learning. We have empirically shown that task diversity does not lead to any significant boost in performance in meta-learning. Instead, limiting task diversity and repeating the same tasks over the training phase allows us to obtain similar performances to the Uniform Sampler without any significant adverse effects. Furthermore, We also show that sophisticated samplers such as OHEM or DPP samplers do not offer any significant boost in performance. In contradiction, we notice that increasing task diversity using the d-DPP Sampler hampers the performance of the meta-learning model. Our experiments using the NDTB and NDB empirically show that a model trained on only a data fragment can perform similarly to a model trained using Uniform Sampler. This is a crucial finding since this questions the need to increase the support set pool to improve the models' performance. We believe that the experiments we performed lay the roadwork to further research for the effect of task diversity domain in meta-learning and lay some groundwork and rules of thumb for task sampling for meta-learning.

Acknowledgements

We would like to thank Sony Corporation for funding this research through the Sony Research Award Program.

References

[1] Erdem Bıyık et al. "Batch active learning using determinantal point processes". In: *arXiv preprint arXiv:1906.07975* (2019).

[2] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1126–1135.

[3] Marta Garnelo et al. "Conditional neural processes". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1704–1713.

[4] Kyle Hsu, Sergey Levine, and Chelsea Finn. "Unsupervised learning via meta-learning". In: *arXiv preprint arXiv:1810.02334* (2018).

[5] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. "Siamese neural networks for one-shot image recognition". In: *ICML deep learning workshop*. Vol. 2. Lille. 2015.

[6] Alexandre Kuhn, Ad Aertsen, and Stefan Rotter. "Higher-order statistics of input ensembles and the response of simple model neurons". In: *Neural computation* 15.1 (2003), pp. 67–101.

[7] Alex Kulesza and Ben Taskar. "Determinantal point processes for machine learning". In: *arXiv preprint arXiv:1207.6083* (2012).

[8] Brenden Lake et al. "One shot learning of simple visual concepts". In: *Proceedings of the annual meeting of the cognitive science society*. Vol. 33. 33. 2011.

[9] Kwonjoon Lee et al. "Meta-learning with differentiable convex optimization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10657–10665.

[10] Chenghao Liu et al. "Adaptive Task Sampling for Meta-learning". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer. 2020, pp. 752–769.

[11] Jialin Liu, Fei Chao, and Chih-Min Lin. "Task augmentation by rotating for meta-learning". In: *arXiv preprint arXiv:2003.00804* (2020).

[12] Tsendsuren Munkhdalai and Hong Yu. "Meta networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 2554–2563.

[13] Alex Nichol, Joshua Achiam, and John Schulman. "On first-order meta-learning algorithms". In: *arXiv preprint arXiv:1803.02999* (2018).

[14] Sachin Ravi and Hugo Larochelle. "Meta-learning for batch mode active learning". In: (2018).

[15] Sachin Ravi and Hugo Larochelle. "Optimization as a model for few-shot learning". In: (2016).

[16] James Requeima et al. "Fast and flexible multi-task classification using conditional neural adaptive processes". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 7959–7970.

[17] Adam Santoro et al. "Meta-learning with memory-augmented neural networks". In: *International conference on machine learning*. PMLR. 2016, pp. 1842–1850.

[18] Amrith Setlur, Oscar Li, and Virginia Smith. "Is Support Set Diversity Necessary for Meta-Learning?" In: *arXiv preprint arXiv:2011.14048* (2020).

[19] 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*. 2016, pp. 761–769.

[20] Jake Snell, Kevin Swersky, and Richard S Zemel. "Prototypical networks for few-shot learning". In: *arXiv preprint arXiv:1703.05175* (2017).

[21] Flood Sung et al. "Learning to compare: Relation network for few-shot learning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1199–1208.

[22] Oriol Vinyals et al. "Matching networks for one shot learning". In: *Advances in neural information processing systems* 29 (2016), pp. 3630–3638.

[23] Mingzhang Yin et al. "Meta-learning without memorization". In: *arXiv preprint arXiv:1912.03820* (2019).

[24] Jaesik Yoon et al. "Bayesian model-agnostic meta-learning". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018, pp. 7343–7353.

[25] Cheng Zhang et al. "Active mini-batch sampling using repulsive point processes". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 5741–5748.

A Appendix

A.1 Dataset

Omniglot Omniglot is a benchmark dataset proposed by [8] for few-shot image classification tasks. Omniglot dataset consists of 20 instances and 1623 characters from 50 different alphabets. We experiment with both 5-way 1-shot and 2-way 1-shot in this work.

***mini*ImageNet** *mini*ImageNet is another benchmark dataset proposed by [15] for few-shot image classification tasks. The *mini*ImageNet dataset involves 64 training classes, 12 validation classes, and 24 test classes. We run under the setting 5-way 1-shot for this experiment.

A.2 Models

This section describes some of the models we used for our experiments and the hyperparameters used for their training.

A.2.1 Prototypical Networks

Prototypical Networks proposed by [20] constructs a prototype for each class and then classifies each query example as the class whose prototype is 'nearest' to it under Euclidean distance. More concretely, the probability that a query example x^* belongs to class k is defined as:

$$p(y^* = k|x^*, \mathcal{S}) = \frac{\exp(-\|g(x^*) - c_k\|_2^2)}{\sum_{k' \in \{1,\ldots,N\}} \exp(-\|g(x^*) - c_{k'}\|_2^2)} \tag{4}$$

Where c_k is the 'prototype' for class k: the average embeddings of class k's support examples.

Hyperparameters In our experiments on Omniglot and *mini*ImageNet under a 5-way, 1-shot setting, we run the model for 100 epochs with a batch size of 32 and a meta-learning rate of 0.001. We use an Adam optimizer to make gradient steps and a StepLR scheduler with step size 0.4 and

gamma 0.5. The same hyperparameters are used for training our model on Omniglot under a 20-way 1-shot setting.

A.2.2 Matching Networks

Matching Networks proposed by [22] labels each query example as a cosine distance-weighted linear combination of the support labels:

$$p(y^* = k|x^*, \mathcal{S}) = \sum_{i=1}^{|\mathcal{S}|} \alpha(x^*, x_i)\Phi_{y_i=k}, \tag{5}$$

where Φ_A is the indicator function and $\alpha(x^*, x_i)$ is the cosine similarity between $g(x^*)$ and $g(x_i)$, softmax normalized over all support examples x_i, where $1 \leq i \leq |\mathcal{S}|$.

We had trouble reproducing the results from matching networks using cosine distance since the convergence seemed to be slow and the final performance dependent on the random initialization. This is similar to what is observed by other repositories such as https://github.com/oscarknagg/few-shot. Since we are focused on the relative performance of the samplers for a given model, this discrepancy would not affect our study of task diversity in any manner.

Hyperparameters In our experiments on Omniglot and *mini*ImageNet under a 5-way, 1-shot setting, we run the model for 100 epochs with a batch size of 32 and an Adam optimizer with a meta-learning rate of 0.001 and a weight decay of 0.0001. The same hyperparameters are used for training our model on Omniglot under a 20-way 1-shot setting.

A.2.3 MAML

MAML proposed by [2] uses a linear layer parametrized by **W** and **b** on top of the embedding function $g(.;\theta)$ and classifies a query example as:

$$p(y^*|x^*, \mathcal{S}) = softmax(\mathbf{b}' + \mathbf{W}' g(x^*; \theta')), \tag{6}$$

where the output layer parameters **W'** and **b'** and the embedding function parameters θ' are obtained by performing a small number of within-episode training steps on the support set $\mathcal{S}$, starting from initial parameter values $(\mathbf{b}, \mathbf{W}, \theta)$.

Hyperparameters In our experiments on Omniglot and *mini*ImageNet under a 5-way, 1-shot setting, we run the model for 150 epochs with a batch size of 32, with the Adam optimizer with a meta-learning rate of 0.001, number of inner adaptations as 1, and step size 0.4. For our experiments on Omniglot under the 20-way 1-shot setting, we set the step size of 0.1 and the number of inner adaptations to 5, batch size of 16, and kept all other hyperparameters constant.

A.2.4 Reptile

Like MAML, Reptile proposed by [13] learns an initialization for the parameters of a neural network model, such that when we optimize these parameters at test time, learning is fast - i.e., the model generalizes from a small number of test tasks. Reptile converges towards a solution ϕ that is close (in Euclidean distance) to each task τ's manifold of optimal solutions. Let ϕ denote the network initialization, and $W = \phi + \Delta\phi$ denote the network weights after performing some sort of update. Let $\mathcal{W}_\tau^*$ denote the set of optimal network weights for task τ. We want to find ϕ such that the distance $D(\phi, \mathcal{W}_\tau^*)$ is small for all tasks:

$$\min_\phi \mathbb{E}_\tau[\frac{1}{2} D(\phi, \mathcal{W}_\tau^*)^2] \tag{7}$$

The official repository seems to train the model with a 5-way 15-shot and test the model on a 5-way 1-shot. However, we do not consider this to be an accurate study for the effect of task diversity. In our work, we train and test the model in a 5-way 1-shot setting to ensure fair and accurate comparison with other models. We believe this to be the source of discrepancy in our performance scores. Since we are focused on the relative performance of the samplers for a given model, this discrepancy would not affect our study of task diversity in any manner.

Hyperparameters In our experiments on Omniglot and *mini*ImageNet under a 5-way, 1-shot setting, we run the model for 150 epochs with a batch size of 32, a learning rate of 0.01, a meta-learning rate of 0.001, number of inner adaptations as 5, and a step size of 0.4. For our experiments on Omniglot under the 20-way 1-shot setting, we set the meta-learning rate to 0.0005 and the number of inner adaptations to 10 and kept all other hyperparameters constant. Furthermore, we only run the model for 50 epochs due to the very high training time.

A.2.5 CNAPs

Conditional Neural Adaptive Processes proposed by [16] is able to efficiently solve new multi-class classification problems after an initial training phase. The proposed approach, based on Conditional Neural Processes (CNPs) mentioned in [3], adapts a small number of task-specific parameters for each new task encountered at test time. These parameters are conditioned on a set of training examples for the new task. They do not require any additional tuning to adapt both the final classification layer and feature extraction process. This allows the model to handle various input distributions. The CNPs construct predictive distributions given x^* as:

$$p(y^*|x^*, \theta, D^\tau) = p(y^*|x^*, \theta, \Psi^\tau = \Psi_\phi(D^\tau)), \tag{8}$$

where θ are global classifer parameters shared across tasks, Ψ^τ are local task-specific parameters, produced by a function $\Psi_\phi(.)$ that acts of D^τ. $\Psi_\phi(.)$ has another set of global parameters ϕ called *adaptation network parameters*. θ and ϕ are the learnable parameters in the model.

Hyperparameters In all our experiments with CNAPs, we run the model for ten epochs with a batch size of 16 and a meta-learning rate of 0.01.

A.2.6 MetaOptNet

MetaOptNet proposed by [9] proposes a linear classifier as the base learner for a meta-learning based approach for few-shot learning. The approach uses a linear support vector machine (SVM) to learn a classifier given a set of labeled training examples. The generalization error is computed on a novel set of examples from the same task. The objective is to learn an embedding model ϕ that minimizes generalization (or test) error across tasks given a base learner $\mathcal{A}$. Formally, the learning objective is:

$$\min_{\phi} \mathbb{E}_{\mathcal{T}}[\mathcal{L}^{meta}(\mathcal{D}^{test}; \theta, \phi), \text{ where } \theta = \mathcal{A}(\mathcal{D}^{test}; \phi)]. \tag{9}$$

The choice of base learner $\mathcal{A}$ has a significant impact on the above equation. The base learner that computes $\theta = \mathcal{A}(\mathcal{D}^{test}; \phi)$ has to be efficient since the expectation has to be computed over a distribution of tasks. This work considers base learners based on multi-class linear classifiers such as SVM, where the base learner's object is convex. Thus, the base learner can be simplified as:

$$\begin{aligned} &\theta = \mathcal{A}(\mathcal{D}^{test}; \phi) = \arg\min_{\{\mathbf{w}_k\}} \min_{\xi_i} \tfrac{1}{2} \textstyle\sum_k \|\mathbf{w}_k\|_2^2 + C \sum_n \xi_n; \text{ subject to:} \\ &\mathbf{w}_{y_n}.f_\phi(x_n) - \mathbf{w}_k.f_\phi(x_n) \geq 1 - \delta_{y_n,k} - \xi_n, \forall n, k \end{aligned} \tag{10}$$

where $\mathcal{D}^{train} = \{(x_n, y_n)\}, C$ is the regularization parameter and $\delta_{.,.}$ is the Kronecker delta function.

Furthermore, the official repository seems to train the model with a 5-way 15-shot and test the model on a 5-way 1-shot. However, we do not consider this to be an accurate study for the effect of task diversity. In our work, we train and test the model in a 5-way 1-shot setting to ensure fair and accurate comparison with other models. We believe this to be the source of discrepancy in our performance scores. Since we are focused on the relative performance of the samplers for a given model, this discrepancy would not affect our study of task diversity in any manner.

Hyperparameters In our experiments on Omniglot and *mini*ImageNet under a 5-way, 1-shot setting, we run the model for 60 epochs with a batch size of 32 and a meta-learning rate of 0.01. We use an SGD optimizer with a momentum of 0.9 and a weight decay of 0.0001 to make gradient steps. We also use a LambdaLR scheduler to train our model. The same hyperparameters are used for training our model on Omniglot under a 20-way 1-shot setting.

A.3 Additional Results

In this section, we present the results with higher precision from our earlier experiments in a Table-1 and Table-2. Subsequently, we also plot convergence curves to aid better visualizations of findings mentioned earlier in Figure-6 and Figure-7.

Dataset	Sampler	*MAML*	*Reptile*	*Protonet*	*Matching Networks*	*MetaOptNet*	*CNAPs*
Omniglot	Uniform Sampler	**98.38 ± 0.17**	94.64 ± 0.32	**97.82 ± 0.23**	**94.71 ± 0.39**	**98.04 ± 0.22**	**95.01 ± 0.35**
	No Diversity Task Sampler	85.46 ± 0.59	81.59 ± 0.57	84.55 ± 0.56	64.41 ± 0.74	84.15 ± 0.57	62.06 ± 0.83
	No Diversity Batch Sampler	97.17 ± 0.25	93.83 ± 0.34	96.67 ± 0.27	76.10 ± 0.65	97.11 ± 0.26	91.07 ± 0.46
	No Diversity Tasks per Batch Sampler	97.76 ± 0.20	94.55 ± 0.31 †	97.18 ± 0.25	93.97 ± 0.40	96.80 ± 0.27	90.84 ± 0.47
	Single Batch Uniform Sampler	93.84 ± 0.37	92.60 ± 0.38	95.95 ± 0.31	92.98 ± 0.44	95.76 ± 0.31	75.86 ± 0.73
	OHTM Sampler	97.74 ± 0.20	93.89 ± 0.34	97.22 ± 0.25	93.48 ± 0.43	96.12 ± 0.29	91.51 ± 0.47
	s-DPP Sampler	97.61 ± 0.21	**94.79 ± 0.30** †	97.22 ± 0.24	92.29 ± 0.44	95.83 ± 0.30	95.00 ± 0.33 †
	d-DPP Sampler	97.69 ± 0.21	94.25 ± 0.33	97.28 ± 0.24	93.71 ± 0.40	95.59 ± 0.30	94.84 ± 0.34 †
MiniImagenet	Uniform Sampler	**48.86 ± 0.62**	41.42 ± 0.56	**48.56 ± 0.60**	**43.84 ± 0.58**	**55.02 ± 0.66**	**64.48 ± 0.71**
	No Diversity Task Sampler	36.70 ± 0.53	32.38 ± 0.48	37.83 ± 0.53	35.08 ± 0.53	36.62 ± 0.55	46.51 ± 0.63
	No Diversity Batch Sampler	48.78 ± 0.60 †	40.80 ± 0.54	47.32 ± 0.62	42.15 ± 0.58	53.50 ± 0.63	60.92 ± 0.68
	No Diversity Tasks per Batch Sampler	48.17 ± 0.62	**41.49 ± 0.56** †	47.73 ± 0.60	42.54 ± 0.53	50.60 ± 0.62	64.11 ± 0.68 †
	Single Batch Uniform Sampler	41.76 ± 0.56	22.96 ± 0.33	41.35 ± 0.56	40.00 ± 0.54	39.10 ± 0.54	45.47 ± 0.67
	OHTM Sampler	48.30 ± 0.58	40.44 ± 0.54	47.45 ± 0.59	43.05 ± 0.55	47.11 ± 0.58	59.62 ± 0.69
	s-DPP Sampler	48.14 ± 0.59	40.19 ± 0.56	47.22 ± 0.58	42.66 ± 0.56	52.74 ± 0.63	63.26 ± 0.69
	d-DPP Sampler	48.99 ± 0.60	40.40 ± 0.54	46.73 ± 0.60	42.37 ± 0.56	48.26 ± 0.60	61.44 ± 0.67

Table 1: Performance metric of our models on different task samplers in the 5-way 1-shot setting.

Dataset	Sampler	*MAML*	*Reptile*	*Protonet*	*Matching Networks*	*MetaOptNet*	*CNAPs*
Omniglot	Uniform Sampler	**91.28 ± 0.22**	90.09 ± 0.22	93.72 ± 0.20	**74.62 ± 0.38**	90.20 ± 0.23	**92.09 ± 0.22**
	No Diversity Task Sampler	83.39 ± 0.29	59.49 ± 0.33	85.84 ± 0.27	26.50 ± 0.32	88.40 ± 0.26	73.82 ± 0.39
	No Diversity Batch Sampler	89.07 ± 0.25	88.23 ± 0.23	93.18 ± 0.20	71.77 ± 0.38	91.24 ± 0.22 ‡	89.56 ± 0.24
	No Diversity Tasks per Batch Sampler	90.77 ± 0.23	**91.15 ± 0.21** ‡	**93.85 ± 0.19** †	61.31 ± 0.41	89.59 ± 0.24	89.99 ± 0.24
	Single Batch Uniform Sampler	82.45 ± 0.31	80.89 ± 0.27	92.67 ± 0.20	54.01 ± 0.40	70.81 ± 0.35	77.54 ± 0.37
	OHTM Sampler	91.25 ± 0.22 †	89.92 ± 0.22	93.33 ± 0.20	72.20 ± 0.38	**91.56 ± 0.23** ‡	89.51 ± 0.25
	s-DPP Sampler	88.79 ± 0.24	85.40 ± 0.25	90.90 ± 0.22	72.86 ± 0.37	91.47 ± 0.22 ‡	90.98 ± 0.22
	d-DPP Sampler	85.36 ± 0.30	85.60 ± 0.25	91.74 ± 0.22	85.36 ± 0.30 ‡	90.40 ± 0.24 †	91.95 ± 0.22 †

Table 2: Performance metric of our models on different task samplers in the 20-way 1-shot setting.

Statistical Results We compare the performance of different models to the Uniform Sampler. All samplers are poorer than the Uniform Sampler and are statistically significant with a confidence interval of 95%. We use the symbol † to represent the instances where the results are not statistically significant and similar to the performance achieved by Uniform Sampler. We only observe four instances where a sampler performs significantly better than the Uniform Sampler, which we represent using the symbol ‡.

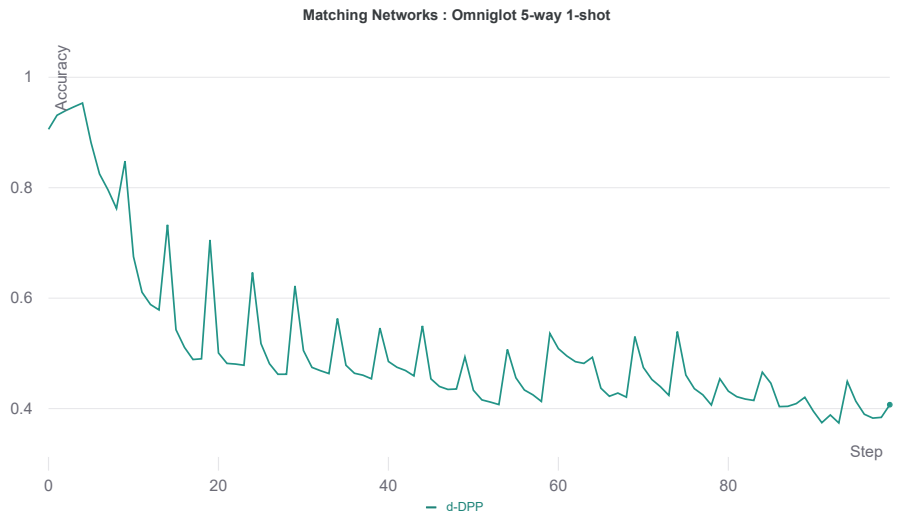


Figure 6: Convergence curve of Matching Networks model on Omniglot 5-way 1-shot.

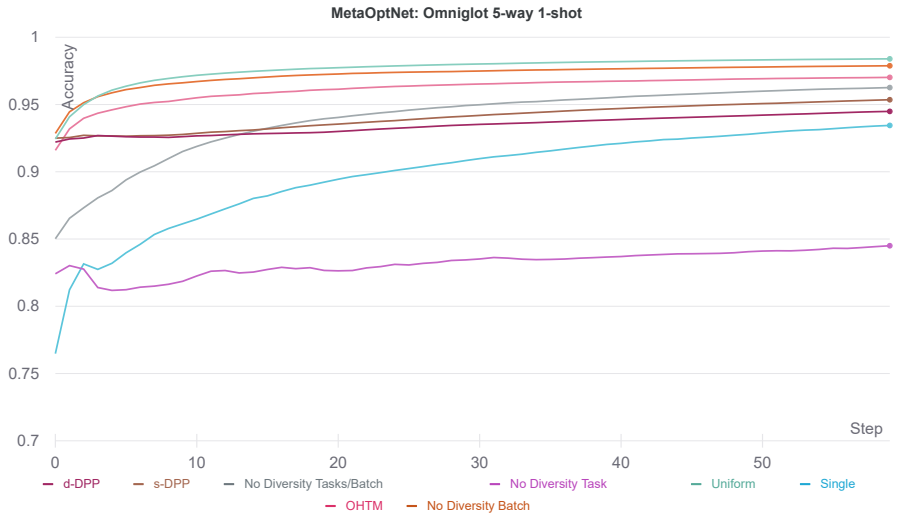


Figure 7: Convergence curve of MetaOptNet model on Omniglot 5-way 1-shot.