The Effect of Diversity in Meta-Learning

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

Google Research, India [1]
Mila, Québec Artificial Intelligence Institute, Université de Montréal [2]
CIFAR, IVADO [3]

Abstract

Recent studies show that task distribution plays a vital role in the meta-learner's performance. Conventional wisdom is that task diversity should improve the performance of meta-learning. In this work, we find evidence to the contrary; (i) our experiments draw into question the efficacy of our learned models: similar manifolds can be learned with a subset of the data (lower task diversity). This finding questions the advantage of providing more data to the model, and (ii) adding diversity to the task distribution (higher task diversity) sometimes hinders the model and does not lead to a significant improvement in performance as previously believed. To strengthen our findings, we provide both empirical and theoretical evidence.

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 (Lake et al. 2011). On the other hand, recent advances in deep learning have primarily relied on datasets with large amounts of labeled examples, primarily due to overfitting concerns in low data regimes. Although 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 impact 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 field of task selection has mainly remained under-explored in this setting. Hence, our exploration of this setting is much warranted. To the best of our knowledge, no previous work attempts to work with task diversity and its effect in the meta-learning setting.

Conventional wisdom is that the model's performance will improve as we train on more diverse tasks. This does seem intuitively sound: training on a diverse and large amount of classes should bring about a more extensive understanding of the world, thus learning multiple concepts of, let's say, the "world model". To test this hypothesis, we define task samplers that either limit task diversity by selecting a subset of overall tasks or improve task diversity using approaches such as Determinantal Point Processes (DPPs) proposed by (Macchi 1975). This problem is interesting since understanding the effect of diversity in meta-learning is closely linked to the model's ability to learn. In hindsight, this study is also an excellent metric to test the efficacy of our models, as will become more substantial in further sections.

1.1 Contributions

In this section, we present the main contributions of the paper:

- We show that limiting task diversity and repeating the same tasks over the training phase allows the model to obtain performances similar to models trained on Uniform Sampler without any adverse effects. (Section 4,5)
- We also show that increasing task diversity using sophisticated samplers such as DPP or Online Hard Task Mining (OHTM) Samplers does not significantly boost performance. Instead, this also harms the performance of the learner in certain instances. (Section 4,5)
- We also propose a suitable theoretical explanation for our findings from the connection to Simpson's paradox phenomenon from the discipline of causality as discussed briefly in Appendix D.1 in (Kumar, Deleu, and Bengio 2022).
- We also propose a metric to compute task diversity in the meta-learning setting. (Section 3)
- Our findings bring into question the efficiency of the model and the advantage it gains with access to more data using samplers such as the standard sampling regime – Uniform Sampler. If we can achieve similar performances with fewer data, the existing models have not taken advantage of the excess data it is provided with.

[*]Work done during an internship at Mila; Correspondence author: ramnathk@google.com.

2 Background

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

2.1 Episodic Few-Shot Learning

In episodic few-shot learning, an episode is represented as a N-way, K-shot classification problem where K is the number of examples per class and N 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 $y_j \in \{0,1\}^N$ 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 K is typically very small (e.g. 1 to 10). The classes change for 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 $p(\tau)$. Here we loosely define a task to be a dataset and loss function $\tau = \{\mathcal{D}_\tau, \mathcal{L}_\tau\}$. Learning how to learn thus becomes:

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

where $\mathcal{L}(\mathcal{D}; \omega)$ measures the performance of a model trained using network parameters ω on dataset $\mathcal{D}$, and $p(\tau)$ indicates the task distribution. In our experiments, we extend this setting such that we vary the task diversity of the train split to study the effects on test split, which remains unchanged (i.e. uniformly sampling test tasks).

2.2 Determinantal Point Processes (DPPs)

A Determinantal Point Process (DPP; Kulesza and Taskar 2012) is a probability distribution over subsets of a ground set $\mathcal{Y}$, where we assume $\mathcal{Y} = \{1, 2, \ldots, N\}$ and $N = |\mathcal{Y}|$. An $\mathbf{L}$-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:

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

where the RHS is the squared volume(vol) 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 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 Prototypical Network (Snell, Swersky, and Zemel 2017) model or the model being trained as discussed in Sec. 2.3.

In a DPP, the cardinality of a sampled subset, $|A|$, is random in general. A k-DPP (Kuhn, Aertsen, and Rotter 2003) is an extension of the DPP 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.

2.3 Task Sampling

In this work, we experiment with eight distinct task samplers, each offering a different level of task diversity. To illustrate the task samplers, we use a 2-way classification problem, and denote each class with a unique alphabet from the Omniglot dataset (Lake et al. 2011). To make our study more theoretically sound and less heuristic in nature, we create a more formal definition of task diversity and discuss it in more detail in Section 3.

Uniform Sampler This is the most widely used Sampler used in the setting of meta-learning (with mutually-exclusive tasks (Yin et al. 2019)). The Sampler creates a new task by sampling uniformly classes. An illustration of this Sampler is shown in Figure 1.

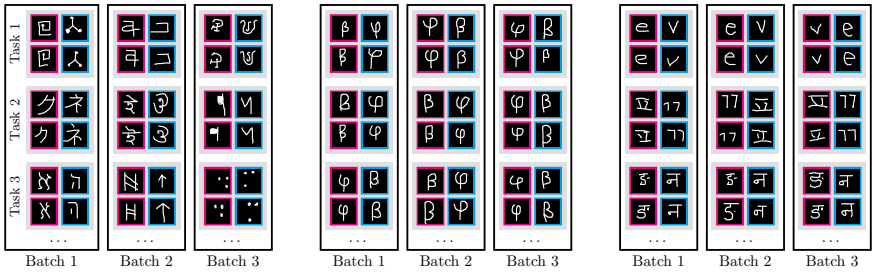
No Diversity Task Sampler In this setting, we uniformly sample one set of the task at the beginning and propagate the same task across all batches and meta-batches. Note that repeating the same class over and over again does not simply repeat the same images/inputs as we episodically retrieve different 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 the labels, as in the No Diversity Task Sampler, to prevent the model from overfitting. 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. Similar to the Samplers above, we also shuffle the labels to reduce overfitting. An illustration of this Sampler is shown in Figure 2.

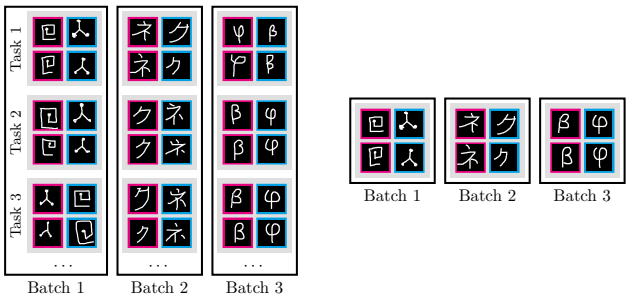
Single Batch Uniform Sampler In this setting, we set the meta-batch size to one. This Sampler is intuitively the same as the No Diversity Task per Batch Sampler, without the repetition of tasks inside a meta-batch. This Sampler would be an ideal ablation study for the repetition of tasks in the meta-learning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

Online Hard Task Mining Sampler This setting is inspired by the works of (Shrivastava, Gupta, and Girshick



(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.



(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.

2016), where they proposed OHEM, which yielded significant boosts in detection performance on benchmarks like PASCAL VOC 2007 and 2012. OHEM sampler samples the hardest tasks from a pool of tasks previously seen. 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. 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 (Kulesza and Taskar 2012). They have also been used in other problems such as the active learning settings (Bıyık et al. 2019) and mini-batch sampling problems (Zhang et al. 2019). These algorithms have also inspired other works in active learning in the batch mode setting (Ravi and Larochelle 2018). In this setting, we use DPP as an off-the-shelf implementation to sample tasks based on their class embeddings. These class 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 class embeddings. The Sampler is motivated by the intuition that selecting the most diverse tasks for a given model will help learn better. An illustration of this Sampler is shown in Figure 3.

3 Study of Diversity

3.1 Preliminaries

Before giving a more formal definition of task diversity, we set a few more fundamental ideas required to better understand our metric. In the domain of meta-learning, there have been no previous proposed definition of Task Diversity, and has remained highly heuristic and intuitive. Our definition could be used to serve as an established notion of "Task Diversity" to be used in future works. In this work, we consider the volume parallelopiped definition as discussed briefly below. Although simple, this definition is very intuitive to our concept of diversity in meta-learning. Our definition is highly robust and does consider diversity across various modalities such as classes, tasks, and batches. In this work, we compute the embedding from a pre-trained protonet model. It would also be possible to compute these embeddings from another neural network approximation function, such as ResNet, VGG, etc., trained on ILSVRC as is commonly used to compare the difference between two images in the computer vision domain. Below we briefly introduce the proposed definition of "Diversity" in the meta-learning domain.

Task Diversity We define task diversity as the diversity among classes within a task. This diversity is defined as the volume of parallelepiped spanned by the embeddings of each of these classes.

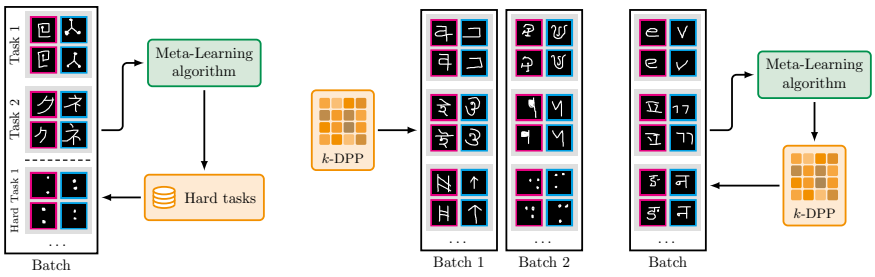
$$\mathcal{TD} \propto [\text{vol}(\mathcal{T})]^2$$

where $\mathcal{T}$ is defined as $\{c_1, \ldots, c_N\}$, where N is the number of ways, and c_i is the feature embedding of the i^{th} class. These feature embeddings are pre-computed using our pre-trained Protonet model, similar to the one used in sDPP. This value is analogous to the probability of selecting a task of the following classes.

Task Embedding We define the task embedding as the mean embedding of class features within that task. The task embedding is computed such that:

$$\mathcal{TE} = \frac{1}{m} \sum_{i=0}^{m} c_i$$

where the task is defined as $\{c_1, \ldots, c_N\}$, where N is the number of ways, and c_i is the embedding of the i^{th} class.



(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.

Batch Diversity We define batch diversity as the diversity among tasks within a mini-batch. This diversity is defined as the volume of parallelepiped spanned by the task embeddings of each of these tasks within a mini-batch:

$$\mathcal{BD} \propto [\text{vol}(\mathcal{B})]^2$$

where $\mathcal{B}$ is defined as $\{t_1, \ldots, t_m\}$, where m is the number of tasks within a mini-batch, and t_i is the feature embedding of the i^{th} task, $\mathcal{TE}_i$.

Batch Embedding We define batch embeddings $\mathcal{BE}$ as the expected value of the embedding where the probability of each batch is proportional to the volume of the embeddings parallelepiped. This definition of probability is analogous to the one used in traditional DPPs.

$$\mathcal{BE} = \mathcal{BD} \sum_i \pi(t_i) \mathcal{TE}_i$$

where $\pi(.)$ is the distribution derived from normalized task diversity $\mathcal{TD}$. By definition, the batch embeddings $\mathcal{BE}$ have been defined such that the embedding is biased towards the most diverse samplers. To compute the overall diversity of our Sampler, we compute the volume of the parallelepiped spanned by the batch embeddings. However, we make a slight modification, such that the length of each batch embeddings is proportional to the average batch diversity, as defined earlier. This is useful when computing the volume, since we would like samplers that result in high batch diversity to encompass a higher volume.

Our process of computing the volume of parallelepiped spanned by the vectors is discussed in Appendix C in (Kumar, Deleu, and Bengio 2022).

3.2 Definition of Diversity

We define the diversity of the Sampler as the volume of the parallelepiped spanned by the batch embeddings.

$$\mathcal{OD} \propto [\text{vol}(\mathcal{BE})]^2 .$$

With this definition, the volume spanned will be reduced if the Sampler has low diversity within a batch. Furthermore, the batch embeddings would be very similar if the model has low diversity across batches, thus reducing the practical volume spanned.

With the following definition in place, we computed the average batch diversity across five batches, with a batch size of 8 with three different seeds. For samplers such as d-DPP and OHTM, we evaluate on the Protonet model since the embeddings would be similar and in the same latent space as those obtained from the other samplers which use the pre-trained Protonet model. Intuitively, $\mathcal{OD}$ measures the volume the embeddings cover in the latent space. The higher the value, the more volume has been covered in the latent space.

SAMPLER	*Diversity*
NO DIVERSITY TASK SAMPLER	0.00
NO DIVERSITY BATCH SAMPLER	0.00
SINGLE BATCH UNIFORM SAMPLER	0.00
NO DIVERSITY TASKS PER BATCH SAMPLER	$\approx$ 0.00
UNIFORM SAMPLER	1.00
OHTM SAMPLER	1.69
D-DPP SAMPLER	12.40
S-DPP SAMPLER	12.86

Table 1: Overall Diversity of Task Samplers.

The average task diversity on the Omniglot dataset, scaled such that the Uniform Sampler has a diversity of 1, has been reported in Table 1. We confirm and show rigorously that our samplers can be broadly divided into three categories:

- **Low-Diversity Task Samplers:** These samplers include those with an overall diversity score less than 1. These include NDT, NDB, NDTB, and SBU Samplers.
- **Standard Sampler:** This serves as our baseline and is the standard Sampler used in the community - the Uniform Sampler.
- **High-Diversity Task Samplers:** These samplers include those with an overall diversity score greater than 1. These include OHTM, sDPP, and dDPP Samplers.

Furthermore, Our approach does have its advantages over other trivial alternatives such as pairwise-distance metrics. Our proposed formulation is agnostic of the batch size. This property is much desired in the meta-learning setting since the meta-training objectives also work with batch averages. Furthermore, our proposed formulation is more computationally efficient in terms of both time and space than other simpler alternatives. The formulation also offers modularity in its approach, and we can study the diversity at each level, be it tasks, meta-batches, or batches, something not possible with other metrics such as pairwise-distance metrics.

4 Experiments

The experiment aims to answer the following questions: (a) How does limiting task diversity affect meta-learning? (b) Do sophisticated samplers such as OHEM or DPP that improve diversity offer any significant boost in performance? (c) What does our finding imply about the efficacy of the current meta-learning models?

To make an exhaustive study on the effect of task diversity in meta-learning, we train on four datasets: Omniglot (Lake et al. 2011), *mini*Imagenet (Ravi and Larochelle 2017), *tiered*ImageNet (Ren et al. 2018), and Meta-Dataset (Triantafillou et al. 2019). With this selection of datasets, we cover both simple datasets, such as Omniglot and *mini*ImageNet, as well as the most difficult ones, such as *tiered*ImageNet and Meta-Dataset. We train three broad classes of meta-learning models on these datasets: Metric-based (i.e., Protonet (Snell, Swersky, and Zemel 2017), Matching Networks (Vinyals et al. 2016)), Optimization-based (i.e., MAML (Finn, Abbeel, and Levine 2017), Reptile (Nichol, Achiam, and Schulman 2018), and MetaOptNet (Lee et al. 2019)), and Bayesian meta-learning models (i.e., CNAPs (Requeima et al. 2019)). More details about the datasets which were used in our experiments are discussed in Appendix A in (Kumar, Deleu, and Bengio 2022). More details about the models and their hyperparameters are discussed in Appendix B in (Kumar, Deleu, and Bengio 2022). Our source code is made available for additional reference [1].

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 3 in the Appendix in (Kumar, Deleu, and Bengio 2022) 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 performance of the models under this setting. Furthermore, the results on the 20-way 1-shot experiments are presented in Table 4 in the Appendix in (Kumar, Deleu, and Bengio 2022) with higher precision. We also extend the same to the meta-regression setting and observe similar trends as further discussed in Appendix D in (Kumar, Deleu, and Bengio 2022). To further establish our findings, we also present our results on notoriously harder datasets such as *tiered*ImageNet and Meta-Dataset. Figure 6 presents the performance of the models on the *tiered*ImageNet. Again, Table 3 in the Appendix in (Kumar, Deleu, and Bengio 2022) presents the same results with higher precision.

Figure 6 presents the performance of the models on the Meta-Dataset Traffic Sign and Meta-Dataset MSCOCO datasets. We only present the results on Traffic Sign and MSCOCO of the Meta-Dataset, as these two sub-datasets are exclusively used for testing and accurately represent the generalization power of the models when trained with different levels of task diversity. Other results on the Meta-Dataset are presented in Table 5. 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.

[1] https://github.com/RamnathKumar181/Task-Diversity-meta-learning

5 Discussion

From our experiments, we show a similar trend on easy meta-classification tasks (Omniglot and *mini*ImageNet as depicted in Figure 4), as well as harder tasks (*tiered*ImageNet and Meta-Dataset as depicted in Figure 6) in the 5-way 1-shot setting. We also extended our study to the 20-way 1-shot setting with the Omniglot dataset (Figure 5). To test the effect of diversity when the number of shots increases, we turn to the meta-regression domain as depicted in Table 2. Furthermore, to study the effect of diversity in the OOD setting, we turn back to our results on Traffic Sign and MSCOCO datasets from Meta-Dataset (Figure 6). Across all our experiments, we notice a general trend, and we discuss this briefly below.

Disparity between Single Batch Uniform and NDTB Sampler An exciting result is the Disparity between Single Batch Uniform Sampler and No Diversity Tasks per Batch Sampler. 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. One might hypothesize that the Single Batch Uniform Sampler obtains the performance observed by the No Diversity Tasks per Batch Sampler if trained for enough epochs. This scenario has been considered and refuted by our experiments in Appendix D in (Kumar, Deleu, and Bengio 2022).

Difference between "Task Difficulty" and "Task Diversity" Prior works have studied the effects of task difficulty on the performance of the model. Classifying diverse classes would be easier for metric-based networks and harder for optimization-based networks (while testing, it might be difficult to reach very different latent spaces after the inner loop optimization). Thus, the concept of diversity and its connection to the difficulty of the tasks becomes model-dependent and not suitable as a robust metric for analogous understanding. It is important that throughout this work, we do not use the concept of difficulty as a definition or analogy for diversity.

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 observation would suggest that the model trained on only a data fragment can perform similarly to that trained on the Uniform Sampler. This phenomenon questions the improvement/addition the additional amount of data has brought.

Declining performance in d-DPP Methodology The performance may degrade over epochs for d-DPP due to the

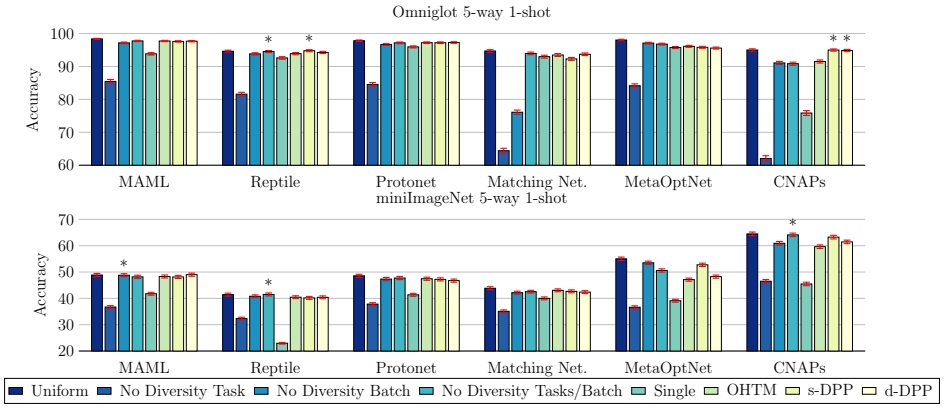


Figure 4: Average accuracy on Omniglot 5-way 1-shot & *mini*ImageNet 5-way 1-shot, with 95% confidence interval. We use the symbol $*$ to represent the instances where the results are not statistically significant (with a p-value $p = 0.05$) and similar to the performance achieved by Uniform Sampler.

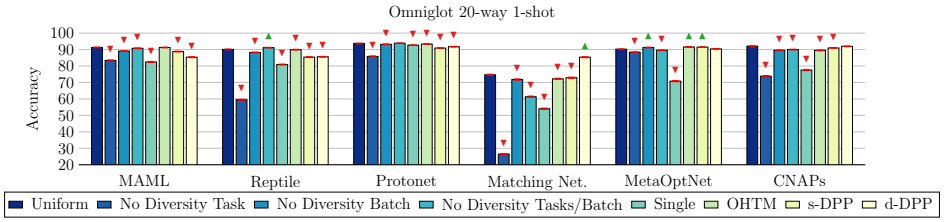


Figure 5: Average accuracy on Omniglot 20-way 1-shot, with a 95% confidence interval. We denote all samplers that are worse than the Uniform Sampler and are statistically significant (with a p-value $p = 0.05$) with ▼, and those that are significantly better than the Uniform Sampler with ▲.

non-stationarity of the sampling process (the DPP gets updated along with the model). This effect may be evident for metric-based methods (matching nets, protonet) since class embeddings directly impact the sampling process through the DPP and the model's performance. One notable exception is MetaOptNet, which explains our highlight in the paper. This behavior hypothesizes that the SVM may be more robust to changes in the embeddings (induced by this non-stationary process) due to the max-margin classifier. We present the convergence graph of the MetaOptNet model on Omniglot 5-way 1-shot run in Figure 14 in the Appendix in (Kumar, Deleu, and Bengio 2022) with an added smoothing factor of 1.

Theoretical Analysis Neural networks in theory are capable of mapping any function given the width of the network is sufficiently large. However, in practice two scenarios could occur derailing the network to a sub-optimal solution: (i) The network is shallow/small and not expressive enough for the optimal solution or (ii) model is expressive enough, but SGD is not able to find the global optima, either due to saddle points, low learning rate etc. Under the assumption that we have a well-defined model, we can intuitively understand why increasing diversity does not help the model better. When the data points are close to each other, the learning of features from one could easily transfer to the other points and achieve a good fit. The diverse data distribution might not be as straightforward since the model would have to learn multiple disjoint concepts to classify these points. This is the crux of Simpson's Paradox. This visualization would be easier to understand in a generic regression setting. We expand on our theoretical analysis further in Appendix D.1 in (Kumar, Deleu, and Bengio 2022).

6 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. A more detailed overview of these methods is discussed in Appendix B.1 in (Kumar, Deleu, and Bengio 2022).

Although research in meta-learning has attracted much attention recently, the effect of task diversity in the domain of meta-learning is still an open question. 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 improve an existing classifier. 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.

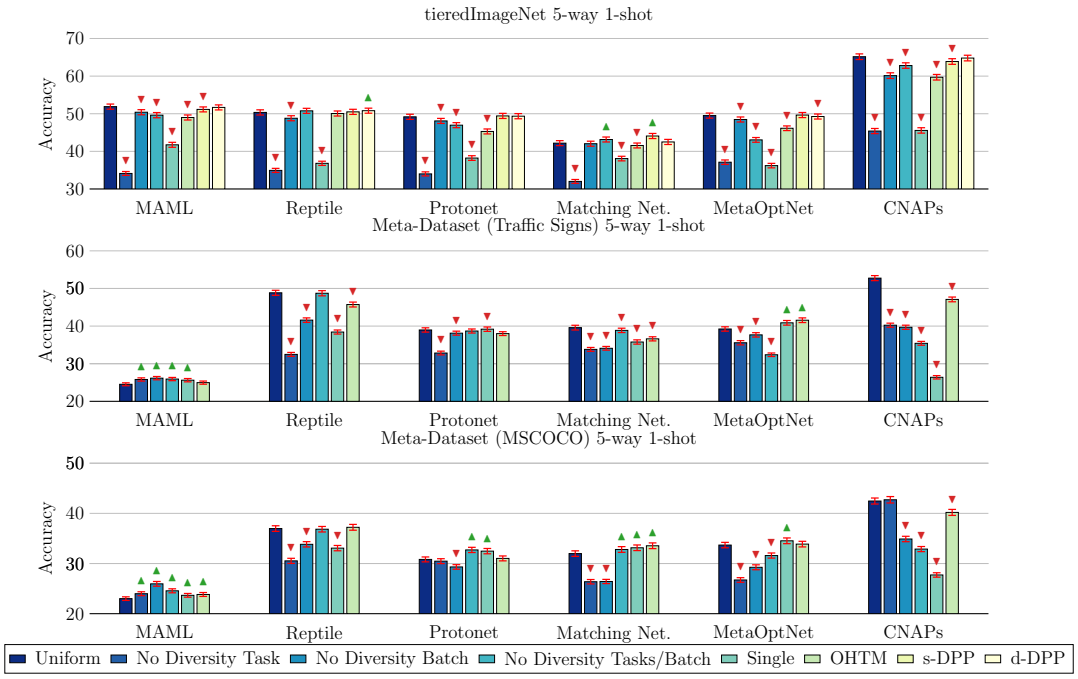


Figure 6: Average accuracy on *tiered*ImageNet 5-way 1-shot, Meta-Dataset Traffic Sign 5-way 1-shot & Meta-Dataset MSCOCO 5-way 1-shot. We denote all samplers that are worse than the Uniform Sampler and are statistically significant (with a p-value $p = 0.05$) with ▼, and those that are significantly better than the Uniform Sampler with ▲.

(Ravi and Larochelle 2018) proposed an approach to sample a batch of samples using a Prototypical Network (Snell, Swersky, and Zemel 2017) as the backbone architecture. The model maximizes the query set, given support set and unlabeled data. Other works such as CACTUs (Hsu, Levine, and Finn 2018), proposed a framework that samples tasks/examples using relatively simple task construction mechanisms such as clustering embeddings. The unsupervised representations learned via these samples perform well 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, (Yin et al. 2019) 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 (Lee et al. 2019; Snell, Swersky, and Zemel 2017) improve the meta-test performance by changing the number of ways or shots of the sampled meta-training tasks, thus increasing the complexity and diversity of the tasks. (Liu et al. 2020) 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, and the Class-based Sampler samples each class independently and does not consider the difficulty of a task as a whole. Our OHTM sampler is similar to the Class-pair based Sampler. (Liu, Chao, and Lin 2020) propose to augment the set of possible tasks by augmenting the predefined set of classes that generate the tasks with varying degrees of rotated inputs as new classes. Closer to our work, (Setlur, Li, and Smith 2020) study a specific sampler by limiting the pool of tasks. Our work, however, has remained fundamentally different, and we expand on this briefly in Appendix D in (Kumar, Deleu, and Bengio 2022). To the best of our knowledge, our work is the first to study the full range of the effect of task diversity in meta-learning.

7 Conclusions and Future Work

In this paper, we have studied the effect of task diversity in meta-learning. We have empirically shown task diversity's effects in the meta-learning domain. We notice two important findings from our research: (i) 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. Our experiments using the NDTB and NDB empirically show that a model trained on even a tiny 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. (ii) We also show that sophisticated samplers such as OHTM 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. We believe that the experiments and task diversity definition we performed and defined lay the roadwork to further research on the effect of task diversity domain in meta-learning and encourage more in-depth studies into the efficacy of our meta-learning methods.

Ethical Statement

Our work studies the effect of task diversity in the meta-learning setting. It helps us understand the efficacy of our models and better design samplers in the future. To the best of our knowledge, this work poses no negative impacts on society.

Acknowledgments

We would like to thank Sony Corporation for funding this research through the Sony Research Award Program. We would also like to thank Dheeraj Mysore Nagaraj from Google Research, India, for his valuable discussions and ideas, which led to the conclusions presented in the theoretical analysis section.

References

Bıyık, E.; Wang, K.; Anari, N.; and Sadigh, D. 2019. Batch active learning using determinantal point processes. *arXiv preprint arXiv:1906.07975*.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C.; and Huang, J.-B. 2019. A Closer Look at Few-shot Classification. In *International Conference on Learning Representations*.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3606–3613.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 1126–1135. PMLR.

Finn, C.; Xu, K.; and Levine, S. 2018. Probabilistic model-agnostic meta-learning. *arXiv preprint arXiv:1806.02817*.

Garnelo, M.; Rosenbaum, D.; Maddison, C.; Ramalho, T.; Saxton, D.; Shanahan, M.; Teh, Y. W.; Rezende, D.; and Eslami, S. A. 2018. Conditional neural processes. In *International Conference on Machine Learning*, 1704–1713. PMLR.

Houben, S.; Stallkamp, J.; Salmen, J.; Schlipsing, M.; and Igel, C. 2013. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In *The 2013 international joint conference on neural networks (IJCNN)*, 1–8. Ieee.

Hsu, K.; Levine, S.; and Finn, C. 2018. Unsupervised learning via meta-learning. *arXiv preprint arXiv:1810.02334*.

Jongejan, J.; Rowley, H.; Kawashima, T.; Kim, J.; and Fox-Gieg, N. 2016. The quick, draw!-ai experiment. *Mount View, CA, accessed Feb*, 17(2018): 4.

Koch, G.; Zemel, R.; Salakhutdinov, R.; et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille.

Kuhn, A.; Aertsen, A.; and Rotter, S. 2003. Higher-order statistics of input ensembles and the response of simple model neurons. *Neural computation*, 15(1): 67–101.

Kulesza, A.; and Taskar, B. 2012. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*.

Kumar, R.; Deleu, T.; and Bengio, Y. 2022. The Effect of Diversity in Meta-Learning. *arXiv preprint arXiv:2201.11775*.

Lacoste, A.; Oreshkin, B.; Chung, W.; Boquet, T.; Rostamzadeh, N.; and Krueger, D. 2018. Uncertainty in multitask transfer learning. *arXiv preprint arXiv:1806.07528*.

Lake, B.; Salakhutdinov, R.; Gross, J.; and Tenenbaum, J. 2011. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10657–10665.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, C.; Wang, Z.; Sahoo, D.; Fang, Y.; Zhang, K.; and Hoi, S. C. 2020. Adaptive Task Sampling for Meta-learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, 752–769. Springer.

Liu, J.; Chao, F.; and Lin, C.-M. 2020. Task augmentation by rotating for meta-learning. *arXiv preprint arXiv:2003.00804*.

Macchi, O. 1975. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1): 83–122.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Munkhdalai, T.; and Yu, H. 2017. Meta networks. In *International Conference on Machine Learning*, 2554–2563. PMLR.

Nichol, A.; Achiam, J.; and Schulman, J. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 722–729. IEEE.

Ravi, S.; and Larochelle, H. 2017. Optimization as a model for few-shot learning. In *International conference on learning representations*.

Ravi, S.; and Larochelle, H. 2018. Meta-Learning for Batch Mode Active Learning. In *International Conference on Learning Representations*.

Ren, M.; Triantafillou, E.; Ravi, S.; Snell, J.; Swersky, K.; Tenenbaum, J. B.; Larochelle, H.; and Zemel, R. S. 2018. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*.

Requeima, J.; Gordon, J.; Bronskill, J.; Nowozin, S.; and Turner, R. E. 2019. Fast and flexible multi-task classification using conditional neural adaptive processes. *Advances in Neural Information Processing Systems*, 32: 7959–7970.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; and Lillicrap, T. 2016. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, 1842–1850. PMLR.

Schroeder, B.; and Cui, Y. 2018. Fgvcx fungi classification challenge 2018. https://github.com/visipedia/fgvcx_fungi_comp. Accessed: 2021-10-29.

Setlur, A.; Li, O.; and Smith, V. 2020. Is Support Set Diversity Necessary for Meta-Learning? *arXiv preprint arXiv:2011.14048*.

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

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1199–1208.

Triantafillou, E.; Zhu, T.; Dumoulin, V.; Lamblin, P.; Evci, U.; Xu, K.; Goroshin, R.; Gelada, C.; Swersky, K.; Manzagol, P.-A.; et al. 2019. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29: 3630–3638.

Yin, M.; Tucker, G.; Zhou, M.; Levine, S.; and Finn, C. 2019. Meta-learning without memorization. *arXiv preprint arXiv:1912.03820*.

Yoon, J.; Kim, T.; Dia, O.; Kim, S.; Bengio, Y.; and Ahn, S. 2018. Bayesian model-agnostic meta-learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 7343–7353.

Zhang, C.; Öztireli, C.; Mandt, S.; and Salvi, G. 2019. Active mini-batch sampling using repulsive point processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5741–5748.