Detailed responses

1. **Performance Concern:**

It seems that the test accuracy of the original MAML method on the 5-way 5-shot setting on the mini-Imagenet dataset is significantly lower (reported 62.71% in Table A5 of [E] but only 60.16% in Table 2 of the revised paper). It also means the proposed method may underperform MAML in this setting (the proposed TA-MAML achieves 62.48% accuracy as shown in Table 2 of the revised paper). As shown in Table 3 of the revised paper, the proposed TA method underperforms the GCP and ATS methods in the 5-shot setting.

Response:

We agree that the accuracies of our baselines (models without attention) denoted by * in Table 2 were lower than those reported in the literature (denoted by # in Table 2). Our careful investigation revealed the cause to be differences in the reported (in the literature) and our experimental setup. Our experimental setup, as discussed below, is more challenging, resulting in lower accuracies for the baselines. The decreased accuracy of the baselines also propagates to their task-attended counterparts. Comparing baseline accuracies reported in the literature and our proposed method (which are under different experimental setups) is unfair. Therefore, we perform an additional set of experiments using the reported setups [B,J,G,C] and show the merit of the proposed approach even on the setups reported in the literature [B,J,G,C]. We also retain the results from the baseline re-runs and the task attended counterparts obtained from our experimental setup. While we present the results in Table 2 of the main paper, we highlight the results obtained for the task-attention framework using the experimental setup reported in the literature for the reviewer's benefit (this was the reviewers' primary concern).

Table 2: Comparison of few-shot classification performance of vanilla ML algorithms with their task attended versions on miniImagenet and tieredImagenet datasets for 5 way (1 and 5 shot) settings on reported experimental setups (denoted by #).

Algorithm	**Test Accuracy**	
	5 way 1 shot	**5 way 5 shot**
miniImagenet		
MAML$^{\#}$ [B]	48.07±1.75	63.15±0.91
CA-MAML$^{\#}$ [L]	47.86±2.50	64.27±1.26
TAML$^{\#}$ [K]	51.77±1.86	65.6±0.93
TA-MAML$^{\#}$	**53.80±1.85**	**66.11±0.11**
MetaSGD$^{\#}$ [J]	50.47±1.87	64.03±0.94
TA-MetaSGD$^{\#}$	**52.60±0.25**	**67.54±0.12**
ANIL$^{\#}$ [C]	46.7±0.4	61.5±0.5

TA-ANIL[#]	**49.53±0.41**	**63.73±0.33**
tieredImagenet		
MAML[#] [G]	47.44±0.18	64.70±0.14
TA-MAML[#]	**51.90±0.19**	**69.43±0.18**

We also note that our approach performs better than GCP and ATS even in 5 shot setup (Table 3 - main paper) when the experimental setup reported in the literature (denoted by #) is used for training and testing our approach. We present the relevant portion of Table 3 below:

Algorithm	**Test Accuracy**	
	5 way 1 shot	**5 way 5 shot**
miniImagenet		
MAML with GCP[#]	46.92 ± 0.83	63.28 ± 0.66
MAML with ATS[#]	47.89 ± 0.77	64.07 ± 0.70
TA-MAML[#] (Ours)	**53.80 ± 1.85**	**66.11 ± 0.11**
MetaSGD with GCP[#]	47.77 ± 0.75	63.50 ±0.71
MetaSGD with ATS[#]	48.59 ± 0.79	64.79 ± 0.74
TA-MetaSGD[#] (Ours)	**52.60 ± 0.25**	**67.54 ± 0.12**

Explanation for the variation in the results:

The literature reports significant variations in the meta-test performances of various ML approaches (Table 7 in supplementary material and presented below). The reported average meta-test accuracies of MAML on the miniImagenet dataset range from 46.47 % to 48.70% (55.16% to 64.39%) for 5 way 1 shot (5 shot) settings. Similarly, for ANIL, meta-test accuracies vary from 46.59 % to 47.82 % (61.5 % to 63.47%) for 5 way 1 shot (5 shot) settings. A careful analysis reveals the different experimental setups resulting in the observed variation. The experimental setups [B,G,C,A,H,I] differ in the number of examples per class in the query set, the number of gradient descent steps in the inner loop, meta-batch size, inductive or transductive batch normalization, etc. Our setup (denoted using *) has the same train and test conditions. Specifically, we set the query examples per class to 15 and gradient steps to 5 for both the meta-train and meta-test phases. However, for 10 way 5 shot setting, we use only 2 gradient steps to reduce the computational burden. More query examples per class (15) during the meta-test provide a robust estimate of the model's generalizability. Further, setting gradient steps to 5 (or 2) can evaluate the quick adaptation capabilities of a learned prior.

Table 7: Variations in the reported performances of MAML and ANIL on miniImagenet dataset on 5 way 1 and 5 shot settings.

Algorithm	**MAML**		**ANIL**	
	5 way 1 shot	**5 way 5 shot**	**5 way 1 shot**	**5 way 5 shot**
Original Papers (Finn et.al [B], Raghu et.al [C])	48.70 ± 1.84%	63.11 ± 0.92%	46.7 ± 0.4%	61.5 ± 0.50%
Antoniou et.al [F]	48.25±0.62%	64.39±0.31%	-	-
Oh et.al [G]	47.44 ± 0.23%	61.75 ± 0.42%	47.82 ± 0.20%	63.04 ± 0.42%
Raghu et.al [C]	46.9 ± 0.2%	63.1 ± 0.4%		
Chen et.al [E]	46.47 ± 0.82%	62.71 ±0.71%	-	-
Arnold et.al [A]	46.88±0.60%	55.16±0.55%	46.59±0.60%	63.47±0.55%
Agarwal et.al [D]	47.13 ± 8.78%	57.69 ± 7.92%		

2. **Motivation concern.**

As mentioned in Question (6), the proposed method only aims to re-weight the tasks with a softmax normalization in each batch, which may be meaningless and problematic. Specifically, there are huge or even infinite tasks in meta learning and different tasks are used in every batch. The normalized weight for one task only represents its importance in the current task batch instead of the global task pool. Thus, this motivation is not much interesting and significant in meta learning. If to re-weight tasks in each batch, many multi-task learning methods like [L, N] can also achieve it, thus it is better to compare with them to show the effectiveness of the proposed TA method.

Response:

We agree with the reviewer that our approach considers the task importance with respect to the current task batch and meta-model only. We are motivated by our hypothesis that the task importance is not only related to the property of the data in the task but also to the property of the current meta-model's configuration. For example, in the initial stage of the meta-training, coarse-grained tasks (tasks composed of semantically distinct classes) may have higher importance than fine-grained tasks (tasks composed of visually similar classes), while this behavior may flip as the training progresses. We agree that global task weighting is an interesting direction that has been studied recently [A, M]. In fact, [A] empirically shows that the hardness or easiness of a task is retained throughout the training of a meta-model. But we have to emphasize the differences in the notion of importance in [A] (overall) and the proposed setup (with respect to the current meta model state). Further, we empirically show that our weighting mechanism imparts better generalizability to the meta-model than the global weighting of the tasks. This is demonstrated in the Tables 3, 4 and 5 (main paper) for various algorithms (MAML, ANIL, MetaSGD), under different few-shot settings (5.1, 5.5), datasets (miniImagenet, miniImagenet noisy) and dataset properties (In distribution, Noisy distribution, and Cross-domain). Thus, we believe that the approach and the results are relevant to the meta-learning community.

We thank the reviewer for drawing our attention to similar literature in allied areas. [L] proposed an optimization method to neutralize conflicts of an average model with individual tasks in multi-task learning. Specifically, they find an optimal update vector that lies within the proximity of the average gradient across the batch of the tasks without conflicting with any task gradient. We first note the subtle difference between the multi-task and meta-learning setup. In multi-task learning, the (meta) train and (meta) test tasks are similar (have the same classes), and the aim is to learn a model that performs well on all the tasks in a training batch (consequently testing batch). On the other hand, meta-learning aims to improve the model's generalizability to unseen meta-test tasks (classes) from the same distribution. Thus, the uniform performance of each task in the training batch is not critical for a meta-learning setup. We hypothesize some tasks may contribute more to the meta-model's knowledge than others, depending on the stage of the training. Consider the stage when the model has already acquired generic knowledge, i.e., distinctive classes like a pen, ball, and a lion could be easily differentiated. A task with such distinctive classes will not contribute significantly to the meta-model's learning at this stage. However, a task containing comparable classes, like three breeds of black dogs, is more challenging and thus contribute more to the meta-model's learning (Fig 9 supplementary).

The idea of constraining the gradients of tasks in a batch to an average gradient [L] is similar to the baseline (TAML) [K] that we had considered. Finding an update vector in [L] is an unconstrained optimization that is hard to solve. To limit the search space, the update vector is constrained to be close to the average gradient vector on a task batch using a hyper-parameter. Similarly, in [K] the loss of each task is pushed towards the average loss of a task batch using a hyper-parameter. Penalizing the tasks that improve the model's generalizability towards average gradient [L] or loss [K] on task-batch may circumvent their knowledge flow to the meta-model. The task models are not penalized in our setup. Penalizing tasks imparts equity among the task models and thus we hypothesize that [L] and [K] are more suitable to the multi-task setup rather than a meta-learning setup. As [L] is a recent approach that outperforms [N] on a multi-task learning setup, we compare our approach with only to [L]. Specifically, we extend [L] to a meta-learning setup by computing the average and weighted average gradients on query loss of the adapted models instead of a model from the previous iteration (as in a multi-task setup). However, [L] and [K] are more computationally feasible as they do not require the training of an additional attention network. Table 2 in the main paper and the relevant portion highlighted below demonstrates that the proposed attention mechanism has better generalizability to unseen tasks than conflict-averse gradient descent adapted for the meta-learning setup (CA-MAML).

Table 2: Comparison of few-shot classification performance of MAML [B], CA-MAML [L], TAML [K], and TA-MAML on miniImagenet dataset for 5 way (1 and 5 shot) settings on reported experimental setups (denoted by #).

Algorithm	**Test Accuracy**	
	5 way 1 shot	**5 way 5 shot**
miniImagenet		
MAML# [B]	48.07±1.75	63.15±0.91
CA-MAML# [L]	47.86±2.50	64.27±1.26
TAML# [K]	51.77±1.86	65.6±0.93
TA-MAML#	**53.80±1.85**	**66.11±0.11**

3. **Efficiency concern.**

As mentioned in Question (9), the proposed method may bring huge computational costs, and at least the authors should clarify this and report the training time of the proposed method.

Response:

The training time for all scheduling /sampling approaches is expected to be higher than their non-scheduling/sampling counterparts. We observe a three-fold increase in the training time from the vanilla setting for a model trained with our strategy and a two-fold increase in the training time if a non-neural scheduling approach [L] is employed. However, our approach significantly outperforms vanilla ML approaches and all state-of-the-art scheduling strategies on various datasets, training setups, and learning paradigms (Tables 2, 3, 4 and 5 - main paper). As training is typically performed offline, the increased computational overhead is expected to be permissible. Further, ours, as well as other approaches, perform vanilla finetuning during meta-testing (i.e., task attention, neural scheduling or conflict resolving mechanism is not employed during meta-testing), resulting in comparable test time (15-20 seconds on 300 tasks for MAML 5-way 1- and 5-shot setups). We also note that we do not pre-train the attention network, unlike state-of-the-art schedulers like ATS [M].

[A] Arnold et.al. Uniform Sampling over Episode Difficulty. NeurIPS, 2021.

[B] Finn et.al. Model-agnostic meta-learning for fast adaptation of deep networks. In ICML, 2017

[C] Raghu et.al. Rapid Learning or Feature Reuse? Towards Understanding the Effectiveness of MAML. In ICLR, 2019.

[D] Agarwal et.al. On sensitivity of meta-learning to support data. In NeurIPS 2021.

[E] Chen et.al. A Closer Look at Few-shot Classification. In ICLR, 2018.

[F] Antoniou et.al. How to train your MAML. In ICLR, 2019.

[G] Oh et.al. BOIL: Towards Representation Change for Few-shot Learning. In ICLR, 2020.

[H] Sun et al. Meta-transfer learning through hard tasks. In TPAMI, 2020.

[I] Oreshkin et al. Tadam: Task dependent adaptive metric for improved few-shot learning. In NeurIPS, 2018.

[J] Li et.al. Meta-sgd: Learning to learn quickly for few-shot learning, In arXiv 2017

[K] Jamal et.al. Task agnostic meta-learning for few-shot learning. In CVPR, 2019.

[L] Liu et.al. Conflict-averse gradient descent for multi-task learning. In NeurIPS, 2021.

[M] Yao et.al. Meta-learning with an Adaptive Task Scheduler. NeurIPS, 2021.

[N] Sener et al. Multi-Task Learning as Multi-Objective Optimization. NeurIPS, 2018.

Not All Tasks are Equal - Task Attended Meta-learning for Few-shot Learning

Anonymous authors
Paper under double-blind review

Abstract

Meta-learning (ML) has emerged as a promising direction in learning models under constrained resource settings like few-shot learning. The popular approaches for ML either learn a generalizable initial model or a generic parametric optimizer through batch episodic training. In this work, we study the importance of tasks in a batch for ML. We hypothesize that the common assumption in batch episodic training where each task in a batch has an equal contribution to learning an optimal meta-model need not be true. We propose to weight the tasks in a batch according to their "importance" in improving the meta-model's learning. To this end, we introduce a training curriculum called task attended meta-training to learn a meta-model from weighted tasks in a batch. The task attention module is a standalone unit and can be integrated with any batch episodic training regimen. Comparison of task-attended ML models with their non-task-attended counterparts on complex datasets, performance improvement of proposed curriculum over state-of-the-art task scheduling algorithms on noisy datasets, and cross-domain few shot learning setup validate its effectiveness.

1 Introduction

The ability to infer knowledge and discover complex representations from data has made deep learning models widely popular in the machine learning community. However, these models are data-hungry, often requiring large volumes of labeled data for training. Collection and annotation of such large amounts of training data may not be feasible for many real life applications, especially in domains that are inherently data constrained, like medical and satellite image classification, drug toxicity estimation, etc. Meta-learning (ML) has emerged as a promising direction for learning models in such settings, where only a limited amount (few-shots) of labeled training data is available. A typical ML algorithm employs an episodic training regimen that differs from the training procedure of conventional learning tasks. This episodic meta-training regimen is backed by the assumption that a machine learning model quickly generalizes to novel unseen data with minimal fine-tuning when trained and tested under similar circumstances (Vinyals et al., 2016). To facilitate such a generalization capacity, a meta-training phase is undertaken, where the model is trained to optimize its performance on several homogeneous tasks/episodes randomly sampled from a dataset. Each episode or task is a learning problem in itself. In the few-shot setting each task is a classification problem, a collection of K support (train) and Q query (test) samples corresponding to each of the N classes. Task-specific knowledge is learned using the support data, and meta-knowledge across the tasks is learned using query samples, which essentially encodes "how to learn a new task effectively." The learned meta-knowledge is generic and agnostic to tasks from the same distribution. It is typically characterized in two different forms - either as an optimal initialization for the machine learning model or a learned parametric optimizer. Under the optimal initialization view, the learned meta-knowledge represents an optimal prior over the model parameters, that is equidistant, but close to the optimal parameters for all individual tasks. This enables the model to rapidly adapt to unseen tasks from the same distribution (Finn et al., 2017; Li et al., 2017; Jamal & Qi, 2019). Under the parametric optimizer view, meta-knowledge pertaining to the traversal of the loss surface of tasks is learned by the meta-optimizer. Through learning task specific and task agnostic characteristics of the loss surface, a parametric optimizer can thus effectively guide the base model to traverse the loss surface and achieve superior performance on unseen tasks from the same distribution (Ravi & Larochelle, 2017).

Initialization based ML approaches accumulate the meta-knowledge by simultaneously optimizing over a batch of tasks. On the other hand, a parametric optimizer sequentially accumulates meta-knowledge across individual tasks. The sequential accumulation process leads to a long oscillatory optimization trajectory and a bias towards the last task, limiting the parametric optimizer's task agnostic potential. However, recently meta-knowledge has been accumulated in a batch mode even for the parametric optimizer (Aimen et al., 2021). Further, under such batch episodic training (for both initialization and optimization views), a common assumption in ML that the randomly sampled episodes of a batch contribute equally to improving the learned meta-knowledge need not hold good. Due to the latent properties of the sampled tasks in a batch and the model configuration, some tasks may be better aligned with the optimal meta-knowledge than others. We hypothesize that proportioning the contribution of a task as per its alignment towards the optimal meta-knowledge can improve the meta-model's learning. This is analogous to classical machine learning algorithms like sample re-weighting, which however, operate at sample granularity. In re-weighting, samples leading to false positives are prioritized and therefore replayed. Hence, the latent properties due to which a sample is prioritized are explicitly defined. For complex task distributions, explicitly handcrafting the notion of "importance" of a task would be hard. To this end, we propose a task attended meta-training curriculum that employs an attention module that learns to assign weights to the tasks of a batch with experience. The attention module is parametrized as a neural network that takes meta-information in terms of the model's performance on the tasks in a batch as input and learns to associate weights to each of the tasks according to their contribution in improving the meta-model. Overall, we make the following contributions,

- We propose a task attended meta-training strategy wherein different tasks of a batch are weighted according to their "importance" defined by the attention module. This attention module is a standalone unit that can be integrated into any batch episodic training regimen.
- We extend the empirical investigation of the batch-mode parametric optimizer (MetaLSTM++) to complex datasets like miniImagenet, FC100, and tieredImagenet and validate its efficiency over its sequential counter-part (MetaLSTM).
- We conduct extensive experiments on miniImagenet, FC100, and tieredImagenet datasets and compare ML algorithms like MAML, MetaSGD, ANIL, and MetaLSTM++ with their task-attended counterparts to validate the effectiveness of the task attention module and its coupling with any batch episodic training regimen.
- We compare the proposed training curriculum with task-disagreement resolving approaches like TAML (Jamal & Qi, 2019) and conflict-averse gradient descent (Liu et al., 2021a) and validate the goodness of the proposed hypothesis. We extend these task-disagreement based approaches to the meta-learning regimen for a fair comparison.
- We further compare task-attended curriculum with state-of-the-art task scheduling approaches and also show the merit of the proposed approach on the miniImagenet-noisy dataset and cross-domain few shot learning (CDFSL) setup.
- We perform exhaustive empirical analysis and visual inspections to decipher the working of the task attention module.

2 Related Work

ML literature is profoundly diverse and may broadly be classified into *initialization* (Finn et al., 2017; Li et al., 2017; Jamal & Qi, 2019; Raghu et al., 2020; Rusu et al., 2019; Sun et al., 2019) and *optimization approaches* (Ravi & Larochelle, 2017) depending on the metaknowledge. However, these approaches assume uniform contribution of tasks in learning a meta-model. In supervised learning, assigning non-uniform priorities to the samples is not new (Kahn & Marshall, 1953; Shrivastava et al., 2016). Self-paced learning (Kumar et al., 2010) and hard example mining (Shrivastava et al., 2016) have popularly been used to reweight the samples and various attributes like losses, gradients, and uncertainty have been used to assign priorities to samples (Lin et al., 2017; Zhao & Zhang, 2015; Chang et al., 2017). Zhao & Zhang (2015) introduce importance

sampling to reduce variance and improve the convergence rate of stochastic optimization algorithms over uniform sampling. They theoretically prove that the reduction in the variance is possible if the sampling distribution depends on the norm of the gradients of the loss function. Chang et al. (2017) conclude that mini-batch SGD for classification is improved by emphasizing the uncertain examples. Lin et al. (2017) propose reshaped cross-entropy loss (focal loss) that down-weights the loss of confidently classified samples. Nevertheless, assigning non-uniform priorities to tasks in meta-learning is under-explored and has recently drawn attention (Kaddour et al., 2020; Gutierrez & Leonetti, 2020; Liu et al., 2020; Yao et al., 2021; Arnold et al., 2021). Gutierrez & Leonetti (2020) propose Information-Theoretic Task Selection (ITTS) algorithm to filter training tasks that are distinct from each other and close to the tasks of the target distribution. This algorithm results in a smaller pool of training tasks. A model trained on the smaller subset learns better than the one trained on the original set. On the other hand, Kaddour et al. (2020) propose probabilistic active meta-learning (PAML) that learns probabilistic task embeddings. Scores are assigned to these embeddings to select the next task presented to the model. These algorithms are, however, specific to meta-reinforcement learning (meta-RL). On the contrary, our focus is on the few shot classification problem. Liu et al. (2020) propose a greedy class-pair potential-based adaptive task sampling strategy wherein task selection depends on the difficulty of all class-pairs in a task. This sampling technique is static and operates at a class granularity. On the other hand, our approach is dynamic and operates at a task granularity. Assigning non-uniform weights to samples prevents overfitting on corrupt data points (Ren et al., 2018b; Jiang et al., 2018). Ren et al. (2018b) used gradient directions to re-weight the data points, and Jiang et al. (2018) learned a curriculum on examples using a mentor network. However, these approaches assume availability of abundant labeled data. Yao et al. (2021) extend Jiang et al. (2018) to the few-shot learning setup. They propose an adaptive task scheduler (ATS) to predict the sampling probability of tasks from a candidate pool containing a subset of tasks sampled from the original (noisy or imbalanced) task distribution (similar to (Jiang et al., 2018). Thus, the sampling probabilities of the tasks are (approximately) global. Another global task sampling approach is Uniform Sampling (Arnold et al., 2021), built on the premise that task difficulty (defined as the negative log-likelihood of the model on the task) approximately follows a normal distribution and is transferred across model parameters during training. They also find sampling uniformly over episode difficulty outperforms other sampling schemes like curriculum, easy and hard mining. Our work is different from these approaches (ATS and Uniform Sampling) as we do not propose a global task sampling strategy but a dynamic task-batch re-weighting mechanism for the current meta-model update. We hypothesize that the task's importance depends on the data contained in it and the current meta-model's configuration. For example, in the initial stage of the meta-models training, coarse-grained tasks (tasks composed of semantically distinct classes) may have higher importance than fine-grained tasks (tasks composed of comparable classes), while this behavior may reverse as the training progresses. Further, our approach differs from Uniform Sampling in the definition of task difficulty, i.e., we neither explicitly handcraft the notion of task difficulty nor assume a normal distribution over it. Instead, we let an attention network learn the suitable weights for the tasks in a batch. Although ATS also dynamically learns the task sampling priority, it maintains a candidate pool to satisfy the global task priority criteria, causing overhead. Further, it performs an additional warm start to the scheduler, utilizes more task batches in a run, and uses REINFORCE for reward estimation; therefore, it is more expensive than the proposed approach. Contrary to our idea is TAML (Jamal & Qi, 2019) - a meta-training curriculum that enforces equity across the tasks in a batch. We show that weighting the tasks according to their "importance" and hence utilizing the diversity present in a batch given the meta-model's current configuration offers better performance than enforcing equity in a batch of tasks.

3 Preliminary

In a typical ML setting, the principal dataset $\mathcal{D}$ is divided into disjoint meta-sets $\mathcal{M}$ (meta-train set), $\mathcal{M}_v$ (meta-validation set) and $\mathcal{M}_t$ (meta-test set) for training the model, tuning its hyperparameters and evaluating its performance, respectively. Every meta-set is a collection of tasks $\mathcal{T}$ drawn from the joint task distribution $P(\mathcal{T})$ where each task $\mathcal{T}_i$ consists of support set $D_i = \{(x_k^c, y_k^c)_{k=1}^K\}_{c=1}^N$ and query set $D_i^* = \{(x_q^{*c}, y_q^{*c})_{q=1}^Q\}_{c=1}^N$. Here (x, y) represents a (sample, label) pair and N is the number of classes, K and Q are the number of samples belonging to each class in the support and query set, respectively. According

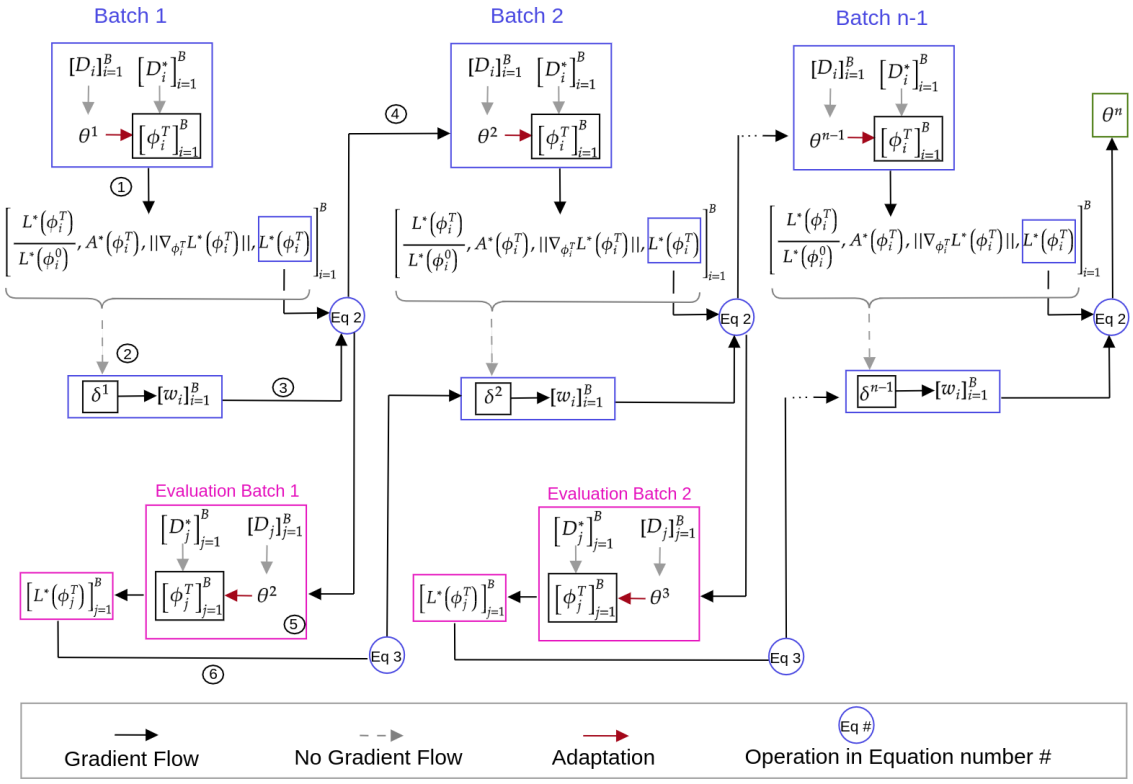

Figure 1: Computational Graph of the forward pass of the meta-model using task attended meta-training curriculum. The output of this procedure is a meta-model θ^n. Gradients are propagated through solid lines and restricted through dashed lines.

to support-query characterization $\mathcal{M}$, $\mathcal{M}_v$ and $\mathcal{M}_t$ could be represented as $\{(D_i, D_i^*)\}_{i=1}^{M}$, $\{(D_i, D_i^*)\}_{i=1}^{R}$, $\{(D_i, D_i^*)\}_{i=1}^{S}$ where M, R and S are the total number of tasks in $\mathcal{M}$, $\mathcal{M}_v$ and $\mathcal{M}_t$ respectively. During meta-training, meta-model θ is adapted on D_i of all tasks in a batch $\{\mathcal{T}_i\}_{i=1}^{B}$ of size B, T times to obtain ϕ_i^T. The adaptation occurs through gradient descent or parametric update on the train loss L using learning rate α. The adapted model ϕ_i^T is then evaluated on D_i^* to obtain test loss L^*, which along with learning rate β, is used to update θ. The output of this episodic training is either an optimal prior or a parametric optimizer, both aiming to facilitate the rapid adaptation of the model on unseen tasks from $\mathcal{M}_t$. The detailed note on initialization and optimization approaches is deferred to the supplementary material.

4 Task Attention in Meta-learning

A common assumption under the batch-wise episodic training regimen adopted by ML is that each task in a batch has an equal contribution in improving the learned meta-knowledge. However, this need not always be true. It is likely that given the current configuration of the meta-model, some tasks may be more important for the meta-model's learning. A contributing factor to this difference is that tasks sampled from complex data distributions can be profoundly diverse. The diversity and latent properties of the tasks coupled with the model configuration may induce some tasks to be better aligned with the optimal meta-knowledge than others. The challenging aspect in the meta-learning setting is to define the "importance" and associate weights to the tasks of a batch proportional to their contribution to improving the meta-knowledge. As human beings, we *learn* to associate importance to events subjective to meta-information about the events and prior experience. This motivates us to define a learnable module that can map the meta-information of tasks to their importance weights.

4.1 Characteristics of Meta-Information

Given a task-batch $\{\mathcal{T}_i\}_{i=1}^{B}$, the task attention module takes as input meta-information about each task ($\mathcal{T}_i$) in the batch, defined as the four tuple below:

$$\mathcal{I} = \left\{ \left(||\nabla_{\phi_i^T} L^*(\phi_i^T)||, L^*(\phi_i^T), A^*(\phi_i^T), \frac{L^*(\phi_i^T)}{L^*(\phi_i^0)} \right) \right\}_{i=1}^{B} \tag{1}$$

where corresponding to each task i in the batch $||\nabla_{\phi_i^T} L^*(\phi_i^T)||$ denotes the norm of gradient, $L^*(\phi_i^T)$ and $A^*(\phi_i^T)$ are the test loss and accuracy of the adapted model respectively, and $\frac{L^*(\phi_i^T)}{L^*(\phi_i^0)}$ is the ratio of the model's test loss post and prior adaptation.

4.1.1 Gradient Norm

Let $P = \left\{\phi_i^T\right\}_{i=1}^{B}$ be the parameters of the models obtained after adapting the initial model (for T iterations) on the support data $\{D_i\}_{i=1}^{B}$ of tasks $\{\mathcal{T}_i\}_{i=1}^{B}$. Also, let $G = \left\{\nabla_{\phi_i^T} L^*(\phi_i^T)\right\}_{i=1}^{B}$ be the gradients of the adapted model parameters w.r.t the query losses $\{L^*(\phi_i^T)\}_{i=1}^{B}$. The gradient norm $\left\{||\nabla_{\phi_i^T} L^*(\phi_i^T)||\right\}_{i=1}^{B}$ is the L_2 norm of the gradients and quantifies the magnitude of the consolidated displacement of the adapted model parameters during a gradient descent update on query data. Larger gradient norm on query dataset could indicate that the model has either not learned the support set or has overfitted. Hence the model is not generalizable on query set compared to the models with low gradient norm. Gradient norm, therefore, carries information about the convergence and generalizability of the adapted models which has been theoretically studied in (Li et al., 2019).

4.1.2 Test Loss

$\{L^*(\phi_i^T)\}_{i=1}^{B}$ represents the empirical error (cross entropy loss) of the adapted base models on unseen query instances and hence characterizes their generalizability. Unlike gradient norm, which characterizes the generalizability in parameter space, query loss quantifies generalizability in the output space as the divergence between the real and predicted probability distributions. As $\{L^*(\phi_i^T)\}_{i=1}^{B}$ is a key component in the meta-update equation, it is an important factor influencing the meta-model's learning. Further, test errors of classes have been widely used to determine their "easy or hardness" (Bengio et al., 2009; Liu et al., 2021b; Arnold et al., 2021). Thus $\{L^*(\phi_i^T)\}_{i=1}^{B}$ acquaints the attention module with the generalizability aspect of task models and their influence in updating the meta-model.

4.1.3 Test Accuracy

$\{A^*(\phi_i^T)\}_{i=1}^{B}$ corresponds to the accuracies of $\{\phi_i^T\}_{i=1}^{B}$ on $\{D_i^*\}_{i=1}^{B}$ scaled in the range [0,1]. $A^*(\phi_i^T)$ evaluates the thresholded predictions (predicted labels) unlike $L^*(\phi_i^T)$, which evaluates the confidence of the model's predictions on the true class labels. Two task models may predict the same class labels but differ in the confidence of the predictions. In such scenarios, neither loss nor accuracy is individually sufficient to comprehend this relationship among the tasks. So, the combination of these two entities is more reflective of the nature of the learned task models.

4.1.4 Loss-ratio

Let $L^*(\phi_i^0)$ be the loss of θ on the D_i^*, and $L^*(\phi_i^T)$ be the loss of the adapted model ϕ_i^T on D_i^*. The loss-ratio $\frac{L^*(\phi_i^T)}{L^*(\phi_i^0)}$ is representative of the relative progress of a meta-model on each task. Higher values (> 1) of the loss-ratio suggests adapting θ to D_i has an adverse effect on generalizing it to D_i^* (negative impact), while lower values (< 1) of the loss-ratio indicates the benefit of adaptation of θ on D_i (positive impact). Loss-ratio of exactly one signifies adaptation attributes to no additional benefit (neutral impact). Therefore, loss-ratio provides information regarding the impact of adaptation on each task for a given meta-model.

4.2 Task Attention Module

We learn a task attention module parameterized by δ, which attends to the tasks that contribute more to the model's learning i.e., the objective of the task attention module is to learn the relative importance of each task in the batch for the meta-model's learning. Thus the output of the module is a $B-$dimensional vector $\mathbf{w} = [w_1, \ldots, w_B]$, ($\sum_{i=1}^{B} w_i = 1$ and $\forall \mathcal{T}_i,\ w_i \geq 0$) quantifying the attention-score (weight - w_i) for each task. The attention vector $\mathbf{w}$ is multiplied with the corresponding task losses of the adapted models $L^*(\phi_i^T)$ on the held-out datasets D_i^* to update the meta-model θ:

$$\theta^{t+1} \leftarrow \theta^t - \beta \nabla_{\theta^t} \sum_{i=1}^{B} w_i L^*(\phi_i^T) \tag{2}$$

After the meta-model is updated using the weighted task losses, we evaluate the goodness of the generated attention weights. We sample a new batch of tasks $\{D_j, D_j^*\}_{j=1}^{B}$ and adapt a base-model ϕ_j using the updated meta-model θ^{t+1} on the train data $\{D_j\}$ of each task. The mean test-loss of the adapted models $\{\phi_j^T\}_{j=1}^{B}$ reflect the goodness of the weights assigned by the attention-module in the previous iteration. The attention module δ is thus updated using the gradients flowing back into it w.r.t to this mean test-loss. The attention network is trained simultaneously with the meta-model in an end to end fashion using the update rule:

$$\delta^{t+1} \leftarrow \delta^t - \gamma \nabla_{\delta^t} \sum_{j=1}^{B} L^*(\phi_j^T) \tag{3}$$

where ϕ_j^T is adapted from θ^{t+1} and γ is the learning rate .

Algorithm 1: Task Attended Meta-Training

Input:
Dataset: $\mathcal{M} = \{D_i, D_i^*\}_{i=1}^{M}$
Models: Meta-model θ, Base-model ϕ, Att-module δ
Learning-rates: α, β, γ
Parameters: Iterations n_{iter}, Batch-size B, Adaptation-steps T
Output: Meta-model θ

1 **Initialization:** $\theta, \delta \leftarrow$ Random Initialization
2 **for** *iteration in* n_{iter} **do**
3 $\{\mathcal{T}_i\}_{i=1}^{B} = \{D_i, D_i^*\}_{i=1}^{B} \leftarrow$ Sample task-batch($\mathcal{M}$)
4 **for** *all* $\mathcal{T}_i$ **do**
5 $\phi_i^0 \leftarrow \theta$
6 $L^*(\phi_i^0), \_ \leftarrow evaluate(\phi_i^0, D_i^*)$ ▷ Compute loss and accuracy of input model on given dataset.
7 $\phi_i^T = adapt(\phi_i^0, D_i)$
8 $L^*(\phi_i^T), A^*(\phi_i^T) \leftarrow evaluate(\phi_i^T, D_i^*)$
9 **end**
10 $[w_i]_{i=1}^{B} \leftarrow Att\_module \left(\left[\frac{L^*(\phi_i^T)}{L^*(\phi_i^0)}, A^*(\phi_i^T), ||\nabla_{\phi_i^T} L^*(\phi_i^T)||, L^*(\phi_i^T) \right]_{i=1}^{B} \right)$
11 $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{i=1}^{B} w_i L^*(\phi_i^T)$
12 $\{D_j, D_j^*\}_{j=1}^{B} \leftarrow$ Sample task-batch($\mathcal{M}$)
13 **for** *all* $\mathcal{T}_j$ **do**
14 $\phi_j^0 \leftarrow \theta$
15 $\phi_j^T = adapt(\phi_j^0, D_j)$
16 **end**
17 $\delta \leftarrow \delta - \gamma \nabla_\delta \sum_{j=1}^{B} L^*(\phi_j^T)$
18 **end**
19 **Return** θ
20 **Function** `adapt`(ϕ_i^t, D_i)**:**
21 $\theta \leftarrow \phi_i^t$
22 **if** θ *is optimal-initialization* **then**
23 **for** *t=1 to T* **do**
24 $\phi_i^{t+1} \leftarrow \phi_i^t - \alpha \nabla_{\phi_i^t} L(\phi_i^t)$
25 **end**
26 **end**
27 **else if** θ *is parametric-optimizer* **then**
28 **for** *t=1 to T* **do**
29 $\phi_i^{t+1} \leftarrow \theta\left(L(\phi_i^t), \nabla_{\phi_i^t} L(\phi_i^t)\right)$ ▷ Parameter updates given by cell state of θ.
30 **end**
31 **end**
32 ***Return*** ϕ_i^T

4.3 Task Attended Meta-Training Algorithm

We demonstrate the meta-training curriculum using the proposed task attention in Figure 1 and formally summarize it in Algorithm 1. As with the classical meta-training process, we first sample a batch of tasks from the task distribution. For each task $\mathcal{T}_i$, we adapt the base-model ϕ_i using the train data D_i for T time-steps (line 7 and lines 20-32 in Algorithm 1). Specifically, for initialization approaches, adaptation is performed by gradient descent on train loss L (lines 22-26 in Algorithm 1). However, for optimization approaches, current loss and gradients are inputted to the meta-model θ, which outputs the updated base-model parameters (lines 27-31 in Algorithm 1). Then we compute the meta-information about the adapted model corresponding to each task. It comprises of the loss $L^*(\phi_i^T)$, accuracy

$A^*(\phi_i^T)$, loss-ratio $\frac{L^*(\phi_i^T)}{L^*(\phi_i^0)}$ and gradient norm $||\nabla_{\phi_i^T} L^*(\phi_i^T)||$ on the test data D_i^*. This meta-information corresponding to each task in a batch is given as input to the task attention module (Figure 1 - Label: ②) which outputs the attention vector (line 10 in Algorithm 1). The attention vector and test losses $\{L^*(\phi_i^T)\}_{i=1}^B$ are used to update meta-model parameters θ according to equation 2 (line 11 in Algorithm 1, Figure 1 - Label: ④). We sample a new batch of tasks $\{D_j, D_j^*\}_{j=1}^B$ and adapt the base-models $\{\phi_j^T\}_{j=1}^B$ using the updated meta-model (lines 12-16 in Algorithm 1, Figure 1 - Label: ⑤). We compute the mean test loss over the adapted base-models $\{L^*(\phi_j^T)\}_{j=1}^B$, which is then used to update the parameters of the task attention module δ according to equation 3 (line 17 in Algorithm 1, Figure 1 - Label: ⑥).

The attention network is designed as a stand-alone module to learn the mapping from the meta-information space to the importance of tasks in a batch. The meta-model is learned according to equation 2 and aims to minimize the weighted loss. It is important to decouple the learning of the attention network from that of the meta-model. If there is information flow from the task attention module to the meta-model, the latter may reduce its weighted loss by learning an initialization that is suboptimal, but for which the task attention network assigns lower weights. This would introduce an undesirable bias to the learning process. To circumvent this bias, we restrict the flow of gradients to the meta-model θ through the task attention module δ by enforcing $\nabla_\theta w_i L^*(\phi_i^T) = w_i \nabla_\theta L^*(\phi_i^T)$ i.e., $\nabla_\theta w_i$ is not computed. Also, gradients flowing through the attention network to the meta-model create additional computational overhead. Specifically, the term $\nabla_\theta \sum_i w_i L^*(\phi_i^T)$ from equation 2 can be expanded as follows -

$$\nabla_\theta \sum_i w_i L^*(\phi_i^T) = \sum_i \nabla_\theta w_i L^*(\phi_i^T) = \underbrace{\sum_i w_i \nabla_\theta L^*(\phi_i^T)}_{\text{Term 1}} + \underbrace{\sum_i L^*(\phi_i^T) \nabla_\theta w_i}_{\text{Term 2}}$$

The $\nabla_\theta w_i$ in Term 2 is computationally expensive as $\nabla_\theta w_i = \nabla_\delta w_i . \nabla_I \delta . \nabla_\phi I . \nabla_\theta \phi$. Restricting the gradient flow avoids these additional computations. We also note that the meta-model and attention network are updated only once during each training iteration, although on different batches of tasks.

5 Experiments and Results

We conduct experiments to demonstrate the merit of the task-attention across multiple datasets, training setups, and learning paradigms. We verify that the proposed regimen could be integrated with various ML approaches like MAML, MetaSGD, MetaLSTM++, and ANIL and further show its superiority over state-of-the-art task-scheduling and conflict-resolving approaches. We also analyze the attention network.

5.1 Dataset and Implementation Details

In line with the state-of-the-art literature (Sun et al., 2020; Arnold et al., 2021), we use miniImagenet, FC100, and tieredImagenet for evaluating the effectiveness of the proposed attention module as they are more challenging datasets comprising of highly diverse tasks. We also test the efficacy of the proposed approach on noisy dataset (miniImagenet-noisy), and under cross-domain few shot learning (CDFSL) miniImagenet → CUB-200 and mini-Imagenet → FGVC-Aircrafts datasets. The details of the datasets are presented in the supplementary material.

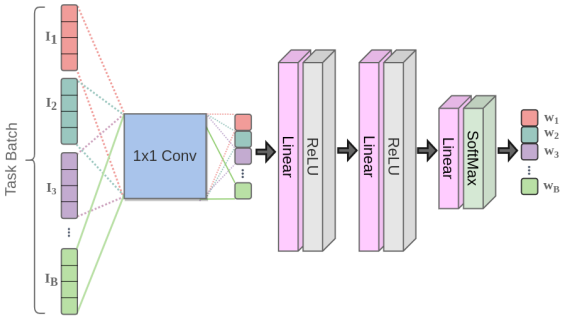

Figure 2: Architecture of Task-attention module.

We use a 4-layer CNN from (Finn et al., 2017) as a base model and a two-layer LSTM (Ravi & Larochelle, 2017) for the parametric optimizer. The architecture of the task-attention module is illustrated in Figure 2 and described as follows.The task attention module is implemented as a 4-layer neural network. The first layer

performs a 1×1 convolution over the input (meta-information) of size B×4 where B denotes the meta-batch size, producing a vector of size B×1 as output. This vector is then passed through two fully connected layers with 32 hidden nodes, each followed by a ReLU activation. This output is then passed through a fully connected layer with B nodes, followed by a softmax activation to produce the normalized attention weights.

We perform a grid search over 30 different configurations for 5000 iterations to find the optimal hyper-parameters for each setting. The search space is shared across all meta-training algorithms and datasets. The meta, base and attention model learning rates are sampled from a log uniform distribution in the ranges $\left[1e^{-4}, 1e^{-2}\right]$, $\left[1e^{-2}, 5e^{-1}\right]$ and $\left[1e^{-4}, 1e^{-2}\right]$ respectively (see appendix for more details). The hyperparameter λ for TAML (Theil) is sampled from a log uniform distribution over the range of $\left[1e^{-2}, 1\right]$. For CA-MAML, c is set as 0.5. The meta-batch size is set to 4 for all settings (Finn et al., 2017; Jamal & Qi, 2019). However, we study its impact in Table 1. All models were trained for 55000 iterations (early stopping was employed for tieredImagenet) using the optimal set of hyper-parameters using an Adam optimizer (Kingma & Ba, 2015). All the experimental results and comparisons correspond to our re-implementation of the ML algorithms integrated into learn2learn library (Arnold et al., 2020) to ensure fairness and uniformity. We believe that integrating the proposed attention module and additional ML algorithms into the learn2learn library will benefit the ML community. We perform individual hyperparameter tuning for all the models over the same hyperparameter space to ensure a fair comparison. The source code is publicly available.[1]

Table 1: Comparison of few-shot classification performance of MAML and TA-MAML on miniImagenet dataset with meta-batch size 4 and 6 and 8 for 5 and 10 way (1 and 5 shot) settings. The ± represents the 95% confidence intervals over 300 tasks. Algorithms denoted by * are rerun on their optimal hyper-parameters on our experimental setup. We observe that TA-MAML consistently performs better than MAML, and an increase in the tasks in a batch improves the performance of both MAML and TA-MAML.

	Test Accuracy (%) on miniImagenet			
	5 Way		10 Way	
Model	1 Shot	5 Shot	1 Shot	5 Shot
	Batch Size 4			
MAML*	46.10 ± 0.19	60.16 ± 0.17	29.42 ± 0.11	41.98 ± 0.10
TA-MAML*	**48.36 ± 0.23**	**62.48 ± 0.18**	**31.15± 0.11**	**43.70 ± 0.09**
	Batch Size 6			
MAML*	47.72 ± 1.041	63.45 ± 1.083	31.55 ± 0.626	46.27 ± 0.64
TA-MAML*	**49.14 ± 1.211**	**65.26 ± 0.956**	**32.62± 0.635**	**46.67 ± 0.63**
	Batch Size 8			
MAML*	47.68±1.20	63.81±0.98	31.54±0.66	46.15±0.58
TA-MAML*	**50.35±1.22**	**65.69±1.08**	**32.00±0.68**	**48.33±0.63**

The literature reports significant variations in the meta-test performances of various ML approaches (Table 7 in supplementary material). The reported average meta-test accuracies of MAML on the miniImagenet dataset range from 46.47 % to 48.70 % (55.16% to 64.39%) for 5 way 1 shot (5 shot) settings. A careful analysis reveals the different experimental setups resulting in the observed variation. Experimental setups (Finn et al., 2017; Oreshkin et al., 2018; Oh et al., 2020) differ in the number of examples per class in the query set, the number of gradient descent steps in the inner loop, meta-batch size, inductive or transductive batch normalization, etc. We conduct two sets of experiments to test the proposed task attention model's efficacy in a fair manner. The first set of experiments use the train and test setups reported in the literature (denoted using $^{\#}$). The second set uses our setup (denoted using *) that has the same train and test conditions. Specifically, we set the query examples per class to 15 and gradient steps to 5 for both the meta-train and meta-test phases. However, for 10 way 5 shot setting, we use only 2 gradient steps to reduce the computational burden. More query examples per class (15) during the meta-test provide a robust estimate of the model's generalizability. Further, setting gradient steps to 5 (or 2) better evaluates the quick adaptation capabilities of a learned prior.

5.2 Influence of Task Attention on Meta-Training

As task-attention (TA) is a standalone module, it can be integrated with any batch episodic training regimen. We, therefore, use MetaLSTM++ (batch mode of MetaLSTM) for our experiments. In (Aimen et al., 2021), authors demonstrated the merit of MetaLSTM++ on MetaLSTM only on Omniglot dataset. We extend upon

[1] https://github.com/taskattention/task-attended-metalearning.git

this empirical investigation by comparing the performance of MetaLSTM and MetaLSTM++ on complex datasets like miniImagenet, FC100, and tieredImagenet (Table 2). It is evident from the results that batch-wise episodic training is more effective than sequential episodic training. We also investigate the performance of models trained with the TA meta-training regimen with their non-TA counterparts on both (our and reported - wherever available) setups. Specifically, we compare MAML, MetaSGD, MetaLSTM++, and ANIL with their task-attended versions on 5 and 10 way (1 and 5 shot) settings on miniImagenet, FC100, and tieredImagenet datasets and report the results in Table 2. We consider 300 meta-test tasks for all approaches unless specified otherwise. For ANIL and its task-attended counterpart, we consider 1000 testing tasks. From Table 2, we observe that models trained with TA regimen generalize better to the unseen meta-test tasks than their non-task-attended versions across all the settings in all datasets. Note that the proposed task attention mechanism aims not to surpass the state-of-the-art meta-learning algorithms but provides new insight into the batch episodic meta-training regimen, which as per our knowledge, is common to all meta-learning algorithms.

We also compare the performance of TA-MAML against TAML - a meta-training regimen that forces the meta-model to be equally close to all the tasks. The results, as presented in Table 2, suggest that TA-MAML performs better than TAML on all benchmarks across all settings. Note that both TAML and TA-MAML are approaches that built upon MAML to address the inequality/diversity of tasks in a batch. Our aim is thus to compare TAML and TA-MAML and not to assess the efficacy of TAML when meta-trained using task attention. Liu et al. (2021a) proposed an optimization method to neutralize conflicts of an average model with individual tasks in a multi-task learning setup. Specifically, they find an optimal update vector that lies in the proximity of the average gradient across the batch of the tasks without conflicting with any task gradient. This method is similar to (Jamal & Qi, 2019) in a meta-learning setup, which constrains the losses of tasks towards the average loss on a task batch. As the update vector is constrained to be close to the average gradient vector on a task batch, information flow from certain useful tasks to the meta-model may decrease. We note that we extend (Liu et al., 2021a) to a meta-learning setup by computing the average and weighted average gradients on query loss of the adapted models instead of a model from the previous iteration (as in a multi-task setup). Table 2 demonstrates that the proposed attention mechanism has better generalizability to unseen tasks than conflict-averse gradient descent adapted for a meta-learning setup (CA-MAML). Our approach utilizes a non-linear model to extract knowledge from multiple meta-information components to learn the weights, which helps it to outperform TAML and CA-MAML.

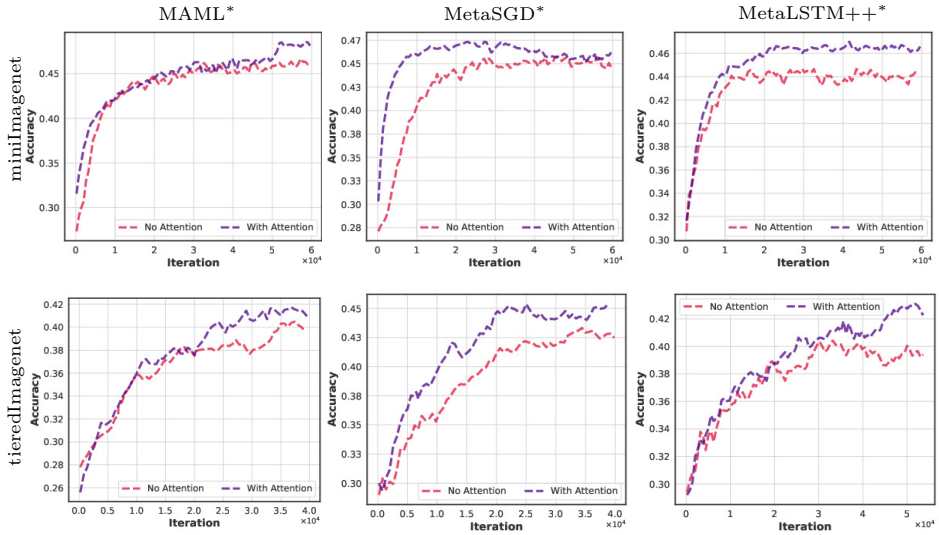

Figure 3: Mean validation accuracies of MAML* (Col-1), MetaSGD* (Col-2) and MetaLSTM++* (Col-3) across 300 tasks with/without attention on 5 way 1 shot setting on miniImagenet (Row-1) and tieredImagenet (Row-2) datasets.

We investigate the influence of the TA meta-training regimen on the model's convergence by analyzing the trend of the model's validation accuracy over iterations. Figure 3 depicts the mean validation accuracy over 300 tasks on miniImagenet and tieredImagenet datasets for a 5 way 1 shot setting across training iterations. We observe that the models meta-trained with TA regimen tend to achieve higher/at-par performance in fewer iterations than the corresponding models meta-trained with the non-TA regimen.

Table 2: Comparison of few-shot classification performance of vanilla ML algorithms with their task attended versions on miniImagenet, FC100 and tieredImagenet datasets for 5 and 10 way (1 and 5 shot) settings. The ± represents the 95% confidence intervals over 300 tasks. Algorithms denoted by * and # are rerun on using the optimal hyper-parameters on our and reported experimental setups, respectively. Attention-based ML algorithms perform better than their corresponding vanilla approaches across all the settings. Further, MetaLSTM++ and TA-MAML perform better than MetaLSTM and TAML (and CA-MAML) , respectively, across all settings and datasets.

	Test Accuracy (%)			
	5 Way		10 Way	
Model	1 Shot	5 Shot	1 Shot	5 Shot
		miniImagenet		
MAML# (Finn et al., 2017)	48.07 ± 1.75	63.15 ± 0.91	-	-
CA-MAML# (Liu et al., 2021a)	47.86 ± 2.50	64.27 ± 1.26	-	-
TAML# (Jamal & Qi, 2019)	51.77 ± 1.86	65.6 ± 0.93	-	-
TA-MAML#	**53.80 ± 1.85**	**66.11 ± 0.11**	-	-
MAML*	46.10 ± 0.19	60.16 ± 0.17	29.42 ± 0.11	41.98 ± 0.10
TAML*	46.26 ± 0.21	53.40 ± 0.14	29.76 ± 0.11	36.88 ± 0.10
TA-MAML*	**48.36 ± 0.23**	**62.48 ± 0.18**	**31.15± 0.11**	**43.70 ± 0.09**
MetaSGD# (Li et al., 2017)	50.47 ± 1.87	64.03 ± 0.94	-	-
TA-MetaSGD#	**52.60 ± 0.25**	**67.54 ± 0.12**	-	-
MetaSGD*	47.65± 0.21	61.60 ± 0.17	30.09± 0.10	42.22 ± 0.11
TA-MetaSGD*	**49.28 ± 0.20**	**63.37 ± 0.16**	**31.50± 0.11**	**44.06 ± 0.10**
MetaLSTM*	41.48 ± 1.02	58.87 ± 0.94	28.62 ± 0.64	44.03 ± 0.69
MetaLSTM++*	48.00 ± 0.19	62.73 ± 0.17	31.16 ± 0.09	45.46 ± 0.10
TA-MetaLSTM++*	**49.18 ± 0.17**	**64.89 ± 0.16**	**32.07± 0.11**	**46.66 ± 0.09**
ANIL# (Raghu et al., 2020)	46.7 ± 0.4	61.5 ± 0.5	-	-
TA-ANIL#	**49.53 ± 0.41**	**63.73 ± 0.33**	-	-
ANIL*	46.92 ± 0.62	58.68 ± 0.54	28.84 ± 0.34	40.95 ± 0.32
TA-ANIL*	**48.84 ± 0.62**	60.80± 0.55	**31.14± 0.34**	**42.52 ± 0.34**
		FC100		
MAML*	36.40 ± 0.38	46.76±0.21	23.93±0.14	31.14 ± 0.07
TAML*	38.00 ± 0.26	48.05± 0.13	21.60± 0.14	33.19± 0.07
TA-MAML*	**39.86± 0.25**	**49.56 ± 0.13**	**25.46± 0.15**	**36.06± 0.08**
MetaSGD*	33.46 ± 0.23	43.96± 0.13	21.40±0.15	30.59± 0.07
TA-MetaSGD*	**35.66±0.25**	**49.49± 0.12**	**23.80±0.15**	**32.08±0.07**
MetaLSTM*	37.20 ± 0.26	47.89 ± 0.13	21.70 ± 0.14	32.11 ± 0.07
MetaLSTM++*	38.60 ±0.23	49.82 ± 0.12	22.80 ± 0.14	33.46 ± 0.08
TA-MetaLSTM++*	**41.53 ±0.28**	**51.17 ±0.13**	**25.33 ±0.15**	**34.18 ±0.08**
ANIL*	34.08 ± 1.29	44.74 ± 0.68	20.65 ± 0.77	27.93 ± 0.42
TA-ANIL*	**38.06 ± 1.26**	**46.94± 0.69**	**23.27± 0.79**	**28.29 ± 0.40**
		tieredImagenet		
MAML# (Oh et al., 2020)	47.44 ± 0.18	64.70 ± 0.14	-	-
TA-MAML#	**51.90 ± 0.19**	**69.43± 0.18**	-	-
MAML*	44.40 ± 0.49	57.07 ± 0.22	27.40 ± 0.25	34.30 ± 0.14
TAML*	46.40 ± 0.40	56.80 ± 0.23	26.40 ± 0.25	34.40 ± 0.15
TA-MAML*	**48.40 ± 0.46**	**60.40 ± 0.25**	**31.00± 0.26**	**37.60± 0.15**
MetaSGD*	52.80 ± 0.44	62.35 ± 0.26	31.90 ± 0.27	44.16 ± 0.15
TA-MetaSGD*	**56.20 ± 0.45**	**64.56 ± 0.24**	**33.20± 0.29**	**47.12 ± 0.16**
MetaLSTM*	37.00 ± 0.44	59.83 ± 0.25	29.80 ± 0.28	39.28 ± 0.13
MetaLSTM++*	47.60 ± 0.49	63.24 ± 0.25	30.70 ± 0.27	47.97 ± 0.16
TA-MetaLSTM++*	**49.00 ± 0.44**	**66.15 ± 0.23**	**32.10± 0.27**	**51.35 ± 0.17**
ANIL*	45.08 ± 1.37	59.71 ±0.77	29.32 ± 0.83	42.76 ± 0.50
TA-ANIL*	**45.96 ± 1.32**	**60.96± 0.72**	**32.68± 0.92**	**47.56 ± 0.51**

5.3 Comparison with Sampling Approaches

We compare our proposed approach with ATS (Yao et al., 2021) and uniform sampling (Arnold et al., 2021) and demonstrate that our weighting mechanism imparts better generalizability to the meta-model than the global weighting of the tasks. Yao et al. (2021) ascertained the merit of ATS over Greedy class-pair (GCP) technique (Liu et al., 2020) on miniImagenet dataset. We extend this comparison and show in Table 3 that the proposed approach performs better than state-of-the-art ATS and GCP in both 1 and 5 shot settings. We also observe that the TA mechanism performs better than uniform sampling on the miniImagenet dataset on 1 and 5 shot settings for MAML and ANIL. ATS has been designed for noisy and imbalanced task distributions. So, we compare the proposed approach with GCP, ATS, and other sampling techniques on the miniImagenet-noisy dataset (Yao et al., 2021) and report the results in Table 3. We observe that task attention outperforms all scheduling algorithms on the miniImagenet-noisy dataset. As ATS is the most competitive baseline for the proposed method on the miniImagenet-noisy dataset, we compare the TA-ANIL and ATS on varying noise ratios for the miniImagenet dataset on 5 way 1 shot setting (Table 4). We observe that the proposed method outperforms ATS on all noise ratios except 0.8. Note that the algorithm used for all sampling approaches is ANIL.

Table 3: Comparison (Test Accuracy (%)) of task attention with GCP, ATS and Uniform Sampling for MAML and MetaSGD (or ANIL) on miniImagenet dataset and various sampling techniques for ANIL on the miniImagenet-noisy dataset for 5 way 1 and 5 shot settings. For miniImagenet, algorithms denoted by * and # are rerun on the optimal hyper-parameters on our and reported experimental setups, respectively.

Model	5 Way 1 Shot	5 Way 5 Shot
	miniImagenet	
MAML with GCP#	46.92 ± 0.83	63.28 ± 0.66
MAML with ATS#	47.89 ± 0.77	64.07 ± 0.70
MAML+UNIFORM (Offline)#	46.67 ± 0.63	62.09 ± 0.55
MAML+UNIFORM (Online)#	46.70 ± 0.61	61.62 ± 0.54
TA-MAML* (Ours)	48.36 ± 0.23	62.48 ± 0.18
TA-MAML# (Ours)	**53.80 ± 1.85**	**66.11 ± 0.11**
MetaSGD with GCP#	47.77 ± 0.75	63.50 ± 0.71
MetaSGD with ATS#	48.59 ± 0.79	64.79 ± 0.74
TA-MetaSGD* (Ours)	49.28 ± 0.20	63.37± 0.16
TA-MetaSGD# (Ours)	**52.60 ± 0.25**	**67.54 ± 0.12**
ANIL+UNIFORM (Offline)#	46.93 ± 0.62	62.75 ± 0.60
ANIL+UNIFORM (Online)#	46.82 ± 0.63	62.63 ± 0.59
TA-ANIL*	48.84 ± 0.62	60.80± 0.55
TA-ANIL#	**49.53 ± 0.41**	**63.73 ± 0.33**
	miniImagenet-noisy	
Uniform	41.67 ± 0.80	55.80 ± 0.71
SPL	42.13 ± 0.79	56.19 ± 0.70
Focal Loss	41.91 ± 0.78	53.58 ± 0.75
GCP	41.86 ± 0.75	54.63 ± 0.72
PAML	41.49 ± 0.74	52.45 ± 0.69
DAML	41.26 ± 0.73	55.46 ± 0.70
ATS	44.21 ± 0.76	59.50 ± 0.71
TA-ANIL* (Ours)	**45.17 ± 0.23**	**62.15 ± 1.01**

5.4 Effectiveness of Task Attention in CDFSL setup

Classical meta-learning approaches assume meta-train and meta-test data belong to the same distribution such that the meta-trained model extends its knowledge to the meta-test set. This is, however, not always the case. The difference in the data acquisition techniques, or evolution of data with time, may cause a discrepancy between the meta-train and meta-test distributions. This realistic setting is popularly termed as cross-domain few-shot learning (CDFSL) (Guo et al., 2020). We conducted experiments to show the merit of the proposed approach in CDFSL setup. Specifically, we train a model using TA meta-training regimen on the miniImagenet dataset and meta-test it on CUB-200 and FGVC-Aircraft datasets. The results reported for 5 way 1 and 5 shot settings in Table 5 indicate that the proposed approach outperforms the state-of-the-art task scheduling approach (Uniform Sampling (Arnold et al., 2021)) on CDFSL setup by a large margin.

5.5 Ablation Studies

To examine the significance of each input given to the task attention model, we conduct an ablation study on 5 way 1 and 5 shot TA-MAML on miniImagenet dataset and re-

Table 4: Comparative analysis of ANIL integrated with ATS and proposed method on miniImagenet dataset with varying noise ratios for 5 way 1 shot setting. BNS is the best non-adaptive scheduler.

Noise ratio	Test Accuracy (%) on miniImagenet-noisy 0.2	0.4	0.6	0.8
ANIL with Uniform	43.46 ± 0.82	42.92 ± 0.78	41.67 ± 0.80	36.53 ± 0.73
ANIL with BNS	44.04 ± 0.81	43.36 ± 0.75	42.13 ± 0.79	38.21 ± 0.75
ANIL with ATS	45.55 ± 0.80	44.50 ± 0.86	44.21 ± 0.76	**42.18 ± 0.73**
TA-ANIL* (Ours)	**47.98 ± 0.26**	**46.69 ± 0.22**	**45.17 ± 0.23**	40.35 ± 1.14

port the results in Table 6. We observe that all the components of meta-information contribute to the learning of a more generalizable meta-model. To further support this observation, we investigate the relationship between the meta-information and weights assigned by the task attention module by analyzing the mean Pearson correlation of each of the components (four tuple) of the meta-information with the attention vector across the training iterations. This is depicted in Figure 4 for TA-MAML on 5 way 1 and 5 shot settings for miniImagenet dataset. We observe that the loss ratio and loss are positively correlated with the attention vector, while accuracy and gradient norm are negatively correlated.

In 5 way 5 shot setting, we observe that the correlation pattern is comparable to 5 way 1 shot setting, but the mean correlation value of grad norm across iterations is less than that of the 5 way 1 shot setting. This could be because the 5 way 5 shot setting is richer in data than the 5 way 1 shot setting, which allows better learning and therefore has low average values of grad norm (Section 4.1.1). The critical observation, however, is that the meta-information components have a weak correlation with the attention weights, indicating that the TA module does not trivially follow any single component of meta-information. We also analyze the ranks of the tasks for maximum and minimum values of : loss, loss ratio, accuracy, and grad norm in a batch, as per the weights across training iterations, and describe results in the supplementary material. The rank analysis also reinforces the same observation. We ascertain the decreasing trend of mean weighted loss across iterations in the supplementary material.

Table 5: Comparative analysis of proposed approach and uniform sampling (Arnold et al., 2021) in a CDFSL setting after training on miniImagenet dataset and tested on CUB-200 and FGVC-Aircraft datasets for 5 way 1 and 5 shot settings.

Model	5 Way 1 Shot	5 Way 5 Shot
	CUB-200	
MAML+ UNIFORM (Online)	35.84 ± 0.54	46.67 ± 0.55
TA-MAML* (Ours)	**42.87 ± 1.18**	**57.49 ± 0.99**
	FGVC-Aircraft	
MAML+ UNIFORM (Online)	26.62 ± 0.39	34.41 ± 0.44
TA-MAML* (Ours)	**29.42 ± 0.78**	**36.34 ± 0.86**

5.6 Analysis of Attention Network

To gain further insights into the operation of the attention module, we also examine the trend of the attention-vector (Figure 5) while meta-training TA-MAML for 5 way 1 and 5 shot settings on the miniImagenet dataset. We plot the maximum and the minimum attention score assigned to the tasks of a batch across iterations together with a few weighted task batches in 5 way 1 shot setting for illustration. We note that the weighted task batches are only intended to demonstrate the change in the tasks' attention scores across iterations. The next experiment presents a more rigorous analysis studying the relationship among classes in a task and attention scores assigned.

Table 6: Effect of ablating components of meta-information in TA-MAML* for 5 way 1 and 5 shot settings on miniImagenet dataset.

Ablation on inputs					
Grad norm	Loss	Loss-ratio	Accuracy	Test Accuracy	
				5 way 1 shot	5 way 5 shot
×	×	×	×	46.10±0.19	60.16±0.17
✓	✓	✓	×	47.30±0.16	60.48±0.16
✓	✓	×	✓	47.62±0.17	62.17±0.17
✓	×	✓	✓	48.10±0.18	60.90±0.20
×	✓	✓	✓	47.30±0.18	61.52±0.16
✓	✓	✓	✓	**48.36±0.23**	**62.48±0.18**

We note that the mean attention score is always 0.25 as we follow a meta-batch size of 4. We observe that the TA module's output follows an interesting trend. Initially, the TA module assigns almost uniform weights to all the tasks of a batch; however, as the iterations increase, it assigns unequal scores to the tasks in a batch, preferring some over the other. This suggests that during the initial phases of the meta-model's training, all tasks have equal contribution towards learning a *generic structure* of the meta-knowledge. As the meta-model's learning proceeds, learning the further *fine-grained meta-knowledge structure* requires prioritizing some tasks in a batch over the others, which are potentially better aligned with learning the optimal meta-knowledge. We study the computational burden imposed by TA regimen in the appendix.

We further decipher the functioning of the black box attention network by analyzing the qualitative relation among weights and the classes of task batches (Figure 9 is presented in appendix due to space constraints).

In Figure 9 left column (col-1) corresponds to the cases where the assignment of attention scores to the tasks is human interpretable. In contrast, the right column (col-2) refers to the uninterpretable attention scores. From the human perspective, tasks containing images from similar classes are hard to distinguish and are assigned higher attention scores indicated by red bounding boxes (Figure 9 col-1). Specifically, (col-1, row-1) task 2 is regarded as most important, possibly because it includes three breeds of dogs followed by task 4, which comprises two species of fish. However, the aforementioned is not a hard constraint, as there are some task batches (Figure 9 col-2) in which the distribution of weights cannot be explained qualitatively.

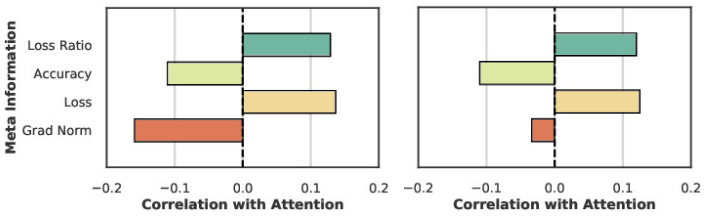

Figure 4: Mean Pearson correlation of TA-MAML* on 5 way 1 shot (left) and 5 shot (right) setting on miniImagenet.

6 Conclusion

In this work we have shown that the batch wise episodic training regimen adopted by ML strategies can benefit from leveraging knowledge about the importance of tasks within a batch. Unlike prior approaches that assume uniform importance for each task in a batch, we propose task attention as a way to learn the relevance of each task according to its alignment with the optimal meta-knowledge. We have validated the effectiveness of task attention by augmenting it to popular initialization and optimization based ML strategies. We have demonstrated through experiments on miniImagenet, FC100 and tieredImagenet datasets that augmenting task attention helps attain better generalization to unseen tasks from the same distribution while requiring fewer iterations to converge. We also show that the task attention is meritorious over existing task scheduling algorithms, even on noisy and CDFSL setups. We also conduct an exhaustive empirical analysis on the distribution of attention weights to study the nature of the meta-knowledge and task attention module. We leave the theoretical motivation of the meta-information components and the proof of convergence of the proposed curriculum as part of our future work. We believe that this end-to-end attention-based meta training paves the way towards efficient and automated meta-training.

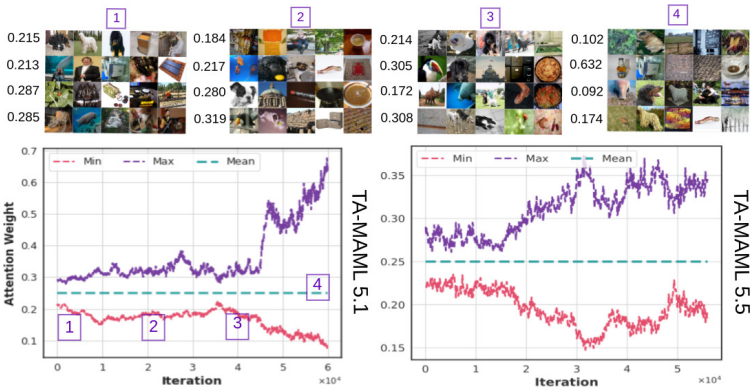

Figure 5: Trend of an attention vector in 5 way 1 shot (left) and 5 shot (right) settings on miniImagenet dataset for TA-MAML*.

References

Mayank Agarwal, Mikhail Yurochkin, and Yuekai Sun. On sensitivity of meta-learning to support data. *Advances in Neural Information Processing Systems*, 34:20447–20460, 2021.

Aroof Aimen, Sahil Sidheekh, Vineet Madan, and Narayanan C Krishnan. Stress Testing of Meta-learning Approaches for Few-shot Learning. In *AAAI Workshop on Meta-Learning and MetaDL Challenge*, 2021.

Antreas Antoniou, Harri Edwards, and Amos Storkey. How to train your maml. In *Seventh International Conference on Learning Representations*, 2019.

Sébastien Arnold, Guneet Dhillon, Avinash Ravichandran, and Stefano Soatto. Uniform sampling over episode difficulty. *Advances in Neural Information Processing Systems*, 34:1481–1493, 2021.

Sébastien MR Arnold, Praateek Mahajan, Debajyoti Datta, Ian Bunner, and Konstantinos Saitas Zarkias. learn2learn: A library for meta-learning research. *CoRR*, 2020.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pp. 41–48. ACM, 2009.

Haw-Shiuan Chang, Erik G. Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 1002–1012, 2017.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2018.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European conference on computer vision*, pp. 124–141. Springer, 2020.

Ricardo Luna Gutierrez and Matteo Leonetti. Information-theoretic task selection for meta-reinforcement learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.

Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 11719–11727. Computer Vision Foundation / IEEE, 2019.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2309–2318. PMLR, 2018.

Jean Kaddour, Steindór Sæmundsson, and Marc Peter Deisenroth. Probabilistic active meta-learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

H. Kahn and A. W. Marshall. Methods of reducing sample size in monte carlo computations. *Oper. Res.*, 1953.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pp. 1189–1197. Curran Associates, Inc., 2010.

Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2019.

Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning, 2017.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 2999–3007. IEEE Computer Society, 2017.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021a.

Chenghao Liu, Zhihao Wang, Doyen Sahoo, Yuan Fang, Kun Zhang, and Steven C. H. Hoi. Adaptive task sampling for meta-learning. In *ECCV*, 2020.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *ICML*, 2021b.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Jaehoon Oh, Hyungjun Yoo, ChangHwan Kim, and Se-Young Yun. Boil: Towards representation change for few-shot learning. In *International Conference on Learning Representations*, 2020.

Boris N. Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. TADAM: task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 719–729, 2018.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of MAML. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

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. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018a.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4331–4340. PMLR, 2018b.

Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick. Training region-based object detectors with online hard example mining. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 761–769. IEEE Computer Society, 2016.

Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 403–412. Computer Vision Foundation / IEEE, 2019.

Qianru Sun, Yaoyao Liu, Zhaozheng Chen, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning through hard tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3630–3638, 2016.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

Huaxiu Yao, Yu Wang, Ying Wei, Peilin Zhao, Mehrdad Mahdavi, Defu Lian, and Chelsea Finn. Meta-learning with an adaptive task scheduler. *Advances in Neural Information Processing Systems*, 2021.

Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1–9. JMLR.org, 2015.