Task Attended Meta-Learning for Few-Shot Learning

Aroof Aimen, Sahil Sidheekh, Bharat Ladrecha, Narayanan C. Krishnan
{2018csz0001, 2017csb1104, 2018csb1080, ckn} @iitrpr.ac.in
Indian Institute of Technology, Ropar

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 episodic training. The former approaches leverage the knowledge from a batch of tasks to learn an optimal prior. 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 weight the tasks in a batch. The task attention is a standalone unit and can be integrated with any batch episodic training regimen. The comparisons of the task-attended ML models with their non-task-attended counterparts on complex datasets like miniImageNet, FC100 and tieredImageNet 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 [49]. 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 [10, 28, 18]. Under the parametric optimizer view, meta-knowledge pertaining to the traversal of the loss surface of individual tasks is learned by the meta-optimizer. Through learning

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

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 [40].

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. Leveraging common knowledge from a batch of tasks to learn the parametric optimizer can help address this issue. We first accumulate meta-knowledge in a batch mode for the parametric optimizer. 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 bootstrapping, which however, operate at sample granularity. In bootstrapping, 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.
- To integrate task attention module with the parametric optimizer, we design a batch-mode parametric optimizer (MetaLSTM++) and experimentally show its merit on miniImageNet, FC100, and tieredImageNet datasets.
- We conduct extensive experiments on miniImageNet, FC100, and tieredImageNet datasets, and comparisons of the ML algorithms with their non-task-attended counterparts to validate the effectiveness of the task attention module and its coupling with any batch episodic training regimen.
- We also perform exhaustive empirical analysis to decipher the working of the task attention module.

2 Related Work

ML literature is profoundly diverse and may broadly be classified into *metric approaches* [49, 45, 48, 16, 23, 8], *initialization approaches* [10, 35, 11, 52, 3, 25, 28, 18, 39, 12, 52, 14, 36, 38, 34, 26, 42, 17, 27, 47, 46], *optimization approaches* [40, 2, 7, 51] and *model approaches* [43, 33, 37, 32] depending on how the meta-knowledge is accumulated. *Metric approaches* learn an embedding from input data and design kernel functions to classify the query samples by finding the maximum similarity sample in the support set. *Initialization approaches* learn an optimal prior on model parameters. The model is thus generalizable to new tasks drawn from the same distribution. *Optimization approaches* learn parametric optimizers to traverse the loss surfaces of tasks during training and guide the model along the loss surfaces of newly sampled tasks from the same distribution. *Model approaches* employ an external memory to store the meta-information gathered from the seen tasks and use it to generalize to unseen tasks.

However, all of these meta-learning approaches follow training strategies that randomly sample tasks with uniform probability. Assigning non-uniform priorities at sample granularity is not new [22, 13, 44]. Various attributes like losses, gradients, uncertainty, etc., have been used to assign priorities to samples [29, 53, 6]. Motivated by human learning, this literature is succeeded by

[5, 19, 24, 50, 41, 20, 9] wherein training of easier samples precede hard ones. These works vary in the procedure of arranging samples throughout training, aiming least overhead in terms of an additional pre-trained model or multiple passes over data. Nevertheless, assigning non-uniform priorities to tasks in meta-learning is under explored. Recent works like [21, 15, 30] focus on sampling tasks based on the information contained in the tasks. [21, 15] are specific to reinforcement learning and [30] propose a class-pair potential based adaptive task sampling strategy. Our work is different from these as we do not propose a task sampling strategy rather a task weighting mechanism for the meta-model update. Contrary to our idea is TAML [18] - 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 $D_i = \{\{x_k^c, y_k^c\}_{k=1}^K\}_{c=1}^N$ 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 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 on $\mathcal{M}$, meta-model θ is adapted on D_i of each $\mathcal{T}_i$ to ϕ_i. The adapted model ϕ_i is then evaluated on D_i^* 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 $\mathcal{M}_t$.

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:

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

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

Gradient norm: Let $P = \left\{(\phi_1^T)_i, \ldots, (\phi_n^T)_i\right\}_{i=1}^B$ be the parameters of the model obtained after adapting the initial model (T iterations) on support data $\{D_i\}_{i=1}^B$. Also, let $G = \left\{(\nabla_{(\phi_1^T, D)} L^*(\phi_1^T, D^*))_i, \ldots, (\nabla_{(\phi_n^T, D)} L^*(\phi_n^T, D^*))_i\right\}_{i=1}^B$ be the gradients of the adapted model

parameters w.r.t the query losses $\{L_i^*\}_{i=1}^B$. The gradient norm $\left\{||\nabla_{\phi_i^T} L_i^*(\phi_i^T)||\right\}_{i=1}^B$ is the L_2 norm of G that carries information about the magnitude of the consolidated displacement of the adapted model parameters during a gradient descent update on query data. This magnitude of consolidated displacement (grad norm) of adapted task model parameters characterizes its generalizability to the unseen query data. Larger gradient norm on query dataset could indicate that the model has either overfitted or has not learned the support set. Hence the model is not generalizable on query set compared to the models with low gradient norm. Thus, the grad norm aids the task attention module in determining the weights of tasks while considering their generalizability.

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

Initialization: $\theta, \delta \leftarrow$ Random Initialization
for *iteration in* n_{iter} **do**
  $\{\mathcal{T}_i\}_{i=1}^B = \{D_i, D_i^*\}_{i=1}^B \leftarrow$ Sample task-batch($\mathcal{M}$)
  for *all* $\mathcal{T}_i$ **do**
    $\phi_i^0 \leftarrow \theta$
    $L_i^{*0}, _ \leftarrow evaluate(\phi_i^0, D_i^*)$
    $\phi_i^T = adapt(\phi_i^0, D_i)$
    $L_i^{*T}, A_i^{*T} \leftarrow evaluate(\phi_i^T, D_i^*)$
  end
  $[w_i]_{i=1}^B \leftarrow$
  $Att_module\left(\left[\frac{L_i^{*T}}{L_i^{*0}}, A_i^{*T}, ||\nabla_{\phi_i^T} L_i^*(\phi_i^T)||, L_i^{*T}\right]_{i=1}^B\right)$
  $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{i=1}^B w_i L_i^*(\phi_i^T)$
  $\{D_j, D_j^*\}_{j=1}^B \leftarrow$ Sample task-batch($\mathcal{M}$)
  for *all* $\mathcal{T}_j$ **do**
    $\phi_j^0 \leftarrow \theta$
    $\phi_j^T = adapt(\phi_j^0, D_j)$
  end
  $\delta \leftarrow \delta - \gamma \nabla_\delta \sum_{j=1}^B L_j^*(\phi_j^T)$
end
Return θ

Test Loss: $\{L_i^{*T}\}_{i=1}^B$ represents the empirical error of the adapted base models on unseen query instances and hence characterizes the generalizability of the adapted models to unseen query data. Unlike grad 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. Moreover, $\{L_i^{*T}\}_{i=1}^B$ is a component of the meta-update so, each tasks' query loss has a direct influence on learning the meta-model and therefore is an essential characteristic of a task. Further, test errors of classes have widely been used to determine their "easy or hardness" [5, 31]. Thus $\{L_i^{*T}\}_{i=1}^B$ acquaints the attention module with the generalizability aspect of task models and their influence in updating the meta-model.

Test Accuracy: $\{A_i^{*T}\}_{i=1}^B$ corresponds to the accuracies of $\{\phi_i^T\}_{i=1}^B$ on $\{D_i^*\}_{i=1}^B$ scaled in the range [0,1]. A_i^{*T} measures the thresholded prediction based on the highest softmax value and the actual class label, unlike L_i^{*T}, which determines the confidence of the predictions. Two task models may predict the same class labels but differ in the confidence of the predictions. In such scenarios, neither loss nor accuracy individually is 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.

Loss-ratio: Let L_i^{*0} be the loss of θ on the D_i^*, and L_i^{*T} be the loss of the adapted model ϕ_i^T on D_i^*. The loss- ratio $\frac{L_i^{*T}}{L_i^{*0}}$ 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 transfer), while lower values (< 1) of the loss ratio indicates the benefit of adaptation of θ on D_i (positive transfer). Loss ratio of exactly one signifies adaptation attributes to no additional benefit (neutral transfer). Therefore, loss-ratio quantifies positive, negative, or neutral transfer of task knowledge to the meta-model given its current configuration and task data.

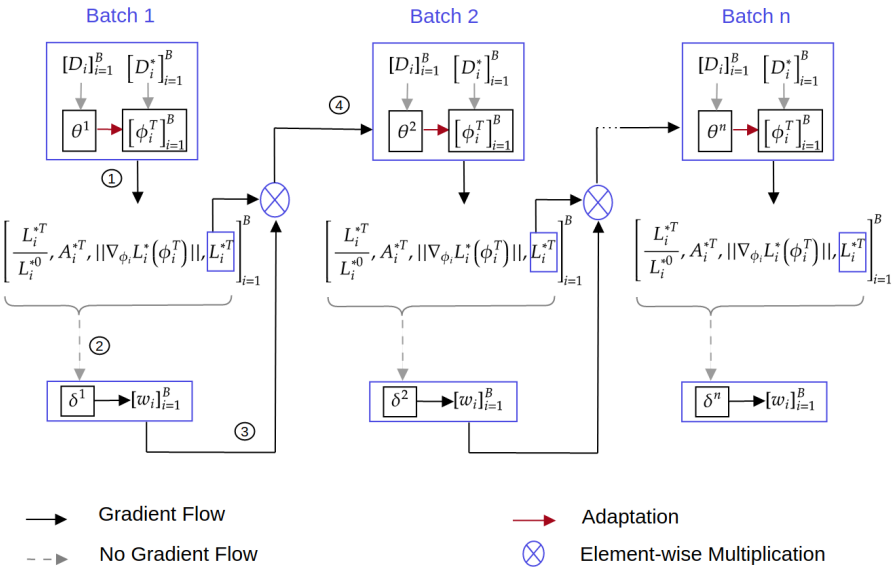


Figure 1: Computational Graph of the forward pass of the meta-model using TA meta-training curriculum. The output of this procedure is a meta-model θ^n. Gradients are propagated through solid lines and restricted through dashed lines.

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 $\mathbf{w} = [w_1, \ldots, w_B]$, $(\sum_{i=1}^{B} w_i = 1)$ 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 $L_i^*(\phi_i^T)$ on the held-out datasets D_i^* to update the meta-model θ:

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

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:

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

where ϕ_j^T is adapted from θ^{t+1}.

4.3 Task Attended Meta-Training Algorithm

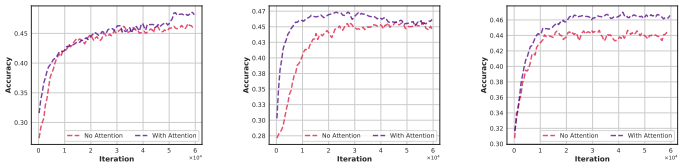
We demonstrate the meta-training curriculum using the proposed task attention in Figure 1, formally summarized in Algorithm 1. As with the classical meta-training process, we first sample a batch of tasks from the task distribution. For each task $\mathcal{T}_i$, we adapt the base-model ϕ_i using the train data $\{D_i\}_{i=1}^{B}$ for T time-steps (line 7). The meta-information about the adapted models for each task is then computed, comprising of the loss L_i^{*T}, the accuracy A_i^{*T}, the loss-ratio $\frac{L_i^{*T}}{L_i^{*0}}$ and gradient

norm on test data $\{D_i^*\}_{i=1}^B$. The 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). The attention vector in combination with test losses $\{L_i^*\}_{i=1}^B$ is used to update meta-model parameters θ (line 11, 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. We compute the mean test loss over the adapted base-models $\{L_j^*(\phi_j^T)\}_{j=1}^B$, which is then used to update the parameters of the task attention module δ (lines 12-17).

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. Thus, it is important to decouple the learning of the attention network from that of the meta-model. The parameters of the meta-model θ should not be directly dependent on that of the task attention module δ. We prevent it by enforcing $\nabla_\theta w_i L_i^*(\phi_i^T) = w_i \nabla_\theta L_i^*(\phi_i^T)$. Restricting the flow of gradients to the meta-model through the task attention module also enables us to evade the computational overhead generated by the product of gradients. Specifically, stopping the gradient flow frees us from computing $\nabla_{\delta^t}\theta^{t+1}.\nabla_{\phi^t}\delta^t.\nabla_{\theta^t}\phi^t$. The leading term $\nabla_{\delta^t}\theta^{t+1}$ in turn requires the computation of $\nabla_{\delta^t}.\nabla_{\theta^t}.\nabla_{\theta^t}$. Figure 1 demonstrates the paths along which gradient backflow is restricted and permitted as dashed and solid lines respectively.

5 Experiments and Results

We consider different few-shot learning settings on the benchmark datasets - miniImageNet, Fewshot Cifar 100 (FC100) and tieredImageNet to test the effectiveness of the proposed attention module. All the experimental results and comparisons correspond to our re-implementation of the ML algorithms integrated into learn2learn library [4] 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 architecture and hyper-parameter details are provided in the implementation details in supplementary material. The source code is publicly available.[1]



(a) MAML (b) MetaSGD (c) MetaLSTM++

Figure 2: Mean validation accuracies of a) MAML b) MetaSGD c) MetaLSTM++ across 300 tasks with/without attention on 5-way 1-shot setting on miniImageNet dataset.

5.1 Influence of Task Attention on Meta-Training

As the task-attention (TA) is a standalone module, it can be integrated with any batch episodic training regimen. To facilitate this integration, we introduced a batch-wise training regimen for parametric optimizer (MetaLSTM++ [1]). The comparative analysis of MetaLSTM and MetaLSTM++ on miniImagenet, FC100, and tieredImagenet is presented in Table 1. It is evident from the results that batch-wise episodic training is effective than sequential episodic training. We also investigate the performance of the models trained with the TA meta-training regimen with their non-TA counterparts. Specifically, we compare MAML, MetaSGD and MetaLSTM++ with TA-MAML, TA-MetaSGD and TA-MetaLSTM++ respectively over 5 and 10-way (1 and 5-shot) settings on miniImageNet, FC100 and tieredImageNet datasets and report the results in Table 1. 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 and all datasets. Note that the proposed TA mechanism

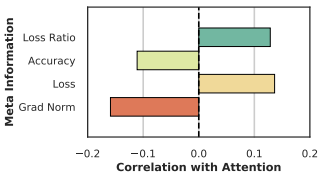


Figure 3: Mean Pearson correlation of TA-MAML on 5-way 1-shot setting on miniImagenet.

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

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. The results, as pre-

Table 1: 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 $\pm$ represents the 95% confidence intervals over 300 tasks. Algorithms denoted by * are rerun on their optimal hyper-parameters. 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, respectively, across all settings and datasets.

	Test Accuracy (%)			
	5-Way		10-Way	
Model	1 Shot	5 Shot	1 Shot	5 Shot
		miniImageNet		
MAML*	46.10 $\pm$ 0.19	60.16 $\pm$ 0.17	29.42 $\pm$ 0.11	41.98 $\pm$ 0.10
TAML*	46.26 $\pm$ 0.21	53.40 $\pm$ 0.14	29.76 $\pm$ 0.11	36.88 $\pm$ 0.10
TA-MAML	**48.36 $\pm$ 0.23**	**62.48 $\pm$ 0.18**	**31.15$\pm$ 0.11**	**43.70 $\pm$ 0.09**
MetaSGD*	47.65$\pm$ 0.21	61.60 $\pm$ 0.17	30.09$\pm$ 0.10	42.22 $\pm$ 0.11
TA-MetaSGD	**49.28 $\pm$ 0.20**	**63.37 $\pm$ 0.16**	**31.50$\pm$ 0.11**	**44.06 $\pm$ 0.10**
MetaLSTM*	41.48 $\pm$ 1.02	58.87 $\pm$ 0.94	28.62 $\pm$ 0.64	44.03 $\pm$ 0.69
MetaLSTM++	48.00 $\pm$ 0.19	62.73 $\pm$ 0.17	31.16 $\pm$ 0.09	45.46 $\pm$ 0.10
TA-MetaLSTM++	**49.18 $\pm$ 0.17**	**64.89 $\pm$ 0.16**	**32.07$\pm$ 0.11**	**46.66 $\pm$ 0.09**
		FC100		
MAML*	36.40 $\pm$ 0.38	46.76$\pm$0.21	23.93$\pm$0.14	31.14 $\pm$ 0.07
TAML*	38.00 $\pm$ 0.26	48.05$\pm$ 0.13	21.60$\pm$ 0.14	33.19$\pm$ 0.07
TA-MAML	**39.86$\pm$ 0.25**	**49.56 $\pm$ 0.13**	**25.46$\pm$ 0.15**	**36.06$\pm$ 0.08**
MetaSGD*	33.46 $\pm$ 0.23	43.96$\pm$ 0.13	21.40$\pm$0.15	30.59$\pm$ 0.07
TA-MetaSGD	**35.66$\pm$0.25**	**49.49$\pm$ 0.12**	**23.80$\pm$0.15**	**32.08$\pm$0.07**
MetaLSTM*	37.20 $\pm$ 0.26	47.89 $\pm$ 0.13	21.70 $\pm$ 0.14	32.11 $\pm$ 0.07
MetaLSTM++	38.60 $\pm$0.23	49.82 $\pm$ 0.12	22.80 $\pm$ 0.14	33.46 $\pm$ 0.08
TA-MetaLSTM++	**41.53 $\pm$0.28**	**51.17 $\pm$0.13**	**25.33 $\pm$0.15**	**34.18 $\pm$0.08**
		tieredImageNet		
MAML*	44.40 $\pm$ 0.49	57.07 $\pm$ 0.22	27.40 $\pm$ 0.25	34.30 $\pm$ 0.14
TAML*	46.40 $\pm$ 0.40	56.80 $\pm$ 0.23	26.40 $\pm$ 0.25	34.40 $\pm$ 0.15
TA-MAML	**48.40 $\pm$ 0.46**	**60.40 $\pm$ 0.25**	**31.00$\pm$ 0.26**	**37.60$\pm$ 0.15**
MetaSGD*	52.80 $\pm$ 0.44	62.35 $\pm$ 0.26	31.90 $\pm$ 0.27	44.16 $\pm$ 0.15
TA-MetaSGD	**56.20 $\pm$ 0.45**	**64.56 $\pm$ 0.24**	**33.20$\pm$ 0.29**	**47.12 $\pm$ 0.16**
MetaLSTM*	37.00 $\pm$ 0.44	59.83 $\pm$ 0.25	29.80 $\pm$ 0.28	39.28 $\pm$ 0.13
MetaLSTM++	47.60 $\pm$ 0.49	63.24 $\pm$ 0.25	30.70 $\pm$ 0.27	47.97 $\pm$ 0.16
TA-MetaLSTM++	**49.00 $\pm$ 0.44**	**66.15 $\pm$ 0.23**	**32.10$\pm$ 0.27**	**51.35 $\pm$ 0.17**

sented in Table 1, 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.

We also 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 2 depicts the mean validation accuracy over 300 tasks on miniImageNet dataset for a 5-way 1-shot setting across training iterations. A similar convergence trend has been obtained for tieredImagenet and is presented in the supplementary material. 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.

5.2 Ablation Studies

To examine the significance of each input given to the task attention model, we conduct an ablation study on 5-way 1-shot TA-MAML on miniImagenet dataset and report the results in Table 2. 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 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 the results are described in the supplementary material. We also 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 of the meta-information with the attention vector across the training iterations. This is depicted in figure 3 for TA-MAML on 5-way 1-shot setting for mini-ImageNet dataset, and results on 5-way 5-shot setting are presented in the supplementary material. We observe that the loss ratio and loss are positively correlated with the attention vector, while accuracy and gradient norm are negatively correlated.

Table 2: Effect of ablating components of meta-information in TA-MAML for 5-way 1-shot setting on miniImagenet dataset.

Ablation on inputs				
Grad norm	Loss	Loss ratio	Accuracy	Test Accuracy
×	×	×	×	46.10± .019
✓	✓	✓	×	47.30± .016
✓	✓	×	✓	47.62±0.17
✓	×	✓	✓	48.10± 0.18
×	✓	✓	✓	47.30 ± 0.18
✓	✓	✓	✓	48.36 ± 0.23

5.3 Analysis of Attention Network

To gain further insights into the operation of the attention module, we examine the trend of the attention-vector (Figure 4) while meta-training TA-MAML for 5-way 1 and 5 shot settings on the miniImageNet dataset. Results for 5-shot setting are deferred to supplementary material. We plot the maximum and minimum attention scores assigned to the tasks of a batch across iterations together with a few weighted task batches for illustration. 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, the TA module 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.

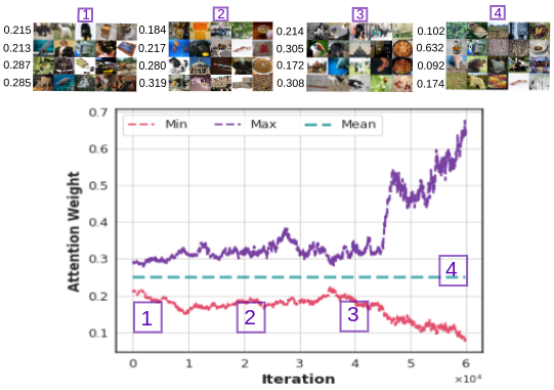


Figure 4: Trend of an attention vector in a 5-way 1-shot setting on miniImageNet dataset for TA-MAML.

6 Summary and Future Work

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 parametric-optimization based ML strategies. To facilitate integration with the latter, we have introduced a batch wise training strategy for a parametric optimizer, that outperforms its previous sequential counterpart. We have demonstrated through few-shot learning 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 conduct an exhaustive empirical analysis on the distribution of attention weights to study the nature of the meta-knowledge and task attention module. We believe this end-to-end attention-based meta training persuades towards fully automated meta-training.

Acknowledgements

The support and the resources provided by 'PARAM Shivay Facility' under the National Supercomputing Mission, Government of India at the Indian Institute of Technology, Varanasi and under Google Tensorflow Research award are gratefully acknowledged.

References

[1] 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.

[2] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *NeurIPS*, 2016.

[3] Antreas Antoniou, Harrison Edwards, and Amos J. Storkey. How to train your MAML. In *ICLR*, 2019.

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

[5] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, 2009.

[6] Haw-Shiuan Chang, Erik G. Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *NeurIPS*, 2017.

[7] Yutian Chen, Matthew W. Hoffman, Sergio Gomez Colmenarejo, Misha Denil, Timothy P. Lillicrap, Matthew Botvinick, and Nando de Freitas. Learning to learn without gradient descent by gradient descent. In *ICML*, 2017.

[8] Debasmit Das and C. S. George Lee. A two-stage approach to few-shot learning for image recognition. *IEEE Trans. Image Process.*, 2020.

[9] Yang Fan, Fei Tian, Tao Qin, Jiang Bian, and Tie-Yan Liu. Learning what data to learn. *CoRR*, 2017.

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

[11] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *NeurIPS*, 2018.

[12] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *ICML*, 2018.

[13] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 1997.

[14] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas L. Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *ICLR (Poster)*, 2018.

[15] Ricardo Luna Gutierrez and Matteo Leonetti. Information-theoretic task selection for meta-reinforcement learning. In *NeurIPS*, 2020.

[16] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *NeurIPS*, 2019.

[17] Shell Xu Hu, Pablo Garcia Moreno, Yang Xiao, Xi Shen, Guillaume Obozinski, Neil D Lawrence, and Andreas C Damianou. Empirical bayes transductive meta-learning with synthetic gradients. In *ICLR*, 2020.

[18] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *CVPR*, 2019.

[19] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. Self-paced curriculum learning. In *AAAI*, 2015.

[20] 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 *ICML*, 2018.

[21] J Kaddour, S Sæmundsson, and MP Deisenroth. Probabilistic active meta-learning. In *NeurIPS*, 2020.

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

[23] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, 2015.

[24] M Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. *NeurIPS*, 2010.

[25] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *CVPR*, 2019.

[26] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *CVPR*, 2019.

[27] Huai-Yu Li, Weiming Dong, Xing Mei, Chongyang Ma, Feiyue Huang, and Bao-Gang Hu. Lgm-net: Learning to generate matching networks for few-shot learning. In *ICML*, 2019.

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

[29] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *TPAMI*, 2020.

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

[31] 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*, 2021.

[32] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *ICLR*, 2018.

[33] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *ICML*.

[34] Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, and Adam Trischler. Rapid adaptation with conditionally shifted neurons. In *ICML*, 2018.

[35] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *CoRR*, 2018.

[36] Jaehoon Oh, Hyungjun Yoo, ChangHwan Kim, and Se-Young Yun. BOIL: towards representation change for few-shot learning. In *ICLR*, 2021.

[37] Boris N. Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. TADAM: task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, 2018.

[38] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of MAML. In *ICLR*, 2020.

[39] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *NeurIPS*, 2019.

[40] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

[41] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, 2018.

[42] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *ICLR*, 2018.

[43] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, 2016.

[44] Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016.

[45] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[46] Qianru Sun, Yaoyao Liu, Zhaozheng Chen, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning through hard tasks. *CoRR*, 2019.

[47] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *CVPR*, 2019.

[48] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018.

[49] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, 2016.

[50] Yixin Wang, Alp Kucukelbir, and David M. Blei. Robust probabilistic modeling with bayesian data reweighting. In *ICML*, 2017.

[51] Olga Wichrowska, Niru Maheswaranathan, Matthew W. Hoffman, Sergio Gomez Colmenarejo, Misha Denil, Nando de Freitas, and Jascha Sohl-Dickstein. Learned optimizers that scale and generalize. In *ICML*, 2017.

[52] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In *NeurIPS*, 2018.

[53] Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *ICML*, 2015.

Task Attended Meta-Learning for Few-Shot Learning Supplementary Material

1 Experiments and Results

1.1 Implementation Details

We use the architecture from [1] for the base model and a two-layer LSTM [4] for the parametric optimizer. The task attention module is a ReLU activated neural network with a 1×1 convolutional layer followed by 2 fully connected layers with 32 neurons, and finally a softmax activation to generate the 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. 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. The hyperparameter λ for TAML (Theil) is sampled from a log uniform distribution over the range of $\left[1e^{-2}, 1\right]$. The number of adaptation steps is fixed to 5 for all settings except for 10-way 5-shot setting, where we use 2 adaptation steps owing to the computational expenses. The meta-batch size is set to 4 for all settings [1, 2]. 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 [3].

Table 1: Comparison of few-shot classification performance of MAML and TA-MAML on miniImageNet dataset with meta-batch size 6 for 5 and 10-way (1 and 5-shot) settings. The $\pm$ represents the 95% confidence intervals over 300 tasks. Algorithms denoted by * are rerun on their optimal hyper-parameters. 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. The hardware constraint restricts the study on a 10-way 5-shot setting, and meta-batch size of 8 or higher.

	Test Accuracy (%)			
	5-Way		10-Way	
Model	1 Shot	5 Shot	1 Shot	5 Shot
miniImageNet (Batch Size 6)				
MAML*	47.72 ± 1.041	63.45 ± 1.083	31.55 ± 0.626	Out of memory
TA-MAML	$\mathbf{49.14 \pm 1.211}$	$\mathbf{65.26 \pm 0.956}$	$\mathbf{32.62 \pm 0.635}$	Out of memory

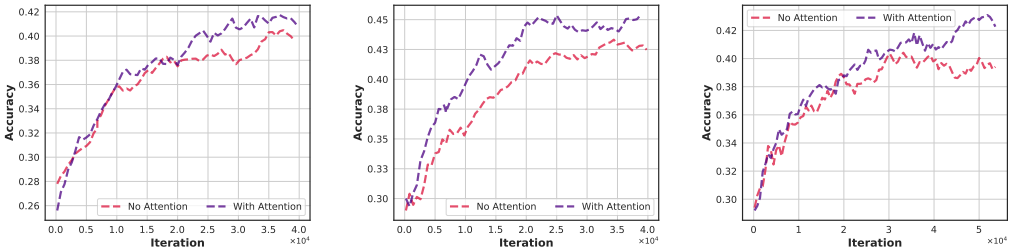
1.2 Influence of Task Attention on Meta-Training

Figure 1 describes the mean validation accuracy over 300 tasks on tieredImageNet dataset for a 5-way 1-shot setting across training iterations. It is observed 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.

1.3 Ablation Studies

We analyze the ranks of the tasks for maximum and minimum values of : loss, loss ratio, accuracy, and grad norm in a batch wrt attention weights throughout meta-training of TA-MAML on a 5-way

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



(a) MAML (b) MetaSGD (c) MetaLSTM++

Figure 1: [Best viewed in color] Mean validation accuracies of a) MAML b) MetaSGD c) MetaLSTM++ across 300 tasks with/without attention on 5-way 1-shot setting on tieredImageNet dataset.

1 and 5 shot settings on miniImagenet dataset (Figure 2 and 3). Specifically, the highest weighted task is given rank one, and the least weighted task in a batch is given the last rank. We observe that the TA module does not assign maximum weight to the tasks with maximum or minimum values of : test loss, loss ratio, grad norm or accuracy throughout meta-training. Thus, the TA module does not trivially learn to assign weights to the tasks based on some component of meta-information but learns useful latent information from all the components to assign importance for the tasks in a batch. For the sake of completeness, we repeat the ablation study on inputs of the attention module of TA-MAML for 5-way 5-shot setting on miniImagenet dataset. Table 2 describes the impact of the individual components of meta-information on meta-test performance.

Table 2: Effect of ablating components of meta-information in TA-MAML for 5-way 5-shot setting on miniImagenet dataset.

Ablation on inputs				
Grad norm	Loss	Loss ratio	Accuracy	Test Accuracy
×	×	×	×	60.16 ± 0.17
✓	✓	✓	×	60.48 ± 0.16
✓	✓	×	✓	62.17 ± 0.17
✓	×	✓	✓	60.90 ± 0.20
×	✓	✓	✓	61.52 ± 0.16
✓	✓	✓	✓	62.48 ± 0.18

We also study the correlation between the meta-information components and weights assigned by the task attention module for TA-MAML on 5-way 5-shot setting for miniImageNet dataset. From Figure 4, 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. It is 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 (Main paper - Section IV(A)).

We illustrate the trend of mean weighted loss across iterations for TA-MAML on 5-way 1 and 5 shot settings on miniImagenet dataset (Figure 5). The trend indicates that the average weighted loss decreases over the meta-training iterations. The shaded region represents a 95% confidence interval over 100 tasks.

1.4 Analysis of Attention Network

To demonstrate the functionality of the task attention module, figure 6 shows the trend of the attention-vector during meta-training of TA-MAML for 5-way 5-shot setting on miniImageNet dataset. We observe that the attention module initially assigns uniform priorities to all tasks to learn the generic meta-knowledge and subsequently inclines towards the tasks aligned more towards optimal meta-knowledge.

We attempt to decipher the functioning of the black box attention network by analyzing the qualitative relation among weights and the classes of task batches (figure 7). We observe that the tasks containing

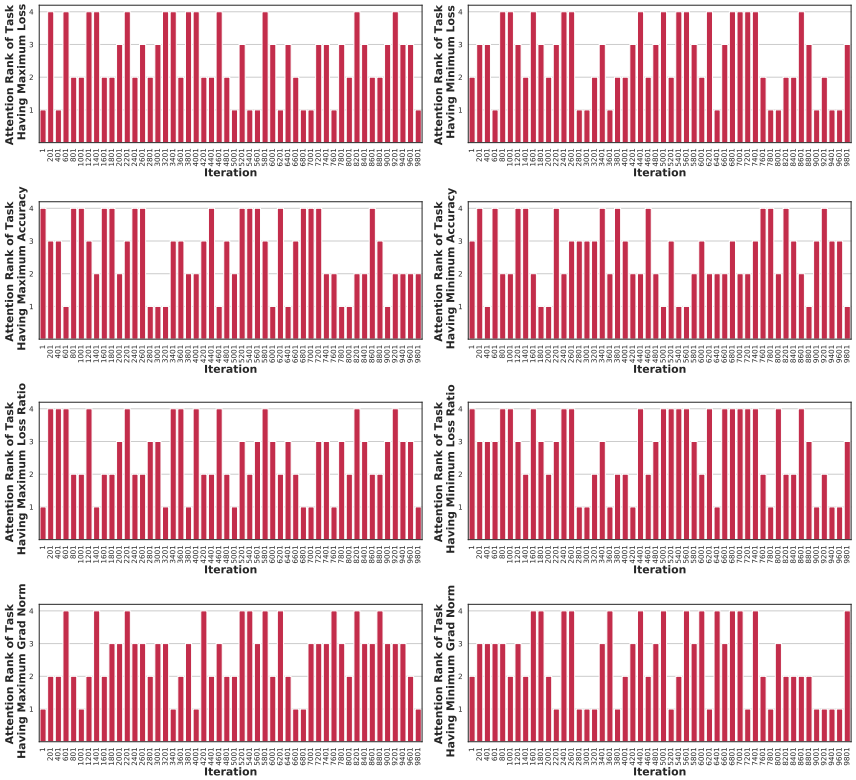


Figure 2: Rank Analysis of tasks for maximum and minimum values of : loss, loss-ratio, accuracy and grad norm throughout the training of TA-MAML for 5-way 1-shot setting on miniImagenet dataset.

images from similar classes are hard to distinguish and given more weight. In figure 7(a) 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 7(b)) in which the distribution of weights cannot be explained qualitatively. More supporting examples corresponding to both cases are described in figure 8.

References

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

[2] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *CVPR*, 2019.

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[4] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

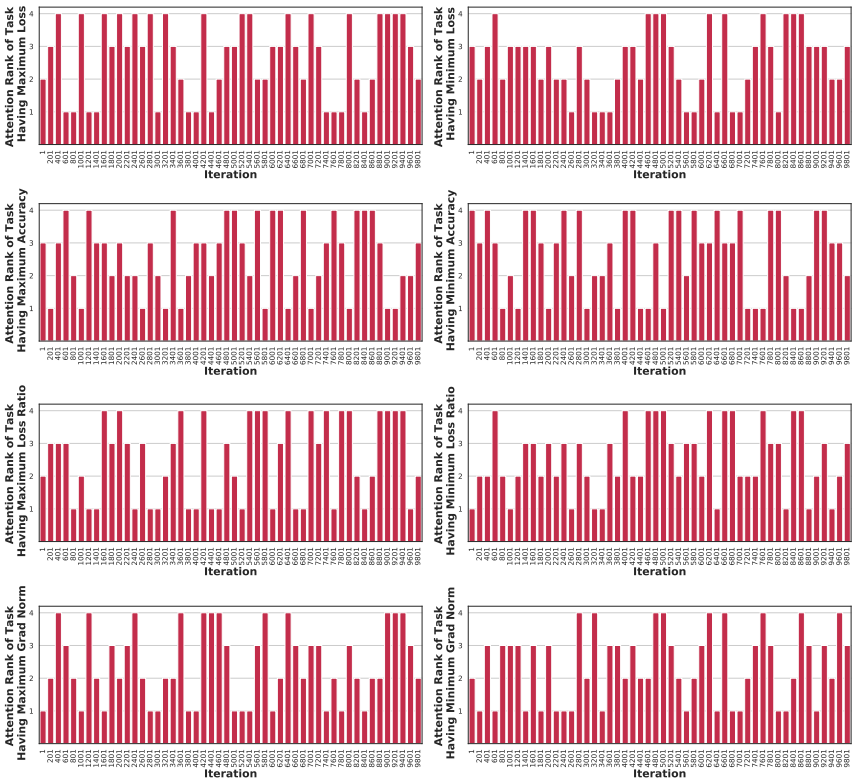


Figure 3: Rank Analysis of tasks for maximum and minimum values of : loss, loss-ratio, accuracy and grad norm throughout the training of TA-MAML for 5-way 5-shot setting on miniImagenet dataset.

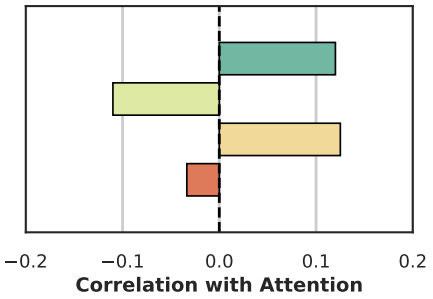


Figure 4: Mean Pearson correlation of TA-MAML on 5-way 5-shot setting on miniImagenet.

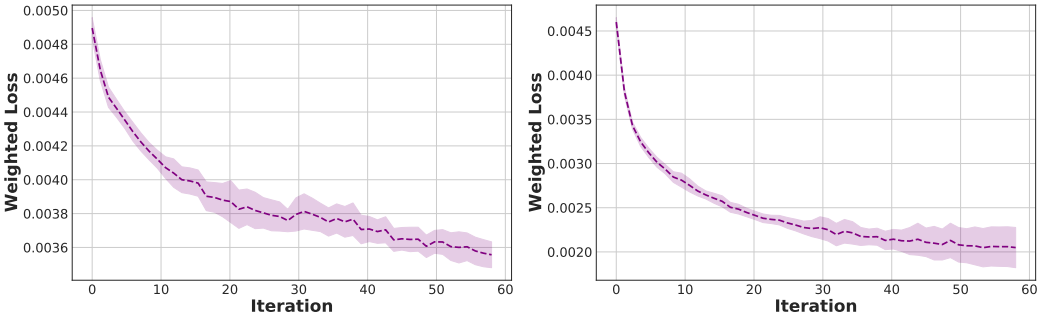


Figure 5: Trend analysis of weighted loss across meta-training iterations for TA-MAML on 5-way 1-shot (left) and 5-shot (right) settings on miniImagenet dataset.

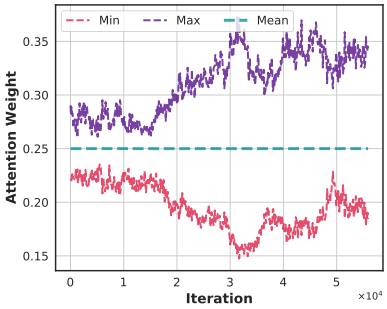


Figure 6: Trend of an attention vector in a 5-way 5-shot setting on miniImageNet dataset for TA-MAML.

(a) (b)

Figure 7: Explanations of TA module in TA-MAML on miniImagenet. a) Higher weights accredited to tasks with comparable classes b) Association of weights and task data is qualitatively uninterpretable.

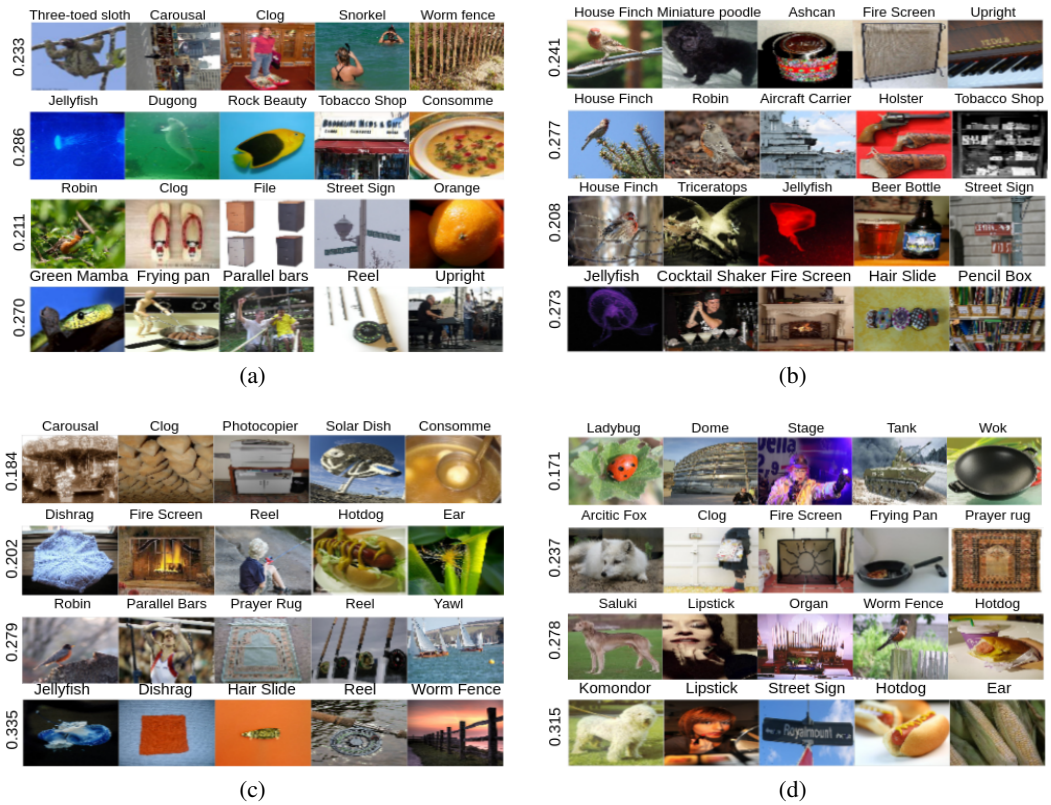


Figure 8: More examples on explanations of TA module in TA-MAML on miniImagenet. (a-b) Higher weights accredited to tasks with comparable classes (c-d) Association of weights and task data is qualitatively uninterpretable.