# RETHINKING THE TRAINING SHOT NUMBER IN ROBUST MODEL-AGNOSTIC META-LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Model-agnostic meta-learning (MAML) has been successfully applied to few-shot learning, but is not naturally robust to adversarial attacks. Previous methods attempted to impose robustness-promoting regularization on MAML's bi-level training procedure to achieve an adversarially *robust* model. They follow the typical MAML practice where training shot number is kept the same with test shot number to guarantee an optimal novel task adaptation. However, as observed by us, introducing robustness-promoting regularization into MAML reduces the intrinsic dimension of features, which actually results in a mismatch between meta-training and meta-testing in terms of affordable intrinsic dimension. Consequently, previous robust MAML methods sacrifice clean accuracy a lot. In this paper, based on our observations, we propose a simple strategy to mitigate the intrinsic dimension mismatch resulted by robustness-promoting regularization, *i.e.*, increasing the number of training shots. Though simple, our method remarkably improves the clean accuracy of MAML without much loss of robustness. Extensive experiments demonstrate that our method outperforms prior arts in achieving a better trade-off between accuracy and robustness. Besides, we observe our method is less sensitive to the number of fine-tuning steps during meta-training, which allows for a reduced number of fine-tuning steps to improve training efficiency.

## 1 INTRODUCTION

Few-shot learning (Finn et al., 2017; Rajasegaran et al., 2020; Li et al., 2021; Dong et al., 2022) aims to train a model which can fast adapt to novel classes with only a few examples. Model-agnostic meta-learning (MAML) (Finn et al., 2017) is a typical meta-learning approach to deal with few-shot learning problems. However, the model trained through MAML is not robust to adversarial attacks. The conventional adversarial training can facilitate MAML with adversarial robustness. However, the limited data in the few-shot setting makes it challenging (Goldblum et al., 2020) to keep both clean accuracy and robustness at a high level at the same time.

In recent years, a series of works (Yin et al., 2018; Goldblum et al., 2020; Wang et al., 2021) paid attention to the robust MAML. Most of them attempted to introduce robustness-promoting regularization (*i.e.*, adversarial loss) into the typical MAML bi-level training framework. For example, compared to MAML, AQ (Goldblum et al., 2020) replaced the clean loss on query data with the adversarial loss to promote the robustness. ADML (Yin et al., 2018) added one more optimization pathway, which minimizes adversarial loss at the fine-tuning stage while minimizing clean loss for meta update. R-MAML (Wang et al., 2021) demonstrated that there is no need to add adversarial loss on support images at the fine-tuning stage and imposed both clean loss and adversarial loss on query data for meta update. We fairly compare the performance of those methods (see Table 1) and find that although those methods imposed adversarial loss in different ways, their performance (*i.e.*, the clean accuracy and robustness) are actually comparable. Compared to plain MAML, all of those robust MAML methods greatly sacrifice the clean accuracy to improve the robustness. Thus, a natural question arises: what is the underlying factor that causes the severe loss of clean accuracy?

Previous robust MAML methods (Goldblum et al., 2020; Yin et al., 2018; Wang et al., 2021) followed the typical meta-learning practice, *i.e.*, matching the training shot number with the test shot number (Cao et al., 2019). However, as illustrated in Table 2, we observe that introducing robustness-promoting regularization, *i.e.*, adversarial loss, into MAML framework reduces the in-

Table 1: Accuracy of 5-way 1-shot models trained by MAML, ADML, AQ, R-MAML and our ITS-MAML on miniImageNet (Vinyals et al., 2016). The $\mathcal{A}_{clean}$ and $\mathcal{A}_{adv}$ denote clean accuracy and robust accuracy respectively. We adopt a 10-step PGD attack (Madry et al., 2017) with power $\epsilon = 2$. The $w^c$ and $w^a$ are the weights of clean loss and adversarial loss, respectively. The original setting of R-MAML is approximately equivalent to $w^c : w^a = 1 : 0.2$ for higher clean accuracy. We also report the results of R-MAML with $w^c : w^a = 1 : 1$ for a fair comparison with other methods.

| Method | $\mathcal{A}_{clean}$ | $\mathcal{A}_{adv}$ |
|---|---|---|
| MAML | **45.00**% | 0.60% |
| ADML | 37.77% | 27.15% |
| AQ | 34.69% | 28.08% |
| R-MAML | 41.58% | 22.92% |
| R-MAML ($w^c : w^a = 1 : 1$) | 37.16% | 27.87% |
| **ITS-MAML (ours)** | **41.68**% | **28.12**% |

Table 2: Intrinsic dimension of models trained by plain MAML and robustness-regularized MAML with different numbers of training shots. Introducing adversarial loss into MAML reduces the intrinsic dimension of clean samples, which may result in the decrease of clean accuracy.

| Method | Number of training shots | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| MAML | 80 | 103 | 139 | 151 | 157 | 160 | 179 |
| Robust MAML | 22 | 41 | 59 | 60 | 71 | 86 | 89 |

trinsic dimension[1] of features. This actually results in *a mismatch between meta-training and meta-testing in terms of affordable intrinsic dimension*. Such mismatch has been demonstrated to be harmful for the few-shot learning accuracy in conventional meta-learning framework (Cao et al., 2019). This observation may explain why previous robust MAML methods fail to achieve a high clean accuracy.

In this paper, we propose a simple way, *i.e.*, *increasing the number of training shots*, to mitigate the intrinsic dimension mismatch resulted by robustness-promoting regularization. Though simple, our method improves clean accuracy of robust MAML remarkably without much loss of robustness. Extensive experiments on miniImageNet (Vinyals et al., 2016), CIFAR-FS (Bertinetto et al., 2018), and Omniglot (Lake et al., 2015) demonstrate that our method performs favourably against previous robust MAML methods considering both clean accuracy and robustness. We also show that our method can achieve a better trade-off between clean accuracy and robustness. Finally, we demonstrate that compared to previous robust MAML methods, our method is less sensitive to the number of fine-tuning (inner-loop) steps in meta-training, and bares almost no drop in accuracy even when the number of fine-tuning steps is greatly reduced, thus improving training efficiency.

In a nutshell, our contributions can be summarized as follows: **1)** We observe that introducing robustness-promoting regularization into MAML framework reduces the intrinsic dimension of features, which may result in a mismatch between meta-training and meta-testing in terms of affordable intrinsic dimension. Such mismatch may harm the clean accuracy of a robust MAML framework. **2)** Based on our observations, we propose a simple yet effective strategy, *i.e.*, increasing the number of training shots, to mitigate the mismatch resulted by introducing robustness-promoting regularization. **3)** Extensive experiments on three few-shot learning benchmarks, *i.e.*, miniImageNet (Vinyals et al., 2016), CIFAR-FS (Bertinetto et al., 2018), and Omniglot (Lake et al., 2015), demonstrate that our method performs favourably against previous robust MAML methods considering both clean accuracy and robustness. We demonstrate that our method can achieve a better trade-off between clean accuracy and robustness and may improve the training efficiency by reducing the fine-tuning steps at the meta-training stage.

[1]The intrinsic dimension means the number of variables needed in a minimal representation of the data.

## 2 RELATED WORK

**Adversarial robustness.** Adversarial robustness refers to the accuracy of the model on adversarially perturbed samples, which are visually indistinguishable from clean samples but can drastically change the model predictions (Ilyas et al., 2019; Xie & Yuille, 2019). Adversarial training is one of the most effective approaches to improve the model's adversarial robustness (Goodfellow et al., 2014; Madry et al., 2017; Zhang et al., 2019b). For example, Goodfellow et al. (2014) proposed for the first time to adopt single-step-based adversarial samples for adversarial training, and Madry et al. (2017) further extended it to multi-step-based adversarial examples for better robustness. Others propose methods for fast adversarial training to improve the training efficiency (Shafahi et al., 2019; Zhang et al., 2019a; Wong et al., 2020). Methods have been proposed to improve the transferablity of robustness in few-shot settings (Chen et al., 2020; Chan et al., 2020; Tian et al., 2020; Rizve et al., 2021). However, learning an adversarially robust meta-model is challenging. The improvement of robustness is often accompanied by the decline of accuracy since a fundamental trade-off between the clean and adversarial distributions exists, as is theoretically proved by (Zhang et al., 2019b). The scarce data in few-shot settings makes it harder to learn the robust decision boundary, resulting in even more vulnerability of the model against adversarial attacks (Xu et al., 2021).

**Adversarially robust model-agnostic meta-learning.** Meta-learning has demonstrated promising performance in few-shot learning (Snell et al., 2017; Huang et al., 2018; Maicas et al., 2018; Sung et al., 2018; Wang et al., 2020). MAML (Finn et al., 2017), as the first to propose an effective meta-learning framework, can learn a well-initialized meta-model to quickly adapt to new few-shot classification tasks. Though widely adopted, MAML naturally lacks adversarial robustness. A few work studied the robustness of MAML including ADML (Yin et al., 2018), AQ (Goldblum et al., 2020) and R-MAML (Wang et al., 2021). However, these methods do not achieve a satisfactory trade-off between clean accuracy and adversarial robustness. For example, AQ trades off accuracy for robustness compared with ADML, while R-MAML achieves a high clean accuracy on condition that the robustness is greatly reduced compared with AQ (see Table 1). In addition, all these methods greatly sacrifice clean accuracy compared with plain MAML.

In this paper, we find that the number of training shots plays an important role in learning a robust yet accurate model. It is possible for robust MAML methods to achieve a significant clean accuracy improvement without much loss of robustness, simply by increasing the number of training shots. As this practice is different from that of plain MAML where the training shot number usually matches with the test shot number for an optimal novel task adaptation ability, we provide insights into why increasing the number of training shots works in the context of robust MAML.

## 3 METHODOLOGY

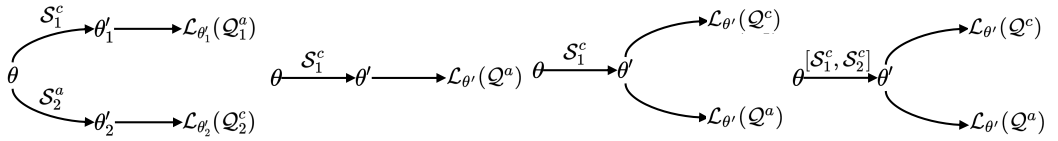
### 3.1 PRELIMINARY

**MAML for few-shot learning.** Few-shot learning aims to enable the model to classify data from novel classes with only a few (*e.g.*, 1) examples to train. Few-shot learning is typically formulated as a $N$-way $K$-shot classification problem, where in each task, we aim to classify samples (denoted as "query") into $N$ different classes, with $K$ samples (denoted as "support") in each class for training. Meta-learning is the conventional way to deal with the few-shot learning problem, while model-agnostic meta-learning (MAML) is one of the most popular and effective meta-learning methods.

Generally speaking, MAML attempts to learn a well-initialized model (*i.e.*, a meta-model), which can quickly adapt to new few-shot classification tasks. During meta-training, the meta-model is fine-tuned over $N$ classes (with each class containing $K$ samples), and then updated by minimizing the validation error of the fine-tuned network over unseen samples from these $N$ classes. The *inner fine-tuning* stage and the *outer meta-update* stage form the *bi-level* learning procedure of MAML. Formally, we consider $T$ few-shot classification tasks, each of which contains a support data set $\mathcal{S} = \{s_1, s_2, \cdots, s_K\}$ ($s_j \in \mathcal{S}$ denotes the $j$-th support sample) for fine-tuning, and a query data set $\mathcal{Q}$ for meta-update. MAML's bi-level optimization problem can be then described as:

$$\underset{\theta}{\text{minimize}} \frac{1}{T} \sum_{i=1}^{T} \mathcal{L}_{\theta_i'}(\mathcal{Q}_i), \text{ subject to } \theta_i' = \arg\min_{\theta} \mathcal{L}_{\theta}(\mathcal{S}_i), \forall i \in \{1, 2, \ldots, T\}, \tag{1}$$

where $\theta$ is the meta-model to be learned, $\theta_i'$ is the fine-tuned parameters for the $i$-th task, $\mathcal{L}_{\theta}(\mathcal{S}_i)$ and $\mathcal{L}_{\theta_i'}(\mathcal{Q}_i)$ are the training error on the support set and the validation error on the query set,




(a) ADML (b) AQ (c) R-MAML (d) ITS-MAML (ours)

Figure 1: Robust meta-learning frameworks of previous arts and our method. $\theta$ denotes the initial model parameters in each episode, $\theta'$ denotes the model parameters after fine-tuning on the support set (inner-loop), and $\mathcal{L}$ is the loss of the fine-tuned model on the query set, adopted for the meta-update of $\theta$ (outer-loop). $\mathcal{S}$ and $\mathcal{Q}$ are the support data and the query data, with the superscripts $c$ and $a$ denoting clean and adversarial samples respectively. The subscripts of $\mathcal{S}$, $\mathcal{Q}$ and $\theta'$ denote the index for different shots. $[\cdot, \cdot]$ denotes the data concatenation. Note that each episode consists of several few-shot classification tasks. We omit the task index in our symbols for brevity.

respectively. Note that the fine-tuning stage (corresponding to the constraint in Eq. 1) usually calls for $M$ steps of gradient update:

$$\theta_i^{(m)} = \theta_i^{(m-1)} - \alpha \nabla_{\theta_i^{(m-1)}} \mathcal{L}_{\theta_i^{(m-1)}}(\mathcal{S}_i),\ m \in \{1, 2, \ldots, M\}, \tag{2}$$

where $\alpha$ is the learning rate of the inner update, $\theta_i^{(0)} = \theta$ and $\theta_i^{(M)} = \theta_i'$.

**Robustness-promoting MAML.** Adversarial training is one of the most effective defense methods to learn a robust model against adversarial attacks (Madry et al., 2017). Suppose $\mathcal{D} = \{\mathcal{D}^c, \mathcal{D}^a\}$ denotes the set of samples used for training, and the adversarial training can be represented as

$$\underset{\theta}{\text{minimize}} \quad w^c \cdot \mathcal{L}_\theta(\mathcal{D}^c) + w^a \cdot \mathcal{G}_\theta(\mathcal{D}^a), \tag{3}$$

where $\theta$ is the parameters of robust model to be learned, $\mathcal{L}_\theta(\mathcal{D}^c)$ is the prediction loss (*e.g.*, cross-entropy loss) on clean sample set $\mathcal{D}^c$, $\mathcal{G}_\theta(\mathcal{D}^a)$ is the adversarial loss (*e.g.*, cross entropy loss (Goodfellow et al., 2014; Madry et al., 2017) and KL divergence (Zhang et al., 2019b)) on adversarial sample set $\mathcal{D}^a$, and $w^c$ and $w^a$ are the weights of clean loss and adversarial loss respectively. Each adversarial sample $x^a \in \mathcal{D}^a$ is obtained by adding perturbations to the clean sample $x \in \mathcal{D}^c$ to maximize its classification loss:

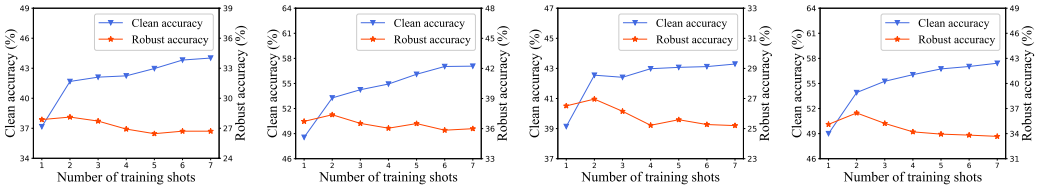
$$x^a = x^c + \underset{\|\delta\|_p \leq \epsilon}{\arg\max}\, \mathcal{G}_\theta(x^c + \delta), \tag{4}$$

where the $p$-norm of the perturbation $\delta$ is limited within the $\epsilon$ bound so that the adversarial samples are visually indistinguishable from the clean samples.

Existing robust MAML methods applied robustness-promoting regularization to MAML's bi-level learning procedure. As illustrated in Fig. 1, AQ (Goldblum et al., 2020) directly replaced the loss on clean query images of plain MAML with adversarial loss. Compared to AQ, ADML (Yin et al., 2018) additionally added another optimization pathway, *i.e.*, fine-tuning with the adversarial loss on support data and evaluating the accuracy on clean query data. Further, R-MAML (Wang et al., 2021) showed that there is no need to impose adversarial loss during the fine-tuning stage and imposed both the cross-entropy loss on clean and adversarial query images. Although R-MAML demonstrated superior clean accuracy compared to ADML, we find that if we treat the clean loss and adversarial loss on query data equally, *i.e.*, setting the learning rate of clean loss to that of adversarial loss, the clean accuracy of R-MAML actually performs comparably to ADML (as shown in Table 1). Thus, R-MAML actually improves clean accuracy by sacrificing robustness (*i.e.*, increasing the learning rate of clean loss). There still remains a question about how to improve the clean accuracy while maintaining good robustness.

### 3.2 Increasing training shot numbers for robust yet accurate MAML

As illustrated in Table 2, we observe that introducing robustness-promoting regularization reduces the intrinsic dimensions of features, which results in a mismatch between meta-training and meta-testing in terms of intrinsic dimension. Such mismatch may degenerate the clean accuracy a lot (Cao et al., 2019), which explains the huge loss of clean accuracy in previous robust MAML frameworks.




(a) 1-shot (miniImageNet) (b) 5-shot (miniImageNet) (c) 1-shot (CIFAR-FS) (d) 5-shot (CIFAR-FS)

Figure 2: Clean accuracy and robust accuracy of models on 5-way 1-shot and 5-way 5-shot meta-testing tasks. The experiments are conducted on miniImageNet and CIFAR-FS. The models are trained with different numbers of training shots. The robust accuracy is computed with a 10-step PGD attack with $\epsilon = 2$ for miniImageNet and $\epsilon = 8$ for CIFAR-FS.

Based on these observations, we propose a simple yet effective way, *i.e.*, increasing the number of training shots, to mitigate the mismatch resulted by adding robustness-promoting regularization, *i.e.*, adversarial loss. Formally, for a $N$-way $K$-shot meta-testing task, our method can be expressed as

$$\underset{\theta}{\text{minimize}} \frac{1}{T}\sum_{i=1}^{T} w^c \cdot \mathcal{L}_{\theta_i'}(\mathcal{Q}^c) + w^a \cdot \mathcal{L}_{\theta_i'}(\mathcal{Q}^a), \text{ subject to } \theta_i' = \arg\min_{\theta} \mathcal{L}_{\theta}(\tilde{\mathcal{S}}_i^c), \tag{5}$$

where $\tilde{\mathcal{S}}_i^c = \{s_1, s_2, \cdots, s_{\tilde{K}}\}$ denotes the support set for the $i$-th task during meta-training. In our method, the number of support images (*i.e.*, training shot number) $\tilde{K}$ for each task at meta-training stage is set to be larger than the number of support images $K$ used in meta-testing stage. For example, if we deal with the 5-way 1-shot setting, $\tilde{K}$ is set to be larger than 1. The $w^c$ and $w^a$ are the weights of clean loss $\mathcal{L}_{\theta_i'}(\mathcal{Q}^c)$ and adversarial loss $\mathcal{L}_{\theta_i'}(\mathcal{Q}^a)$, respectively. Usually, we find that $w^c = w^a = 1$ is sufficient to achieve a good trade-off between clean accuracy and robustness.

As illustrated in Table 2, we find as we increase the training shot number $\tilde{K}$, the intrinsic feature dimension of robust MAML also increases, which effectively mitigates the intrinsic dimension mismatch between meta-training and meta-testing. Correspondingly, as shown in Fig. 2, with increasing the number of training shots, the clean accuracy steadily increases before reaching a bound, while the robustness slightly decreases but still remains high. Further analysis can be found in Sec. 4.5.

# 4 Experiments

## 4.1 Setup

**Dataset.** We conduct experiments on three widely-used few-shot learning benchmarks, *i.e.*, miniImageNet (Vinyals et al., 2016), CIFAR-FS (Bertinetto et al., 2018), and Omniglot (Lake et al., 2015). The *miniImageNet* contains 100 classes with 600 samples in each class. The whole dataset is split into 64, 16 and 20 classes for training, validation and testing respectively. We adopt the training set for meta-training, and randomly select 2000 unseen tasks from the testing set for meta-testing. Each image is downsized to $84 \times 84 \times 3$ in our experiments. *CIFAR-FS* has the same dataset splitting as miniImageNet, *i.e.*, 64, 16 and 20 classes for training, validation and testing respectively, with each class containing 600 images. We also adopt the training set for meta-training, and randomly select 4000 unseen tasks from the testing set for meta-testing. Each image is resized to $32 \times 32 \times 3$. *Omniglot* includes handwritten characters from 50 different alphabets, with a total of 1028 classes of training data and 423 classes of testing data. We randomly select 2000 unseen tasks from the testing data for meta-testing. Each image has a size of $28 \times 28 \times 1$.

**Training.** We verify our method based on two kinds of architectures, *i.e.*, a four-layer convolutional neural network as in (Wang et al., 2021) and a ResNet-12 (He et al., 2016). All the models in our experiments are trained for 12 epochs unless specified. During meta-training, each episode consists of 4 randomly selected tasks. For 5-way 1-shot meta-testing tasks, the number of support images per class in each task of meta-training is 1 for previous methods, and 2 for ITS-MAML. For 5-way 5-shot meta-testing tasks, the number of support images per class in each task of meta-training is 5 for previous methods, and 6 for ITS-MAML. The number of query images per class is 15 for all methods on miniImageNet and CIFAR-FS. On Omniglot, since each class only contains 20 samples, the number of query images per class is thus set to 9 for 5-way 1-shot meta-testing tasks and 5 for 5-way 5-shot meta-testing tasks to avoid data repetition in a task. The learning rate is set to 0.01

Table 3: Accuracy of meta-models trained by different methods under different types of attacks on miniImageNet. The best results for each test case are marked in bold.

| Method | Model | 5-way 1-shot accuracy (%) | | | | 5-way 5-shot accuracy (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Clean | FGSM | PGD | CW | Clean | FGSM | PGD | CW |
| MAML | 4-layer CNN | **45.00** | 3.71 | 0.60 | 0.24 | **64.28** | 9.63 | 0.96 | 0.33 |
| ADML | 4-layer CNN | 37.77 | 29.96 | 27.15 | 26.79 | 56.02 | 41.96 | 35.90 | 35.66 |
| AQ | 4-layer CNN | 34.69 | 30.53 | 28.08 | 26.20 | 50.28 | 40.39 | 36.21 | 36.26 |
| R-MAML | 4-layer CNN | 37.16 | 29.95 | 27.87 | 26.13 | 55.85 | 42.63 | 36.30 | 34.93 |
| **ITS-MAML (2-shot)** | 4-layer CNN | 41.68 | **31.74** | **28.12** | **27.84** | 53.37 | **43.68** | **37.38** | **38.22** |
| **ITS-MAML (6-shot)** | 4-layer CNN | **43.82** | 30.60 | 26.71 | 26.34 | **57.03** | 42.81 | 35.84 | 35.40 |
| MAML | ResNet-12 | **47.32** | 8.46 | 0.49 | 0.96 | **69.90** | 14.18 | 1.83 | 7.20 |
| ADML | ResNet-12 | 40.08 | 34.56 | 31.48 | 29.84 | 61.80 | 47.77 | 45.14 | 44.13 |
| AQ | ResNet-12 | 35.55 | 31.64 | 29.27 | 29.79 | 58.27 | 46.25 | 43.50 | 44.84 |
| R-MAML | ResNet-12 | 39.92 | 33.82 | 29.36 | 30.23 | 62.63 | 47.92 | 44.28 | 46.30 |
| **ITS-MAML (2-shot)** | ResNet-12 | 42.87 | **35.29** | **32.60** | **31.71** | 60.35 | **50.33** | **48.19** | **48.21** |
| **ITS-MAML (6-shot)** | ResNet-12 | **45.70** | 32.96 | 28.24 | 29.85 | **63.53** | 47.08 | 43.87 | 45.10 |

Table 4: Accuracy of meta-models trained by different methods under different types of attacks on CIFAR-FS. A 4-layer CNN is adopted for all methods.

| Method | 5-way 1-shot accuracy (%) | | | | 5-way 5-shot accuracy (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | Clean | FGSM | PGD | CW | Clean | FGSM | PGD | CW |
| MAML | **49.93** | 9.57 | 0.15 | 0.08 | **65.63** | 18.18 | 0.70 | 1.22 |
| ADML | 40.41 | 36.67 | 26.05 | 25.79 | 56.52 | 49.63 | 32.08 | 32.86 |
| AQ | 32.08 | 30.69 | 26.49 | 23.08 | 51.40 | 48.19 | 35.08 | 33.90 |
| R-MAML | 39.14 | 35.79 | 26.51 | 25.77 | 56.09 | 50.78 | 32.67 | 34.38 |
| **ITS-MAML (2-shot)** | 42.55 | **38.52** | **26.96** | **27.60** | 53.91 | **51.46** | **36.47** | **37.21** |
| **ITS-MAML (6-shot)** | **43.12** | 38.38 | 25.37 | 26.24 | **57.12** | 51.07 | 33.62 | 34.05 |

for fine-tuning, and 0.001 for the meta-update. Following (Wang et al., 2021), the number of fine-tuning steps is set to 5 for meta-training, and 10 for meta-testing for all methods unless specified. As (Wang et al., 2021) showed that adopting the FGSM attack (Goodfellow et al., 2014) instead of the PGD attack (Madry et al., 2017) during meta-training can improve the training efficiency without significantly affecting the model performance, we also adopt the FGSM attack (Goodfellow et al., 2014) as the training attack. The training attack power $\epsilon$ is set to 2 for miniImageNet, and 10 for CIFAR-FS and Omniglot. We evaluate our method under different kinds of attacks for testing. The testing attack power is 2 for miniImageNet, 8 for CIFAR-FS and 10 for Omniglot unless specified. For a fair comparison, we set $w^c : w^a = 1 : 1$ in Eq. 5 for all methods unless specified.

### 4.2 Comparisons with previous robust MAML methods

We evaluate the performance of our proposed method (denoted as "ITS-MAML") under 5-way 1-shot and 5-way 5-shot settings, and compare our method with plain MAML method and previous typical robust MAML methods, *i.e.*, AQ, ADML and R-MAML. We demonstrate the experimental results on three benchmarks in Tables 3, 4 and 5, respectively. In those tables, we show the accuracy on clean images ("Clean") and robust accuracy under different types of attacks, *i.e.*, FGSM (Goodfellow et al., 2014), PGD (Madry et al., 2017), and CW (Carlini & Wagner, 2017). The accuracy under different types of attacks reflects the model's robustness level.

Based on the experiment results, we have three observations:

1) Compared to MAML, all the robust MAML methods (including ours) significantly improve the model's robustness. For example, for 4-layer CNN on 5-way 1-shot task of miniImageNet, compared to MAML baseline, AQ, ADML, R-MAML and ITS-MAML improve the accuracy under PGD

Table 5: Accuracy of meta-models trained by different methods under different types of attacks on Omniglot. A 4-layer CNN is adopted for all methods.

| Method | 5-way 1-shot accuracy (%) | | | | 5-way 5-shot accuracy (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | Clean | FGSM | PGD | CW | Clean | FGSM | PGD | CW |
| MAML | 93.02 | 65.20 | 22.72 | 26.64 | 97.78 | 91.75 | 60.25 | 54.39 |
| ADML | 90.58 | 89.84 | 78.27 | 77.19 | 97.46 | 96.53 | 91.06 | **91.50** |
| AQ | 90.09 | 88.72 | 81.69 | 80.92 | 97.02 | 96.44 | 91.25 | 90.11 |
| R-MAML | 89.60 | 87.65 | 77.44 | 77.57 | 97.12 | 96.09 | 90.28 | 89.72 |
| **ITS-MAML (2-shot)** | 93.56 | **91.43** | **85.45** | **83.68** | 97.02 | **96.80** | **91.87** | 91.46 |
| **ITS-MAML (6-shot)** | **94.23** | 86.89 | 79.90 | 77.34 | **98.19** | 96.19 | 90.77 | 88.52 |

attack by about 27%, 27%, 27% and 28%, respectively. These results verify introducing robustness-promoting regularization (*i.e.*, adversarial loss) into MAML contributes to the model's robustness.

2) By comparing previous robust MAML methods (*e.g.*, ADML and R-MAML), we find that generally they achieve comparable results for both clean accuracy and robustness, indicating whether applying the adversarial loss to the fine-tuning procedure doesn't bring a huge difference. For example, for 4-layer CNN on 5-way 1-shot task of miniImageNet, the clean accuracy for ADML and R-MAML is 37.77% and 37.16% respectively, while the robust accuracy is 27.15% and 27.87% (under PGD attack). Generally, AQ performs worse than ADML and R-MAML on clean accuracy, which indicates imposing clean loss for meta-update may be necessary for a high clean accuracy.

3) Compared to previous robust MAML methods, our increasing training shot number strategy ("ITS-MAML") remarkably improves the clean accuracy while maintaining good robustness (sometimes the robustness is even better than previous methods). For example, on 5-way 1-shot task on CIFAR-FS, in terms of clean accuracy, "ITS-MAML (2-shot)" outperforms AQ and R-MAML by more than 10% and 3%. If we increase the training shot number (*i.e.*, from 2-shot to 6-shot), we observe that "ITS-MAML (6-shot)" further improves the clean accuracy by around 1%. The robust accuracy decreases slightly, but still remains at a high level, which is competitive among previous robust MAML methods. For 5-way 5-shot task, we observe the similar trend. Thus, if we care more about the model's robustness, we may lower the number of training shots. However, if we want to achieve a robust yet accurate model, increasing the training shot number is a good choice.

### 4.3 Better trade-off between clean accuracy and robustness

R-MAML (Wang et al., 2021) demonstrated superior clean accuracy compared to AQ and ADML, when the learning rate of clean loss is set 5 times the learning rate of adversarial loss, *i.e.*, the ratio $w^c : w^a$ is set to $1:0.2$ (see Eq. 5). We find that when this ratio is set to be consistent with ADML and ITS-MAML, *i.e.*, $1:1$, the high clean accuracy of R-MAML no longer exists (see Table 1). It implies that there exists a trade-off between clean accuracy and robustness with the change of this ratio. We compare our ITS-MAML with R-MAML under different ratios of $w^c : w^a$ and show corresponding clean accuracy and robustness in Fig. 3. To demonstrate the effectiveness of increasing training shot number, we also show the results of ITS-MAML under different training shot numbers (keeping $w^c : w^a = 1:1$) in Fig. 3. From Fig. 3 (left), we observe that the larger the ratio of $w^c : w^a$, the higher the clean accuracy and the lower the robust accuracy of the model. Compared with R-MAML, our ITS-MAML achieves obviously better trade-off (as shown, the curve of ITS-MAML is above that of R-MAML). In addition, Fig. 3 (right) demonstrates that with the increase of training shot numbers, the clean accuracy of ITS-MAML steadily increases before reaching a bound, while the robustness remains relatively stable at a high level, which further verifies the effectiveness of our method.

### 4.4 More efficient training

The number of fine-tuning steps (*i.e.*, $M$ in Eq. 2) during meta-training affects the training efficiency of the model. Larger $M$ leads to higher computation costs and a lower training efficiency. We expect that ITS-MAML can achieve good performance even if the number of fine-tuning steps $M$ is reduced during training, thus improving the training efficiency. To this end, we train ADML, AQ, R-MAML ($w^c : w^a = 1:0.2$ as in (Wang et al., 2021)) and ITS-MAML models with $M$ ranging from 1 to 5.

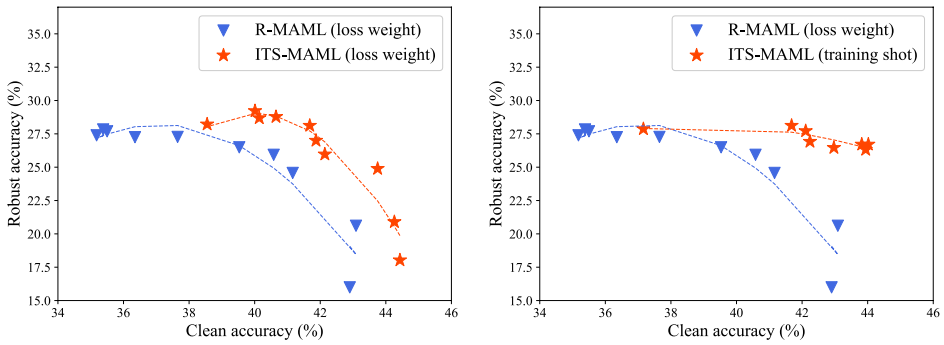



Figure 3: Clean accuracy *vs.* robust accuracy. **Left**: The results are obtained by varying the value of $w^c : w^a$ from $1:0.1$ to $1:5$ for both R-MAML and ITS-MAML ($\tilde{K} = 2$). **Right**: The results are obtained by varying the value of $w^c : w^a$ for R-MAML, and by varying the training shot number $\tilde{K}$ from 1 to 7 for ITS-MAML.

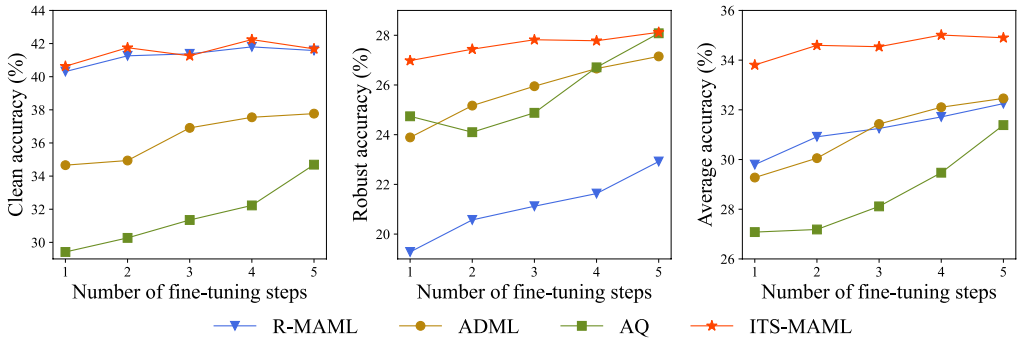



Figure 4: Accuracy of models trained by ADML, AQ, R-MAML ($w^c : w^a = 1 : 0.2$ as in (Wang et al., 2021)) and ITS-MAML with different number of fine-tuning steps during meta-training. **Left**: clean accuracy. **Middle**: Robust accuracy, which is computed with a 10-step PGD attack with $\epsilon = 2$. **Right**: Average accuracy, which denotes the average value of clean accuracy and robust accuracy.

We then evaluate the 5-way 1-shot accuracy of the models on miniImageNet. The results are shown in Fig. 4. We observe that previous methods may suffer from an obvious performance degradation (either clean accuracy or robustness) with reducing $M$. Compared to previous methods, our ITS-MAML is less sensitive to the number of fine-tuning steps during meta-training. This allows for a reduced number of fine-tuning steps to improve the training efficiency.

### 4.5 Increasing training shot number brings higher intrinsic dimension

As discussed, introducing adversarial loss into MAML framework reduces the intrinsic dimension of features, resulting in a mismatch between meta-training and meta-testing in terms of affordable intrinsic dimension. We provide more empirical evidences in Table 6. We observe that the intrinsic dimension of clean samples increases with the increase of training shot number, either for MAML or robust MAML (or ITS-MAML). As the mismatch of intrinsic dimension between meta-training and meta-testing may harm the accuracy Cao et al. (2019), for 1-shot testing task, the clean accuracy of MAML decreases as the training shot number increases (Fig. 5). Moreover, with the same number of training shots, robust MAML always learns features of clean samples with lower intrinsic dimension, which harms its clean accuracy. For example, for a 1-shot meta-testing task, in order to achieve a high clean accuracy of robust MAML, the preferred intrinsic dimension of clean embeddings of ITS-MAML should be close to that of plain MAML when the training shot number is set to 1, *i.e.*, 80 in the table. As the number of training shots increases in robust MAML, the intrinsic dimension

Table 6: Intrinsic feature dimensions of models trained by plain MAML and the proposed ITS-MAML with different numbers of training shots. The intrinsic dimension is estimated through principal component analysis (PCA) of features trained on miniImageNet (Fan et al., 2010). In our experiments, the intrinsic dimension is set to the number of principal components which retain at least $90\%$ of the variance. The "Clean", "Adversarial" and "Noise" refer to the clean samples, adversarial samples and the adversarial noise (added to the clean samples), respectively.

| Method | | Number of training shots | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| MAML | Clean | 80 | 103 | 139 | 151 | 157 | 160 | 179 |
| | Adversarial | 78 | 98 | 138 | 142 | 144 | 149 | 175 |
| | Noise | 82 | 106 | 131 | 96 | 192 | 156 | 147 |
| ITS-MAML | Clean | 22 | 41 | 59 | 60 | 71 | 86 | 89 |
| | Adversarial | 23 | 42 | 61 | 61 | 72 | 88 | 92 |
| | Noise | 4 | 7 | 7 | 8 | 8 | 9 | 10 |

of clean embeddings increases towards approaching the one of plain MAML when the training shot number is set to 1, thus achieving a remarkable improvement of clean accuracy as in Fig. 5.

We also investigate how the intrinsic dimension of adversarial noise changes as the training shot number increases. We expect the intrinsic dimension of adversarial noise should be as low as possible because it provides no useful information to classify different samples. However, we observe in Table 6 that the intrinsic dimension of noise in plain MAML is quite high (even higher than the intrinsic dimension of clean samples). This may explain why plain MAML exhibits quite worse robustness to adversarial attacks (see Table 3). In contrast, robust MAML has remarkably lower intrinsic dimension of adversarial noise, which is consistent with the superior robust accuracy performance. Further, the intrinsic dimension of adversarial samples aligns well with that of clean samples for robust MAML. However, for plain MAML, the intrinsic dimension of adversarial samples is obviously lower than that of clean samples, which indicates without adversarial loss, partial key variations of features may be destroyed by adversarial noise. Moreover, as the training shot number increases, the intrinsic dimension of adversarial noise (see "Noise" lines in the table) slightly increases. This may explain why the robustness slightly decreases for our ITS-MAML. We hope this perspective may inspire future investigation on understanding adversarial robustness.

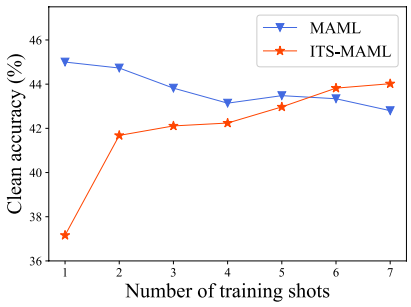



Figure 5: Clean accuracy for 1-shot testing of meta-models trained by MAML and ITS-MAML with different number of training shots.

## 5 Conclusion

In this paper, we observe that introducing adversarial loss into MAML framework reduces the intrinsic dimension of features, which results in a mismatch between meta-training and meta-testing in terms of affordable intrinsic dimension. Based on this observation, we propose a simple yet effective strategy, *i.e.*, increasing the number of training shots, to mitigate the intrinsic dimension mismatch resulted by introducing adversarial loss. Extensive experiments on few-shot learning benchmarks demonstrate that compared to previous robust MAML methods, our method can achieve superior clean accuracy while maintaining high-level robustness. Further, empirical studies show that our method may achieve a better trade-off between clean accuracy and robustness and better training efficiency. We hope our new perspective may inspire future research in the field of robust meta-learning.

## REFERENCES


Luca Bertinetto, Joao F Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *ICLR*, 2018.

Tianshi Cao, Marc T Law, and Sanja Fidler. A theoretical analysis of the number of shots in few-shot learning. In *ICLR*, 2019.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Alvin Chan, Yi Tay, and Yew-Soon Ong. What it thinks is important is important: Robustness transfers through input gradients. In *CVPR*, pp. 332–341, 2020.

Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *CVPR*, pp. 699–708, 2020.

Junhao Dong, Yuan Wang, Jian-Huang Lai, and Xiaohua Xie. Improving adversarially robust few-shot image classification with generalizable representations. In *CVPR*, pp. 9025–9034, 2022.

Mingyu Fan, Nannan Gu, Hong Qiao, and Bo Zhang. Intrinsic dimension estimation of data by principal component analysis. *arXiv preprint arXiv:1002.2050*, 2010.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pp. 1126–1135. PMLR, 2017.

Micah Goldblum, Liam Fowl, and Tom Goldstein. Adversarially robust few-shot learning: A meta-learning approach. *Advances in Neural Information Processing Systems*, 33:17886–17895, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Po-Sen Huang, Chenglong Wang, Rishabh Singh, Wen-tau Yih, and Xiaodong He. Natural language to structured query generation via meta-learning. *arXiv preprint arXiv:1803.02400*, 2018.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Junjie Li, Zilei Wang, and Xiaoming Hu. Learning intact features by erasing-inpainting for few-shot classification. In *AAAI*, volume 35, pp. 8401–8409, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Gabriel Maicas, Andrew P Bradley, Jacinto C Nascimento, Ian Reid, and Gustavo Carneiro. Training medical image analysis systems like radiologists. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 546–554. Springer, 2018.

Jathushan Rajasegaran, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Mubarak Shah. Self-supervised knowledge distillation for few-shot learning. *arXiv preprint arXiv:2006.09785*, 2020.

Mamshad Nayeem Rizve, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Exploring complementary strengths of invariant and equivariant representations for few-shot learning. In *CVPR*, pp. 10836–10846, 2021.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, pp. 1199–1208, 2018.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *ECCV*, pp. 266–282. Springer, 2020.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Guangting Wang, Chong Luo, Xiaoyan Sun, Zhiwei Xiong, and Wenjun Zeng. Tracking by instance detection: A meta-learning approach. In *CVPR*, pp. 6288–6297, 2020.

Ren Wang, Kaidi Xu, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Chuang Gan, and Meng Wang. On fast adversarial robustness adaptation in model-agnostic meta-learning. *arXiv preprint arXiv:2102.10454*, 2021.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Cihang Xie and Alan Yuille. Intriguing properties of adversarial training at scale. *arXiv preprint arXiv:1906.03787*, 2019.

Han Xu, Yaxin Li, Xiaorui Liu, Hui Liu, and Jiliang Tang. Yet meta learning can adapt fast, it can also break easily. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 540–548. SIAM, 2021.

Chengxiang Yin, Jian Tang, Zhiyuan Xu, and Yanzhi Wang. Adversarial meta-learning. *arXiv preprint arXiv:1806.03316*, 2018.

Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. *Advances in Neural Information Processing Systems*, 32, 2019a.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, pp. 7472–7482. PMLR, 2019b.