ROBUST FEW-SHOT LEARNING WITH ADVERSARIALLY QUERIED META-LEARNERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Previous work on adversarially robust neural networks requires large training sets and computationally expensive training procedures. On the other hand, few-shot learning methods are highly vulnerable to adversarial examples. The goal of our work is to produce networks which both perform well at few-shot tasks and are simultaneously robust to adversarial examples. We adapt adversarial training for meta-learning, we adapt robust architectural features to small networks for meta-learning, we test pre-processing defenses as an alternative to adversarial training for meta-learning, and we investigate the advantages of robust meta-learning over robust transfer-learning for few-shot tasks. This work provides a thorough analysis of adversarially robust methods in the context of meta-learning, and we lay the foundation for future work on defenses for few-shot tasks.

1 INTRODUCTION

For safety-critical applications like facial recognition, traffic sign detection, and copyright control, adversarial attacks pose an actionable threat (Zhao et al., 2018; Eykholt et al., 2017; Saadatpanah et al., 2019). Conventional adversarial training and pre-processing defenses aim to produce networks that resist attack (Madry et al., 2017; Zhang et al., 2019; Samangouei et al., 2018), but such defenses rely heavily on the availability of large training datasets. In applications that require *few-shot learning*, such as face recognition from few images, recognition of a video source from a single clip, or recognition of a new object from few example photos, the conventional robust training pipeline breaks down.

When data is scarce or new classes arise frequently, neural networks must adapt quickly (Duan et al., 2017; Kaiser et al., 2017; Pfister et al., 2014; Vartak et al., 2017). In these situations, *meta-learning* methods achieve few-shot learning by creating networks that learn quickly from little data and with computationally cheap fine-tuning. While state-of-the-art meta-learning methods perform well on benchmark few-shot classification tasks, these naturally trained neural networks are highly vulnerable to adversarial examples. In fact, we will see below that even robust classifiers, when adapted to a new task, fail to resist attacks unless appropriate measures are taken.

We study robust few-shot image classification by meta-learning. We begin by exploring several obvious defenses for few shot learning: adversarial training, robust architectural features, and pre-processing defenses, and find that all three provide relatively weak security in the few-shot setting. Specifically, feature denoising layers, architectural features that achieve state-of-the-art adversarial robustness on ImageNet, are not effective on the lightweight architectures used by meta-learning algorithms, and pre-processing defenses, such as DefenseGAN and image superresolution, dramatically decrease natural accuracy without achieving robustness.

We propose a new approach, called *adversarial querying*, in which the network is exposed to adversarial attacks during the query step of meta-learning. This algorithm-agnostic method produces a feature extractor that is robust, even without adversarial training during fine-tuning. In the few-shot setting, we show that adversarial querying out-performs standard defenses by a wide margin in terms of both clean accuracy and robustness.

Model	$\mathcal{A}_{nat}$	$\mathcal{A}_{adv}$
AT transfer learning (R2-D2 backbone)	39.13%	25.33%
ADML	47.75%	18.49%
Naturally Trained R2-D2	72.59%	0.00%
AQ R2-D2 (ours)	57.87%	**31.52%**

Table 1: The R2-D2 meta-learning method, adversarially trained transfer learning (ADML), and our adversarially queried (AQ) R2-D2 classifier on 5-shot Mini-ImageNet. The transfer learning model was trained on all training data (except the hold-out classes) simultaneously, and then fine-tuned on few-shot classes. All R2-D2 models are fine-tuned with a ridge regression head as in (Bertinetto et al., 2018), and we re-implement ADML from (Yin et al., 2018). Natural accuracy is denoted $\mathcal{A}_{nat}$, and robust accuracy, $\mathcal{A}_{adv}$, is computed with respect to a 20-step PGD attack as in (Madry et al., 2017) with $\epsilon = \frac{8}{255}$. A description of our training regime can be found in Appendix A.1.

2 Related Work

2.1 Learning with less data

Before the emergence of meta-learning, a number of approaches existed to cope with few-shot data. One simple approach is *transfer learning*, in which pre-trained feature extractors are created using large datasets, and then fine-tuned on new tasks using less data (Bengio, 2012). Metric learning methods avoid overfitting to the small number of training examples in new classes by instead performing classification using nearest-neighbors in feature space with a feature extractor that is trained on a large corpus of data and not re-trained when classes are added (Snell et al., 2017; Gidaris & Komodakis, 2018; Mensink et al., 2012). Metric learning methods are computationally efficient when adding many low-shot classes, since the feature extractor network is not re-trained.

Meta-learning algorithms create a "base" model that quickly adapts to new tasks by fine-tuning. This model is created using a set of training tasks $\{\mathcal{T}_i\}$ that can be sampled from a task distribution. Each task comes with *support* data, $\mathcal{T}_i^s$, and *query* data, $\mathcal{T}_i^q$. In practice, each task is taken to be a classification problem involving only a small subset of classes in a large many-class dataset. The number of examples per class in the support set is called the *shot*, so that fine-tuning on five support examples per class is 5-shot learning.

An iteration of training begins by sampling tasks $\{\mathcal{T}_i\}$ from the task distribution. The base model is fine-tuned on the support data for the sampled tasks, and then used to make predictions on the query data. Then, the base model parameters are updated to improve the accuracy of the resulting fine-tuned model. This requires backpropagation through the fine-tuning steps. See Algorithm 1 for a formal treatment.

Algorithm 1: The meta-learning framework

Require: Base model, F_θ, fine-tuning algorithm, A, learning rate, γ, and distribution over tasks, $p(\mathcal{T})$.
Initialize θ, the weights of F;
while not done **do**
  Sample batch of tasks, $\{\mathcal{T}_i\}_{i=1}^n$, where $\mathcal{T}_i \sim p(\mathcal{T})$ and $\mathcal{T}_i = (\mathcal{T}_i^s, \mathcal{T}_i^q)$.
  for $i = 1, ..., n$ **do**
    Fine-tune model on $\mathcal{T}_i$ (inner loop). New network parameters are written $\theta_i = A(\theta, \mathcal{T}_i^s)$.
    Compute gradient $g_i = \nabla_\theta \mathcal{L}(F_{\theta_i}, \mathcal{T}_i^q)$.
  Update base model parameters (outer loop): $\theta \leftarrow \theta - \frac{\gamma}{n} \sum_i g_i$

Note that the fine-tuned parameters, $\theta_i = A(\theta, \mathcal{T}_i^s)$, in the above algorithm, are a function of the base model's parameters so that the gradient computation in the outer loop may backpropagate through A. For validation after training, the base model is fine-tuned on the support set of hold-out tasks, and accuracy on the query set is reported. In this work, we report performance on OmniGlot, Mini-ImageNet, and CIFAR-FS (Lake et al., 2015; Vinyals et al., 2016; Bertinetto et al., 2018).

We focus on four meta-learning algorithms: MAML, R2-D2, MetaOptNet, and ProtoNet. (Finn et al., 2017; Bertinetto et al., 2018; Lee et al., 2019; Snell et al., 2017). During fine-tuning, MAML uses SGD to update all parameters, minimizing cross-entropy loss. Since unrolling SGD steps into a deep computation graph is expensive, a first-order variants ignore second-order derivatives. We use the original MAML formulation. R2-D2 and MetaOptNet, on the other hand, only update the final linear layer during fine-tuning, leaving the "backbone network" that extracts these features frozen at test time. R2-D2 replaces SGD with a closed-form differentiable solver for regularized ridge regression, while MetaOptNet achieves its best performance when replacing SGD with a solver for SVM. Because the objective of these linear problems is convex, differentiable convex optimizers can be efficiently deployed to find optima, and differentiate these optima with respect to the backbone features at train time. ProtoNet takes an approach inspired by metric learning. It constructs class prototypes as the centroids in feature space for each task. These centroids are then used to classify the query set in the outer loop of training. Because each class prototype is a simple geometric average of feature representations, it is easy to differentiate through the fine-tuning step.

2.2 Robust learning with less data

Several authors have tried to learn robust models in the data scarce regime. The authors of (Shafahi et al., 2019) study robustness properties of transfer learning. They find that retraining earlier layers of the network during fine-tuning impairs the robustness of the network, while only retraining later layers can largely preserve robustness. ADML is the first attempt at achieving robustness through meta-learning. ADML is a MAML variant, specifically designed for robustness, which employs adversarial training (Yin et al., 2018). However, this method for robustness is only compatible with MAML, an outdated meta-learning algorithm. Moreover, ADML is computationally expensive, and the authors only test their method against a weak attacker. We implement ADML and test it against a strong attacker. We show that our methods achieve both higher robustness and natural accuracy.

Sample results comparing baseline robust learning methods are shown in Table 1, which shows that clean meta-learning and a direct application of adversarial training to meta-learning (the ADML method) achieve low levels of robustness. While simple robust transfer learning achieves more robustness, the adversarial querying procedure does significantly better in terms of both clean and robust accuracy.

3 Evaluating the robustness of existing few-shot methods

In this section, we benchmark existing methods for robust learning with scarce data in terms of both natural and robust accuracy. Following standard practices, we assess the robustness of models by attacking them with ℓ_∞-bounded perturbations. We craft image perturbations using the projected gradient descent attack (PGD) since it has proven to be one of the most effective algorithms both for attacking as well as for adversarial training (Madry et al., 2017). This attack is a more powerful version of the one-step attack used in ADML (Yin et al., 2018). A detailed description of the PGD attack can be found in Algorithm 2. We consider perturbations of ℓ_∞ radius of $\frac{8}{255}$, and a step size of $\frac{2}{255}$ as described by Madry et al. (2017).

Adversarial training is the industry standard for creating robust models that maintain good clean-label performance (Madry et al., 2017). This method involves replacing clean examples with adversarial examples during the training routine. A simple way to harden models to attack is *adversarial training*, which solves the minimax problem

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\max_{\|\delta\|_p<\epsilon} \mathcal{L}_\theta(\mathbf{x}+\delta, y) \right], \tag{1}$$

where $\mathcal{L}_\theta(\mathbf{x}+\delta, y)$ is the loss function of a network with parameters θ, $\mathbf{x}$ is an input image with label y, and δ is an adversarial perturbation. Adversarial training finds network parameters that keep the loss low (and class labels correct) even when adversarial perturbations are added.

3.1 Naturally trained meta-learners are not robust

Similarly to classically trained classifiers, we expect that few-shot learners are highly vulnerable to attack when adversarial defenses are not employed. We test prominent meta-learning algorithms

Algorithm 2: PGD Attack

Require: network, F_θ, input data, $(\mathbf{x}, y)$, perturbation, δ, number of steps, n, step size, γ, and attack bound, ϵ.
Initialize $\delta \in \mathcal{B}_\epsilon(\mathbf{x})$ randomly;
for $i = 1, ..., n$ **do**
  Compute $g = \text{sign}\left(\nabla_{\mathbf{x}} \mathcal{L}_\theta\left(\mathbf{x} + \delta, y\right)\right)$.
  Update $\delta = \delta + \gamma g$.
  If $\|\delta\|_p > \epsilon$, then project δ onto the surface of $\mathcal{B}_\epsilon(\mathbf{x})$.
return perturbed image $\mathbf{x} + \delta$

against a 20-step PGD attack as in (Madry et al., 2017). Table 2 contains 5-shot natural and robust accuracy on the Mini-ImageNet and CIFAR-FS datasets (Vinyals et al., 2016; Bertinetto et al., 2018).

Model	$\mathcal{A}_{nat}$ MI	$\mathcal{A}_{adv}$ MI	$\mathcal{A}_{nat}$ CIFAR-FS	$\mathcal{A}_{adv}$ CIFAR-FS
ProtoNet	70.23%	0.00%	79.66%	0.00%
R2-D2	73.02%	0.00%	82.81%	0.00%
MetaOptNet	78.12%	0.00%	84.11%	0.00%

Table 2: 5-shot MiniImageNet (MI) and CIFAR-FS results comparing naturally trained meta-learners. $\mathcal{A}_{nat}$ and $\mathcal{A}_{adv}$ are natural and robust test accuracy respectively, where robust accuracy is computed with respect to a 20-step PGD attack.

We find that these algorithms are completely unable to resist the attack. Interestingly, MetaOptNet uses SVM for fine-tuning, which is endowed with a wide margins property. The failure of even SVM to express robustness during testing suggests that using robust fine-tuning methods on naturally trained meta-learners is insufficient for robust performance. To further examine this, we consider MAML, which updates the entire network during fine-tuning. We use a naturally trained MAML model and perform adversarial training during fine-tuning (see Table 3). Adversarial training is performed with 7-PGD as in (Madry et al., 2017). If adversarial fine-tuning yielded robust classification, then we could avoid expensive adversarial training variants during meta-learning.

Model	$\mathcal{A}_{nat}$	$\mathcal{A}_{adv}$	$\mathcal{A}_{nat(adv-tuned)}$	$\mathcal{A}_{adv(adv-tuned)}$
1-shot Mini-ImageNet	45.04%	0.03%	33.18%	0.20%
5-shot Mini-ImageNet	60.25%	0.03%	32.45%	1.55%
1-shot Omniglot	91.50%	68.46%	91.60%	74.66%
5-shot Omniglot	97.12%	82.28%	97.71%	87.94%
5-shot Omniglot AQ	97.27%	**95.85%**	97.51%	**96.14%**

Table 3: MAML models on Mini-ImageNet and Omniglot datasets. $\mathcal{A}_{nat}$ and $\mathcal{A}_{adv}$ are natural and robust test accuracy respectively, where robust accuracy is computed with respect to a 20-step PGD attack. $\mathcal{A}_{nat(adv-tuned)}$ and $\mathcal{A}_{adv(adv-tuned)}$ are natural and robust test accuracy with 7-PGD training during fine-tuning. The bottom row is an adversarially queried model for comparison.

While clean trained MAML models with adversarial fine-tuning are slightly more robust than their naturally fine-tuned counterparts, they achieve almost no robustness on Mini-ImageNet even with adversarial fine-tuning. Omniglot is an easier dataset for robustness, so we include an adversarially queried (AQ) MAML model for comparison. The adversarially queried model achieves far superior robustness. We conclude from these experiments that naturally trained meta-learners are vulnerable to adversarial examples, and an analysis of robust techniques for few-shot learning is in order.

3.2 Transfer learning from adversarially trained models is less robust than robust meta-learning

We have observed that few-shot learning methods with a non-robust feature extractor break under attack. But what if we use a robust feature extractor? In the following section, we consider both transfer learning and meta-learning with a robust feature extractor.

In order to compare transfer learning and meta-learning, we train the backbone networks from meta-learning algorithms on all training data simultaneously in the fashion of standard adversarial training using 7-PGD (not meta-learning). We then fine-tune using the head from a meta-learning algorithm on top of the transferred feature extractor. We compare the performance of these feature extractors to that of those trained using adversarially queried meta-learning algorithms with the same backbones and heads. This experiment provides a direct comparison of feature extractors produced by transfer learning and robust meta-learning (see Table 3.2). Meta-learning exhibits far superior robustness than transfer learning on all algorithms we test.

Model	$\mathcal{A}_{nat}$ Transfer	$\mathcal{A}_{adv}$ Transfer	$\mathcal{A}_{nat}$ Meta	$\mathcal{A}_{adv}$ Meta
MAML	32.79%	18.03%	**33.45%**	**23.07%**
ProtoNet	31.14%	22.31%	**52.04%**	**27.99%**
R2-D2	39.13%	25.33%	**57.87%**	**31.52%**
MetaOptNet	50.23%	22.45%	**60.71%**	**28.08%**

Table 4: Adversarially trained transfer learning and adversarially queried meta-learning on 5-shot Mini-ImageNet. $\mathcal{A}_{nat}$ and $\mathcal{A}_{adv}$ are natural and robust test accuracy respectively, where robust accuracy is computed with respect to a 20-step PGD attack.

4 Adversarial Querying: a robust meta-learning technique

We now adapt adversarial training to the meta-learning paradigm by introducing the query data, but not support data, to adversarial attack (see Algorithm 3). This approach yields fast performance during deployment, as adversarial training (which is roughly 10X slower than standard training) is not required to adapt to a new task. Adversarial querying is algorithm agnostic. We test this method on the MAML, ProtoNet, R2-D2, and MetaOptNet algorithms on the Mini-ImageNet and CIFAR-FS datasets (see Table 4).

Algorithm 3: Adversarial Querying

Require: Base model, F_θ, fine-tuning algorithm, A, learning rate, γ, and distribution over tasks, $p(\mathcal{T})$.
Initialize θ, the weights of F;
while not done **do**
  Sample batch of tasks, $\{\mathcal{T}_i\}_{i=1}^n$, where $\mathcal{T}_i \sim p(\mathcal{T})$ and $\mathcal{T}_i = (\mathcal{T}_i^s, \mathcal{T}_i^q)$.
  for $i = 1, ..., n$ **do**
    Fine-tune model on $\mathcal{T}_i$. New network parameters are written $\theta_i = A(\theta, \mathcal{T}_i^s)$.
    Construct adversarial query data, $\widehat{\mathcal{T}_i^q}$, by maximizing $\mathcal{L}(F_{\theta_i}, \widehat{\mathcal{T}_i^q})$ constrained to $\|\widehat{\mathbf{x}_j^q} - \mathbf{x}_j^q\|_p < \epsilon$ for query examples, $\mathbf{x}_j^q$, and their associated adversaries, $\widehat{\mathbf{x}_j^q}$.
    Compute gradient $g_i = \nabla_\theta \mathcal{L}(F_{\theta_i}, \widehat{\mathcal{T}_i^q})$.
  Update base model parameters: $\theta \leftarrow \theta - \frac{\gamma}{n} \sum_i g_i$

Model	$\mathcal{A}_{nat}$ MI	$\mathcal{A}_{adv}$ MI	$\mathcal{A}_{nat}$ CIFAR-FS	$\mathcal{A}_{adv}$ CIFAR-FS
ProtoNet AQ	52.04%	27.99%	63.53%	40.11%
R2-D2 AQ	57.87%	**31.52%**	69.25%	**44.80%**
MetaOptNet AQ	60.71%	28.08%	71.07%	43.79%

Table 5: Comparison of adversarially queried (AQ) meta-learners on 5-shot Mini-ImageNet (MI) and CIFAR-FS. $\mathcal{A}_{nat}$ and $\mathcal{A}_{adv}$ are natural and robust test accuracy respectively, where robust accuracy is computed with respect to a 20-step PGD attack.

In our tests, R2-D2 outperforms MetaOptNet in robust accuracy despite having a less powerful backbone architecture. In Section 4.2, we dissect the effects of backbone architecture and classification

head on the disparity between R2-D2 and MetaOptNet in robust performance. In Section 4.4, we verify that adversarial querying generates networks robust to a wide array of strong attackers.

Adversarial querying can also be used to construct meta-learning analogues for other variants of adversarial training. We explore this by substituting the cross-entropy loss for the TRADES loss (Zhang et al., 2019). We refer to this method as meta-TRADES. While meta-TRADES can marginally outperform our initial adversarial querying method in robust accuracy with a careful hyperparameter choice, λ, we find that networks trained with meta-TRADES severely sacrifice natural accuracy (see Table 4).

Model	$\mathcal{A}_{nat}$ MI	$\mathcal{A}_{adv}$ MI	$\mathcal{A}_{nat}$ CIFAR-FS	$\mathcal{A}_{adv}$ CIFAR-FS
R2-D2 Adversarial Queried	**57.87%**	31.52%	**69.25%**	44.80%
R2-D2 TRADES ($1/\lambda = 1$)	56.02%	30.96%	66.29%	45.59%
R2-D2 TRADES ($1/\lambda = 3$)	51.51%	**32.30%**	61.41%	**46.54%**
R2-D2 TRADES ($1/\lambda = 6$)	34.29%	22.04%	58.32%	45.89%

Table 6: 5-shot Mini-ImagNet (MI) and CIFAR-FS results comparing meta-TRADES to adversarial querying. $\mathcal{A}_{nat}$ and $\mathcal{A}_{adv}$ are natural and robust test accuracy respectively, where robust accuracy is computed with respect to a 20-step PGD attack.

4.1 For better natural and robust accuracy, only fine-tune the last layer.

High performing meta-learning models, like MetaOptNet and R2-D2, fix their feature extractor and only update their last linear layer during fine-tuning. In the setting of transfer learning, robustness is a feature of early convolutional layers, and re-training these early layers leads to a significant drop in robust test accuracy (Shafahi et al., 2019). We verify that re-training only the last layer leads to improved natural and robust accuracy in adversarially queried meta-learners by training a MAML model but only updating the final layer during fine-tuning including during the inner loop of meta-learning. We find that the model trained by only fine-tuning the last layer decisively outperforms the traditional MAML algorithm (AQ) in both natural and robust accuracy (see Table 4.1).

Layers updated	$\mathcal{A}_{nat}$	$\mathcal{A}_{adv}$	$\mathcal{A}_{nat(adv-tuned)}$	$\mathcal{A}_{adv(adv-tuned)}$
All layers	33.45%	23.07%	33.03%	23.29%
FC Only	40.06%	25.15%	39.94%	25.32%

Table 7: Adversarially queried MAML compared with a MAML variant with only the last layer re-trained during fine-tuning on 5-shot Mini-ImageNet. $\mathcal{A}_{nat}$ and $\mathcal{A}_{adv}$ are natural and robust test accuracy respectively, where robust accuracy is computed with respect to a 20-step PGD attack. $\mathcal{A}_{nat(adv-tuned)}$ and $\mathcal{A}_{adv(adv-tuned)}$ are natural and robust test accuracy respectively with 7-PGD training during fine-tuning. Layers are fine-tuned for 10 steps with a learning rate of 0.01.

4.2 The R2-D2 head, not embedding, is responsible for superior robust performance.

The naturally trained MetaOptNet algorithm outperforms R2-D2 in natural accuracy, but previous research has found that performance discrepancies between meta-learning algorithms might be an artifact of different backbone networks (Chen et al., 2019). On natural meta-learning, we confirm that MetaOptNet with the R2-D2 backbone performs similarly to R2-D2 (see Table 4.2). In our adversarial querying experiments, we saw that MetaOptNet was less robust than R2-D2. This discrepancy remains when we train MetaOptNet with the R2-D2 backbone (see Table 4.2). We conclude that MetaOptNet's backbone is not responsible for its inferior robustness. These experiments suggest that ridge regression may be a more effective fine-tuning technique than SVM for robust performance. ProtoNet with R2-D2 backbone also performs worse than the other two adversarially queried models with the same backbone architecture.

Model	1-shot MI	5-shot MI	1-shot CIFAR	5-shot CIFAR
R2-D2	55.22%	73.02%	68.36%	82.81%
MetaOptNet	**60.65%**	**78.12%**	**70.99%**	**84.11%**
MetaOptNet (R2-D2 backbone)	55.78%	73.15%	68.37%	82.71%

Table 8: Natural test accuracy of naturally trained R2-D2, MetaOptNet, and the MetaOptNet head with R2-D2 backbone on the Mini-ImageNet (MI) and CIFAR-FS (CIFAR) datasets.

Model	1-shot MI	5-shot MI	1-shot CIFAR	5-shot CIFAR
R2-D2	**20.59%**	**31.52%**	**32.33%**	**44.80%**
MetaOptNet	18.37%	28.08%	30.74%	43.79%
MetaOptNet (R2-D2 backbone)	18.81%	24.68%	29.57%	41.90%
ProtoNet (R2-D2 backbone)	18.24%	28.39%	26.48%	40.59%

Table 9: Robust test accuracy of adversarially queried R2-D2, MetaOptNet, and the MetaOptNet and heads with R2-D2 backbone on Mini-ImageNet (MI) CIFAR-FS (CIFAR) datasets. Robust accuracy is computed with respect to a 20-step PGD attack.

4.3 Enhancing robustness with robust architectural features

In addition to adversarial training, architectural features have been used to enhance robustness (Xie et al., 2019). Feature denoising blocks pair classical denoising operations with learned 1×1 convolutions to reduce the feature noise in feature maps at various stages of a network, and thus reduce the success of adversarial attacks. Massive architectures with these blocks have achieved state-of-the-art robustness against targeted adversarial attacks on ImageNet. However, when deployed on small networks for meta-learning, we find that denoising blocks do not improve robustness. We deploy denoising blocks identical to those in Xie et al. (2019) after various layers of the R2-D2 network. The best results for the denoising experiments are achieved by adding a denoising block after the fourth layer in the R2-D2 embedding network (See Table 10).

Model	$\mathcal{A}_{nat}$	$\mathcal{A}_{adv}$
R2-D2	73.02%	0.00%
R2-D2 AQ	57.87%	**31.52%**
R2-D2 AQ Denoising	57.68%	31.14%

Table 10: 5-shot MiniImageNet results for our highest performing R2-D2 with feature denoising blocks. $\mathcal{A}_{nat}$ and $\mathcal{A}_{adv}$ are natural and robust test accuracy respectively, where robust accuracy is computed with respect to a 20-step PGD attack.

4.4 Resistance to other attacks

We test our method by exposing our adversarially queried R2-D2 model to a variety of powerful adversarial attacks. We implement the momentum iterated fast gradient sign method (MI-FGSM), DeepFool, and 20-step PGD with 20 random restarts (Dong et al., 2018; Moosavi-Dezfooli et al., 2016; Madry et al., 2017). Our adversarially queried model indeed is nearly as robust against the strongest ℓ_∞ bounded attacker as it is against the 20-step PGD attack with a single random start we tested against previously. Note that DeepFool is not ℓ_∞ bounded and thus the perturbed images are outside of the robustness radius enforced during adversarial querying.

5 Pre-processing defenses as an alternative to adversarial training

Recent works have proposed pre-proccessing defenses for sanitizing adversarial examples before feeding them into a naturally trained classifier. If successful, these methods would avoid the expen-

Model	$\mathcal{A}_{nat}$	$\mathcal{A}_{DF}$	$\mathcal{A}_{MI}$	$\mathcal{A}_{20-PGD}$
R2-D2	73.02%	7.91%	0.01%	0.0%
R2-D2 AQ	57.87%	**14.45%**	**31.87%**	**30.31%**
R2-D2 AT (Transfer Learning)	39.13%	0.42%	24.01%	19.75%

Table 11: 5-shot MiniImageNet results against DeepFool (DF) (2 iteration) ℓ_∞ attack, MI-FGSM (MI) ($\epsilon = 8/255$) attack, and PGD attack with 20 random restarts (20-PGD). We compare R2-D2 trained with adversarial-querying (AQ) to the transfer learning R2-D2 as in section 3.2.

sive adversarial querying procedure during training. While this approach has found success in the mainstream literature, we find that it is ineffective in the few-shot regime.

In DefenseGAN, a GAN trained on natural images is used to sanitize an adversarial example by replacing (possible corrupted) test images with the nearest image in the output range of the GAN (Samangouei et al., 2018). Unfortunately, GANs are not expressive enough to preserve the integrity of testing images on complex datasets involving high-res natural images, and recent attacks have critically compromised the performance of this defense (Ilyas et al., 2017; Athalye et al., 2018). We found the expressiveness of the generator architecture used in the original DefenseGAN setup to be insufficient for even CIFAR-FS, so we substitute a stronger ProGAN generator to model the CIFAR-100 classes (Karras et al., 2017).

The supperesolution defense first denoises data with sparse wavelet filters and then performs superresolution (Mustafa et al., 2019). This defense is also motivated by the principle of projecting adversarial examples onto the natural image manifold. We test the superresolution defense using the same wavelet filtering and superresolution network (SRResNet) used by Mustafa et al. (2019) and first introduced by Ledig et al. (2017). Like with the generator for DefenseGAN, we train the SRResNet on the entire CIFAR-100 dataset before applying the superresolution defense.

We find that these methods are not well suited to the few-shot domain, in which the generative model or superresolution network may not be able to train on the little data available. Morever, even after training the generator on all CIFAR-100 classes, we find that DefenseGAN with a naturally trained R2-D2 meta-learner performs significantly worse in both natural and robust accuracy than an adversarially queried meta-learner of the same architecture. Similarly, the superresolution defense achieves little robustness. The results of these experiments can be found in Table 5.

Model	$\mathcal{A}_{nat}$	$\mathcal{A}_{adv}$
R2-D2	83.30%	0.00%
R2-D2 AQ	69.25%	**44.80%**
R2-D2 with SR defense	35.15%	23.00%
R2-D2 with DefenseGAN	35.15%	28.05%

Table 12: 5-shot CIFAR-FS results comparing the superresolution defense (SR defense) and DefenseGAN. $\mathcal{A}_{nat}$ and $\mathcal{A}_{adv}$ are natural and robust test accuracy respectively, where robust accuracy is computed with respect to a 20-step PGD attack. Both methods perform worse than their adversarially queried counterpart.

6 Discussion & Conclusion

Naturally trained networks for few-shot learning are vulnerable to adversarial attacks, and existing robust transfer learning methods do not perform well on few-shot tasks. Naturally trained networks suffer from adversarial vulnerability even when adversarially trained during fine-tuning. We thus identify the need for an investigation into robust few-shot methods. We particularly study robustness in the context of meta-learning. We develop an algorithm-agnostic method, called adversarial querying, for hardening meta-learning models. We find that meta-learning models are most robust when the feature extractor is fixed, and only the last layer is retrained during the fine tuning stage. We further identify that choice of classification head matters for robustness. We hope that this paper serves as a starting point for developing new adversarially robust methods for few-shot applications.

REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 17–36, 2012.

Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International conference on learning representations*, 2019.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. doi: 10.1109/cvpr.2018.00957. URL `http://dx.doi.org/10.1109/cvpr.2018.00957`.

Yan Duan, Marcin Andrychowicz, Bradly Stadie, OpenAI Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel, and Wojciech Zaremba. One-shot imitation learning. In *Advances in neural information processing systems*, pp. 1087–1098, 2017.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning models. *arXiv preprint arXiv:1707.08945*, 2017.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1126–1135. JMLR. org, 2017.

Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4367–4375, 2018.

Andrew Ilyas, Ajil Jalal, Eirini Asteri, Constantinos Daskalakis, and Alexandros G Dimakis. The robust manifold defense: Adversarial training using generative models. *arXiv preprint arXiv:1712.09196*, 2017.

Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events. *arXiv preprint arXiv:1703.03129*, 2017.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2017.

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

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and et al. Photo-realistic single image super-resolution using a generative adversarial network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. doi: 10.1109/cvpr.2017.19. URL `http://dx.doi.org/10.1109/CVPR.2017.19`.

Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10657–10665, 2019.

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.

Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *European Conference on Computer Vision*, pp. 488–501. Springer, 2012.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Aamir Mustafa, Salman H. Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks, 2019.

Tomas Pfister, James Charles, and Andrew Zisserman. Domain-adaptive discriminative one-shot learning of gestures. In *European Conference on Computer Vision*, pp. 814–829. Springer, 2014.

Parsa Saadatpanah, Ali Shafahi, and Tom Goldstein. Adversarial attacks on copyright detection systems. *arXiv preprint arXiv:1906.07153*, 2019.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*, 2018.

Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David Jacobs, and Tom Goldstein. Adversarially robust transfer learning. *arXiv preprint arXiv:1905.08232*, 2019.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

Manasi Vartak, Arvind Thiagarajan, Conrado Miranda, Jeshua Bratman, and Hugo Larochelle. A meta-learning perspective on cold-start recommendations for items. In *Advances in neural information processing systems*, pp. 6904–6914, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 501–509, 2019.

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

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.

Yue Zhao, Hong Zhu, Qintao Shen, Ruigang Liang, Kai Chen, and Shengzhi Zhang. Practical adversarial attack against object detector. *arXiv preprint arXiv:1812.10217*, 2018.

A Appendix

A.1 Training hyperparameters

We train ProtoNet, R2-D2, and MetaOptNet models for 60 epochs with SGD. We use a learning rate of 0.1, momentum (Nesterov) of 0.9, and a weight decay term of $5(10^{-4})$ for the parameters of both the head and the embedding. We decrease the learning rate to 0.06 after epoch 20, 0.012 after epoch 40, and 0.0024 after epoch 50. MAML is trained for 60000 epochs with meta learning rate of 0.001 and fine-tuning learning rate of 0.01. Fine-tuning is performed for 10 steps per task.