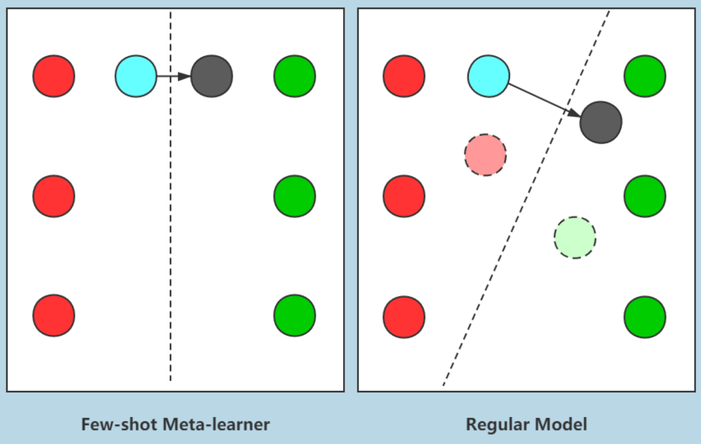
Task-based Focal Loss for Adversarially Robust Meta-Learning

Yufan Hou, Lixin Zou, Weidong Liu
{hou-yf18, zoulx15}@mails.tsinghua.edu.cn, liuwd@tsinghua.edu.cn

Problem

Due to the lack of training samples for each task, few-shot meta-learners are more vulnerable to adversarial attacks than regular models. In this work we study on a typical meta-learner, **Model-Agnostic Meta-Learning(MAML)**, to explore adversarial robustness improvement of meta-learners.



Basic Concepts

Adversarial attack is a technique that attempts to fool models by supplying deceptive input. The objective of mainstream white-box attack $\mathcal{A}$ is to maximize loss on perturbed example with perturbation restriction, which can be expressed formally as:

$$\mathcal{A}(x) \rightarrow \max_{x_a : ||x_a - x|| \leq \epsilon} \mathcal{L}_{CE}(x_a) \tag{1}$$

Correspondingly **adversarial robustness** is to evaluate the ability of defending against an adversary who will attack the model.

Related Works

Adversarial attacks:
FGSM The attack generates adversarial examples x_a via a one-step gradient using the loss function $\mathcal{L}$ of victim model:

$$x_a = x + \epsilon \text{sign}(\nabla_x \mathcal{L}(x, y)). \tag{2}$$

The multi-step version of FGSM is PGD attack.
C&W In Carlini & Wagner's method the attack is viewed as an optimization problem which solves

$$\min_{x_a} \|x - x_a\|_p - c\mathcal{L}(x_a, y). \tag{3}$$

Adversarial robustness of meta-learners:
ADML The method fine-tunes tasks to generate extra adapted parameters, and utilizes antagonistic correlations to make the inner gradient update and the meta-update arm-wrestle with each other.
Adversarial Querying(AQ) It was found that attacking both support and query data like ADML is not necessary and relatively time-consuming. In AQ, only query data is perturbed to harden several meta-learning models.

Key References

[1] C. Finn, P. Abbeel, and S. Levine: *Model-agnostic meta-learning for fast adaptation of deep networks.*
[2] I. J. Goodfellow, J. Shlens, and C. Szegedy: *Explaining and harnessing adversarial examples.*
[3] C. Yin, J. Tang, Z. Xu, and Y. Wang: *Adversarial meta-learning.*
[4] M. Goldblum, L. Fowl, and T. Goldstein: *Robust few-shot learning with adversarially queried meta-learners.*

Focal Loss for Adversarially Trained Meta-learner

Since we can use cross entropy loss $\mathcal{L}_{CE}$ to represent focal loss as

$$\mathcal{L}_{FL} = (1 - p_t)^\gamma (-log(p_t)) = (1 - exp(-\mathcal{L}_{CE}))^\gamma \cdot \mathcal{L}_{CE} \tag{4}$$

and regard the term $M_{FL} = (1 - exp(-\mathcal{L}_{CE}))^\gamma$ as modulating factor. When applied to meta-learner, the loss terms in formula (4) represent sum of loss on query examples in the task rather than a single example. Considering that the objective of mainstream white-box attack $\mathcal{A}$ is formula (1), we introduce a fixed adversary to describe adversarial robustness of a task τ:

$$\mathcal{L}_{AR}(\tau) = \max\{\mathcal{L}_{CE}(f_{\theta_\tau}, \mathcal{A}(x_q)) - \mathcal{L}_{CE}(f_{\theta_\tau}, x_q), \delta\} \tag{5}$$

where $\delta > 0$ is a small number which ensures positivity of $\mathcal{L}_{AR}$. We utilize such loss with multiple k to replace $\mathcal{L}_{CE}$ in the modulating factor M_{FL} of focal loss (4), and give definition of our modulating factor based and task-based adversarial focal loss $\mathcal{L}_{TAFL}$ on each task τ:

$$M_{TAFL}(\tau) = (1 - exp(-k\mathcal{L}_{AR}(\tau)))^\gamma, \mathcal{L}_{TAFL}(f_{\theta_\tau}, x_q) = M_{TAFL}(\tau) \cdot \mathcal{L}_{CE}(f_{\theta_\tau}, \mathcal{A}(x_q)) \tag{6}$$

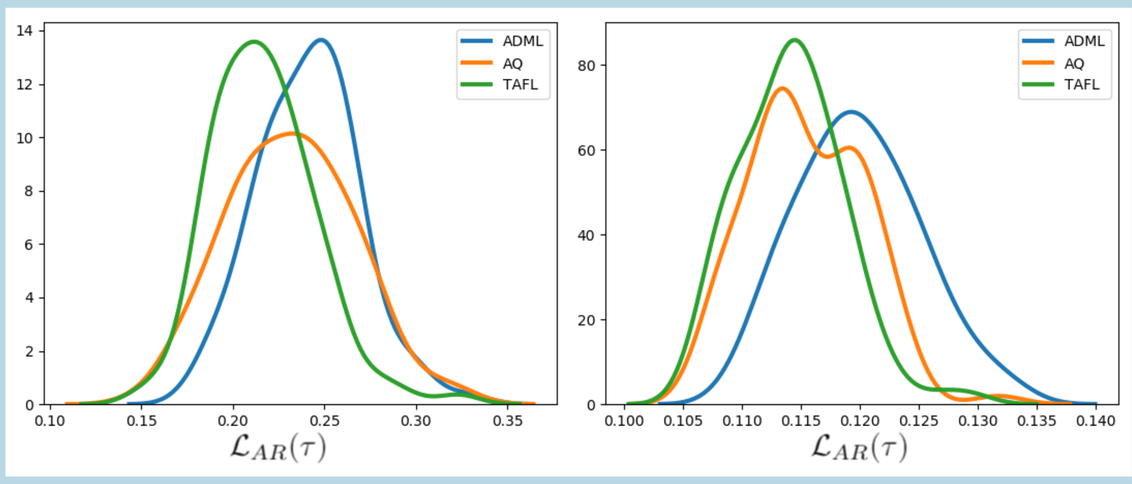
For a batch of tasks, the modulating factors $M_{TAFL}(\tau)$ are linearly normalized within the batch. Such factors are not function of θ to be minimized during gradient descent optimization in MAML.

Experimental Design and Results

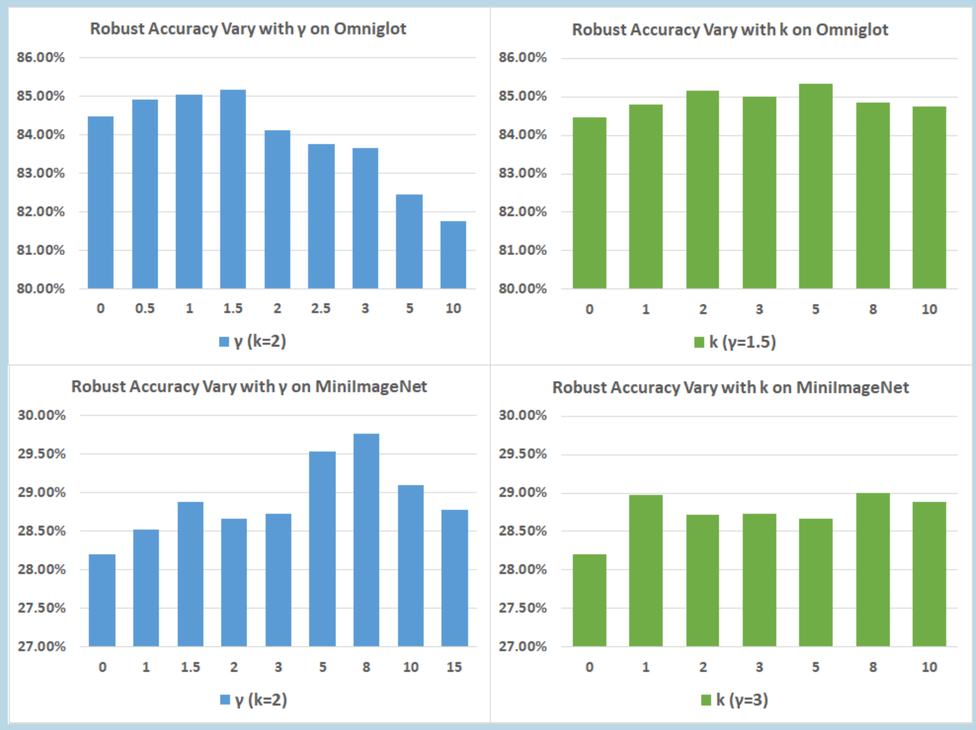
As shown in Table 1, images with small perturbations generated by any attack can fool the model to produce nearly no correct classification. It is demonstrated that our method achieves better results than baseline defenses, ADML and AQ. Our method can improve robust accuracy by at least about 0.5% compared with baselines. The promotion can reach up to 1% on part of test cases.

Table 1: Robust accuracy against three typical attacks on **MiniImageNet** dataset.

Model/Attack	**MiniImageNet** dataset (5-way 1-shot)		
	PGD	**MI-FGSM**	**C&W**
MAML	$0.42 \pm 0.06\%$	$0.01 \pm 0.01\%$	$14.38 \pm 0.36\%$
ADML	$28.53 \pm 0.48\%$	$28.19 \pm 0.56\%$	$26.77 \pm 0.41\%$
AQ	$28.20 \pm 0.48\%$	$27.94 \pm 0.54\%$	$26.82 \pm 0.42\%$
TAFL(ours)	$\mathbf{29.53 \pm 0.60\%}$	$\mathbf{28.94 \pm 0.61\%}$	$\mathbf{27.75 \pm 0.44\%}$



We plot the distribution of $\mathcal{L}_{AR}(\tau)$ over tasks when testing different defense methods in figure above. Here 100 batches of tasks are sampled to estimate the distribution via kernel density estimation(KDE) method. Compared with other two baselines AQ and ADML, our proposed TAFL can apparently reduce the proportion of tasks with high $\mathcal{L}_{AR}(\tau)$ when attacked by same PGD attack. This actually benefits from the feature of focusing more on tasks with high $\mathcal{L}_{AR}(\tau)$ in our method.



As for parameter sensitivity, we evaluate the performance of TAFL under different parameters γ and k. The model achieves higher performance by increasing k or γ at the beginning, which means the focusing effect appears. However, the performance declines when γ continues to rise. It can be interpreted that the model with high γ overly ignores and performs poorly on tasks with low $\mathcal{L}_{AR}$. The impact of parameter k is much smaller and almost irregular unlike γ.