清華大學
Tsinghua University

Task-based Focal Loss for Adversarially Robust Meta-Learning

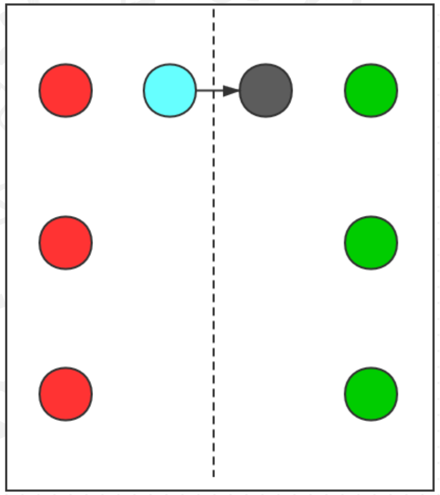
Yufan Hou, Lixin Zou, Weidong Liu
Department of Computer Science and Technology
Tsinghua University, Beijing, China

ICPR 2020

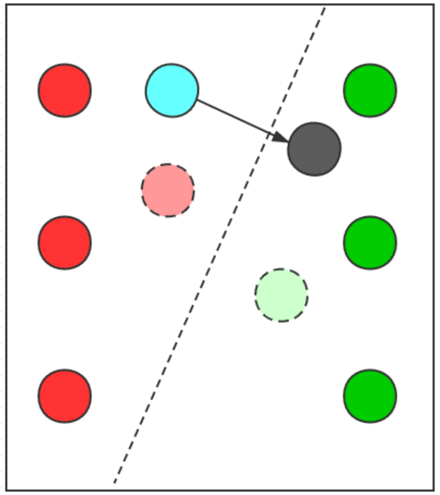
Content

Background

- **Adversarial attack:**
  - a technique that attempts to fool models by supplying deceptive input
  - white-box attack: maximize loss on perturbed example with perturbation restriction
- **Adversarial robustness:**
  - evaluate the ability of defending against an adversary who will attack the model
- **Problem of robust (few-shot) meta-learner:**
  - meta-learners designed to learn with less training data, are easier to attack
  - Edmunds et al., revealed that simple attack can disturb MAML with success rate over 80%
- **Our focus:**
  - select MAML as a typical meta-learner
  - improve adversarial robustness of MAML



Few-shot Meta-learner



Regular Model

Related Work

Algorithm 1 Model-Agnostic Meta-Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters

```
randomly initialize θ
while not done do
    Sample batch of tasks T_i ~ p(T)
    for all T_i do
        Evaluate ∇_θ L_{T_i}(f_θ) with respect to K examples
        Compute adapted parameters with gradient descent: θ'_i = θ − α∇_θ L_{T_i}(f_θ)
    end for
    Update θ ← θ − β∇_θ Σ_{T_i~p(T)} L_{T_i}(f_{θ'_i})
end while
```

Meta-learning:

- Model-Agnostic Meta-Learning(MAML)
- Bayesian Model-Agnostic Meta-Learning
- Hierarchically Structured Meta-learning(HSML)
- many models are derived from MAML

Adversarial attacks:

- FGSM: Fast Gradient Sign Method $x_a = x + \epsilon \mathrm{sign}(\nabla_x \mathcal{L}(x, y)).$
- PGD: improved version, iteratively generates perturbation & conduct projection
- C&W attack: optimization-based attack $\min_{x_a} \|x - x_a\|_p - c\mathcal{L}(x_a, y).$

Robust meta-learner:

- ADML: perturb both support and query data, make the inner gradient update and the meta-update arm-wrestle with each other
- Adversarial Querying: only perturb query data, more efficient and more robust

Method

❖ Motivation: Focal Loss

- ❖ vast number of easy negatives overwhelms the object detector during training
- ❖ proposed to make the model focus on hard examples

$$\mathcal{L}_{FL}(p_t) = -(1-p_t)^{\gamma} log(p_t)$$

❖ TAFL: Task-based Adversarial Focal Loss

❖ I. Sample->Task:

- ❖ use cross entropy loss to represent focal loss

$$\mathcal{L}_{CE} = -log(p_t)$$

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

- ❖ extract the modulating factor term

$$M_{FL} = (1 - exp(-\mathcal{L}_{CE}))^{\gamma}$$

- ❖ applied to meta-learner? loss term represents sum of loss in a task rather than an example

Method

❖ II. Classification difficulty->Adversarial robustness:

- ❖ **objective of white-box attacks $\mathcal{A}$:**

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

- ❖ **introduce difference between loss on clean and perturbed query data $\mathcal{L}_{AR}(\tau)$ to replace $\mathcal{L}_{CE}$**

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

- ❖ **rewrite modulating factor term and construct meta update loss**

$$M_{TAFL}(\tau) = (1 - exp(-k\mathcal{L}_{AR}(\tau)))^\gamma$$

$$\mathcal{L}_{TAFL}(f_{\theta_\tau}, x_q) = M_{TAFL}(\tau) \cdot \mathcal{L}_{CE}(f_{\theta_\tau}, \mathcal{A}(x_q))$$

- ❖ **such factors are not function of θ to be minimized during gradient descent optimization**

Experiment Design & Results

Tsinghua University 清華大學

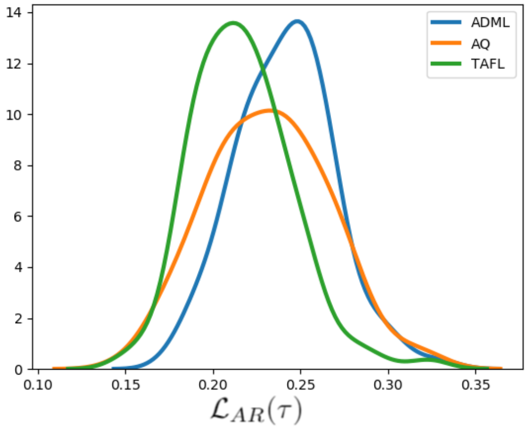
- **Experimental setup :**
  - datasets: Omniglot / MiniImageNet / CUB
  - sample 100 batches of test tasks, calculated with 95% confidence intervals
- **Robust accuracy:**
  - baselines: MAML, ADML, Adversarial Querying
  - 3 attacks for test: PGD, MI-FGSM, C&W

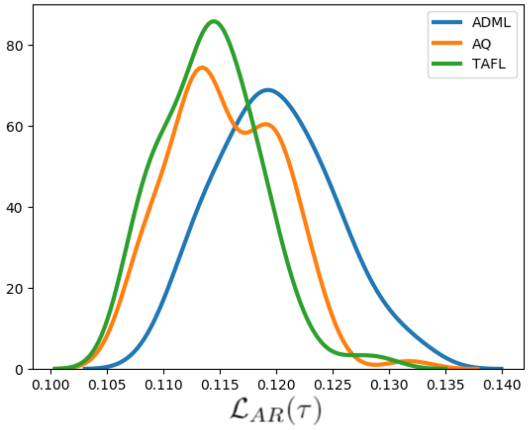
Attack	Omniglot (5-way 1-shot)	MiniImageNet (5-way 1-shot)
PGD	$\epsilon = 0.1, step = 30$	$\epsilon = 0.01, step = 30$
MI-FGSM	$\epsilon = 0.1, step = 30$	$\epsilon = 0.01, step = 30$
C&W	$c = 10.0, step = 60$	$c = 1.0, step = 30$

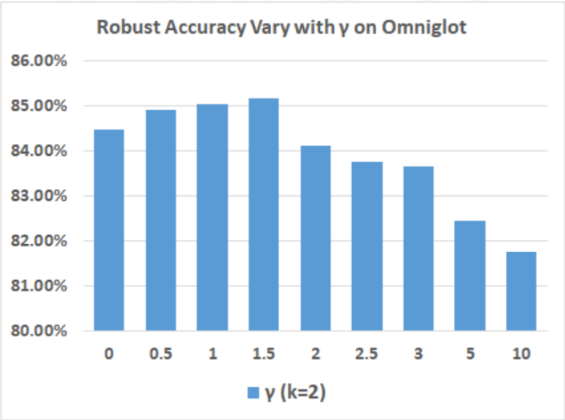
Model/Attack	MiniImageNet dataset (5-way 1-shot) PGD	MI-FGSM	C&W
MAML [7]	$0.42 \pm 0.06\%$	$0.01 \pm 0.01\%$	$14.38 \pm 0.36\%$
ADML [11]	$28.53 \pm 0.48\%$	$28.19 \pm 0.56\%$	$26.77 \pm 0.41\%$
AQ [12]	$28.20 \pm 0.48\%$	$27.94 \pm 0.54\%$	$26.82 \pm 0.42\%$
TAFL(ours)	$\mathbf{29.53 \pm 0.60\%}$	$\mathbf{28.94 \pm 0.61\%}$	$\mathbf{27.75 \pm 0.44\%}$

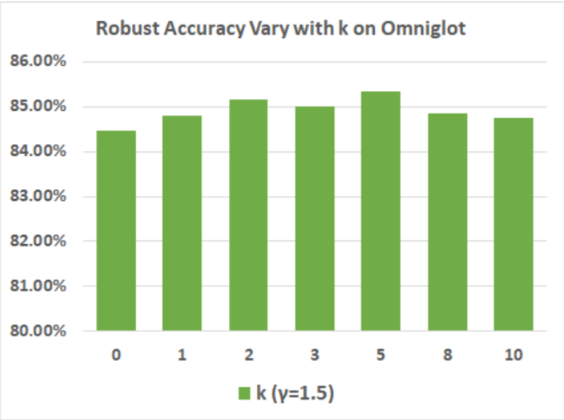
Experiment Design & Results

- **Visualization on adversarial robustness loss($\mathcal{L}_{AR}$):**
  - distribution of $\mathcal{L}_{AR}$ over tasks when testing different defense methods
  - estimate the distribution via kernel density estimation(KDE) method
  - our method reduce the proportion of tasks with high $\mathcal{L}_{AR}$
- **Effects of different parameters:**
  - γ is a more sensitive parameter
  - robust accuracy increases first, and then reduces with γ increases









Conclusion

- proposed TAFL focus more on tasks which are hard to protect
- the proportion of adversarially fragile tasks are reduced via focal effect
- result in promotion of overall adversarial robustness

清華大學
Tsinghua University

THANK YOU

ICPR 2020