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 & Related Work

Method

Experiments

Conclusion
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
Related Work

- **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
    \[ x_a = x + \epsilon \text{sign}(\nabla_x \mathcal{L}(x, y)). \]
  - **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

---

Algorithm 1 Model-Agnostic Meta-Learning

**Require:** \( p(\mathcal{T}) \): distribution over tasks

**Require:** \( \alpha, \beta \): step size hyperparameters

1. randomly initialize \( \theta \)
2. **while** not done **do**
3. Sample batch of tasks \( \mathcal{T}_i \sim p(\mathcal{T}) \)
4. **for** all \( \mathcal{T}_i \) **do**
5. Evaluate \( \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_\theta) \) with respect to \( K \) examples
6. Compute adapted parameters with gradient descent: \( \theta'_i = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_\theta) \)
7. **end for**
8. Update \( \theta \leftarrow \theta - \beta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'}_i) \)
9. **end while**
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 $\mathcal{L}_{CE}$

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

- rewrite modulating factor term and construct meta update loss

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

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

- such factors are not function of $\theta$ to be minimized during gradient descent optimization
**Experiment Design & Results**

- **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

<table>
<thead>
<tr>
<th>Attack</th>
<th>Omniglot (5-way 1-shot)</th>
<th>MiniImageNet (5-way 1-shot)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGD</td>
<td>(\epsilon = 0.1, \text{step} = 30)</td>
<td>(\epsilon = 0.01, \text{step} = 30)</td>
</tr>
<tr>
<td>MI-FGSM</td>
<td>(\epsilon = 0.1, \text{step} = 30)</td>
<td>(\epsilon = 0.01, \text{step} = 30)</td>
</tr>
<tr>
<td>C&amp;W</td>
<td>(c = 10.0, \text{step} = 60)</td>
<td>(c = 1.0, \text{step} = 30)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model/Attack</th>
<th>MiniImageNet dataset (5-way 1-shot)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PGD</td>
</tr>
<tr>
<td>MAML [7]</td>
<td>0.42 ± 0.06%</td>
</tr>
<tr>
<td>ADML [11]</td>
<td>28.53 ± 0.48%</td>
</tr>
<tr>
<td>AQ [12]</td>
<td>28.20 ± 0.48%</td>
</tr>
<tr>
<td>TAFL (ours)</td>
<td>29.53 ± 0.60%</td>
</tr>
</tbody>
</table>
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:**
  - $\gamma$ is a more sensitive parameter
  - robust accuracy increases first, and then reduces with $\gamma$ 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
THANK YOU