Distribution-restrained Softmax Loss for the Model Robustness

Chen Li
Inspur Electronic Information Industry

AI Safety-SafeRL 2023 (IJCAI-23)
Introduction

- **AI models are vulnerable**
  - Adversarial Noise
    - “panda” + [adversarial noise] = “gibbon”
  - Adversarial Rotation
    - “vulture” + [rotation] = “orangutan”
  - Adversarial Photographer
    - “not hotdog” + [photograph] = “hotdog”

- **Robustness plays an important role in AI safety**
  - Stop sign
    - Adversarial perturbation
    - Confidence: 0.9153
  - Flowerpot
    - Confidence: 0.8374

Introduction

- Supervised Machine Learning requires labeled training data
- Usually, training data can be error-prone, adding ‘label noise’ to training sets
- DNN models should be trained with noisy labels operate effectively

\[ x^* = \arg \max_x (a_i(x) - R_\theta(x)) \]

\[ \hat{X}_{adv}^n = X_{adv}^{n-1} - \alpha \cdot \text{sign}(\nabla_{X^n} J(X^n | y_{target})) \]

\[ X_{adv}^n = \text{clip} \left( \hat{X}_{adv}^n, [X - \epsilon, X + \epsilon] \right) \]
Loss to Robustness


- CE are implicitly weighed more than samples with predictions that agree more with provided labels in the gradient update
- MAE treats every sample equally, which makes it more robust to noisy labels
- MAE can concurrently cause increased difficulty in training, and lead to performance drop
- GCE is a trade-off loss function between performance and robustness
AI Robustness v. s. Human Robustness
**Relationship Between Adversarial Examples and 2\textsuperscript{nd} Softmax**

*Green bar:* the probabilities of softmax outputs after attack  
*Red bar:* the probabilities of second largest softmax outputs before attack

- After attacks, the second largest softmax probabilities trends to be the largest one
- One hypothesis: as the second largest softmax probabilities decrease, it becomes more difficult for an adversarial attack to manipulate them into becoming the largest one
Distribution-Restrained Softmax Loss Function (DRSL)

\[
L(f(x; \theta), y) = -1_y^T \log \left( \text{softmax}(f(x; \theta)) \right)
\]

\[
L(f(x; \theta), y) = -1_y^T \log \left( \text{softmax}(f(x; \theta)) \right) + \tau \cdot d(\text{softmax}(f(x; \theta)), \text{avg})
\]

- Model accuracy has nothing to do with the softmax distribution
- DRSL restrain the softmax distribution
Robustness of CE/DRSL based model

- DRSL is more robust than other methods in same precision level
- DRSL can be extended to more models
Visualization of Distribution-Restrained Softmax

- All the loss functions achieve a similar precision level
- DRSL is difficult to attack
- DRSL is robust in label noise
- After attacks, adversarial examples of other two models are diffusion in the reduced space, while DRSL’s are concentrated at the tip of clusters
We identified a significant factor that affects the robustness of models: the distribution characteristics of softmax values for non-real label samples.

The results after an attack are highly correlated with the distribution characteristics.

After the distribution diversity of softmax is suppressed in loss function, a significant improvement of model robustness were found.

DRSL can be applied not only to classification models but also to other softmax-inclusive models, such as generative models, which inspires us to further investigate and explore of the method.

\[
L(f(x; \theta), y) = -1^T_y \log(\text{softmax}(f(x; \theta))) + \tau \cdot d(\text{softmax}(f(x; \theta)), \text{avg})
\]
Thank You

Thanks to all authors of the present work