Robust AI Project Team

Weekly Report

Heran Zhu

Electronic Information School, Wuhan University

Sep 10 th, 2021

基于优化的攻击方法

白盒攻击

- **L-BFGS**: 有目标攻击, Intriguing properties of neural networks
- **DeepFool**: 无目标攻击, a simple and accurate method to fool deep neural networks
- **UAP**: 无目标攻击, Universal adversarial perturbations
- **CW**: 有/无目标攻击, Towards Evaluating the Robustness of Neural Networks

黑盒攻击

- **Grad.Est.**: 有/无目标攻击, Exploring the space of black-box attacks on deep neural networks
- **ZOO**: 有/无目标攻击, ZOO:Zeroth pder optimization based black-box attacks to deep neural networks without training substitute models.
- **IS**: 有/无目标攻击, Simple black-box adversarial perturbations on deep neural networks

基于优化的攻击方法

需要解决的含约束条件的优化问题

$$\begin{aligned} &\textbf{Minimize } ||r||_2 \\ &\textbf{subject to } f(x+r) = l \textbf{ or } f(x+r) \neq f(x) \\ &x + r \in [0,1]^m \end{aligned} \tag{1}$$

求解满足约束条件的最小对抗扰动 r，就可以产生对抗样本

基于优化的攻击策略

4种基于优化的白盒攻击 L-BFGS, CW, DeepFool, UAP

- **1. L-BFGS**: 有目标攻击

$$\textbf{Minimize } c \cdot ||r||_2 + loss_f(x+r, l) \textbf{ subject to } x + r \in [0,1]^m \tag{2}$$

将对抗样本$x+r$经过分类器的预测输出定向为目标标签l

- **2. CW**: 有目标/无目标攻击

$$\textbf{Minimize } ||r||_2 + c \cdot f(x+r) \textbf{ subject to } x + r \in [0,1]^m \tag{3}$$

提出了7种目标函数 f 来进行优化

基于优化的攻击策略

- **3. DeepFool**: 无目标攻击

$$\begin{aligned} &\textbf{Minimize } ||r||_2 \\ &\textbf{subject to } sign(f(x_0 = r)) \neq sign(f(x_0)) \\ &\qquad or \exists k : \omega_k^T(x_0 + r) + b_k \geq \omega_{\hat{k}(x_0)}^T(x_0 + r) + b_{\hat{k}(x_0)} \end{aligned} \tag{4}$$

- **3. UAP**: 无目标攻击

$$\begin{aligned} &\Delta v_i \leftarrow \arg\min_r \text{ s.t. } \hat{k}(x_i + v + r) \neq \hat{k}(x_i) \\ &v \leftarrow \mathcal{P}_{p,\zeta} = \arg\min_{v'} ||v - v'||_2 \text{ subject to } ||v'||_p \leq \zeta \end{aligned} \tag{5}$$

Thanks!