Improved Network Robustness with Adversary Critic

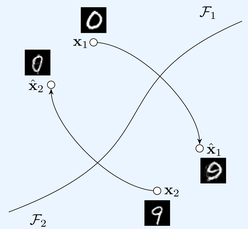
Alexander Matyasko, Lap-Pui Chau
Nanyang Technological University, Singapore

NANYANG TECHNOLOGICAL UNIVERSITY
School of Electrical & Electronic Engineering

Neural Information Processing Systems Foundation

Abstract

- Ideally, what confuses neural network should be confusing to humans.
- Experiments with adversarial examples show that imperceptible noise can change the prediction.
- To address this gap in perception, we propose a novel approach for learning robust classifier.

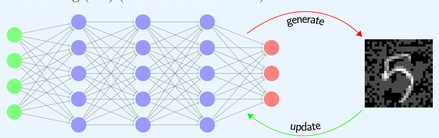


Main idea

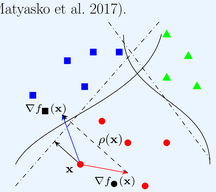
Adversarial examples for the robust classifier should be indistinguishable from the regular data of the adversarial target.

Previous Work

- Adversarial Training (AT) (Goodfellow et al. 2015).



- Margin Maximization (Matyasko et al. 2017).



Robust Optimization

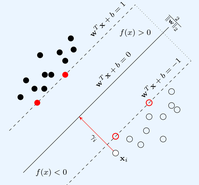
Robust optimization seeks a solution robust to the worst-case input perturbations:

$$\min_{\mathbf{W}} \max_{\boldsymbol{r}_i \in \mathcal{U}_i} \sum_{i=1}^{N} \mathcal{L}(f(\boldsymbol{x}_i + \boldsymbol{r}_i), y_i)$$

where $\mathcal{L}$ is a training loss, $\boldsymbol{r}_i$ is an arbitrary (even adversarial) perturbation for the input $\boldsymbol{x}_i$, and $\mathcal{U}_i$ is an uncertainty set, e.g. l_p-norm ϵ-ball $\mathcal{U}_i = \{\boldsymbol{r}_i : \|\boldsymbol{r}_i\|_p \leq \epsilon\}$.
Selecting a good uncertainty set $\mathcal{U}$ for robust optimization is crucial. Poorly chosen uncertainty set may result in an overly conservative model. Most importantly, each perturbation $\boldsymbol{r} \in \mathcal{U}$ should leave the "true" class of the original input $\boldsymbol{x}$ unchanged. To ensure that the changes of the network prediction are "mistakes", (Goodfellow et al. 2015) argue in favor of a max-norm perturbation constraint for image classification problems. However, simple disturbance models (e.g. l_p-norm ϵ-ball used in adversarial training) are inadequate because the distance to the decision boundary for different inputs may significantly vary.

Main limitation

Uncertainty set $\mathcal{U}$ needs to be carefully selected, so that each perturbation $\boldsymbol{r}$ is label non-changing.



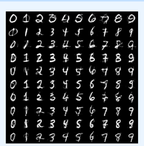
SVM is equivalent to the RO with hinge loss.

Our Approach

Classifier is robust if its adversarial examples are indistinguishable from the regular data of the adversarial target. So, the changes introduced by the adversarial noise should be associated with removing identifying characteristics of the original label and adding identifying characteristics of the adversarial label. Then, we propose the following mathematical problem:

$$\min \sum_{i=1}^{N} \mathcal{L}(f(\boldsymbol{x}_i), y_i) + \lambda \mathcal{D}\left[p_{\text{data}}(\boldsymbol{x}, y), p_{\text{adv}}(\boldsymbol{x}, y)\right]$$

where $p_{\text{data}}(\boldsymbol{x}, y)$ and $p_{\text{adv}}(\boldsymbol{x}, y)$ is the distribution of the natural and the adversarial examples.

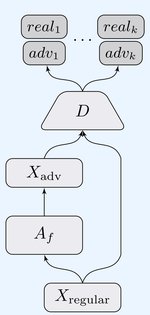


Adversary Critic

- The objective for the adversary critic is to distinguish between natural and adversarial examples.
- We implement a multiclass adversary critic as a k-output neural network with the objective for the k-output to distinguish between natural and adversarial examples for the target k:

$$\mathcal{L}(f^*, D_k) = \min_{D_k} \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x}\,|\,y_k)} \left[\log D_k(\boldsymbol{x})\right] + \mathbb{E}_{y, y \neq y_k} \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x}\,|\,y)} \left[\log\left(1 - D_k\left(\mathcal{A}_f(\boldsymbol{x}; y_k)\right)\right)\right]$$

where $\mathcal{A}_f(\boldsymbol{x}, y_k)$ is the targeted adversarial attack on the classifier f which transforms the input $\boldsymbol{x}$ to the adversarial target y_k. Note that the second term in the above equation is computed by transforming the regular inputs $(\boldsymbol{x}, y) \sim p_{\text{data}}(\boldsymbol{x}, y)$ with the original label y different from the adversarial target y_k.
- To improve stability of the critic, we also add the gradient penalty to its objective (Gulrajani et al. 2017).



Classifier and Adversarial Attack

- The objective for the classifier f is to correctly classify regular examples subject to that its adversarial examples confuse the critic D:

$$\mathcal{L}(f, D^*) = \min_{f} \mathbb{E}_{\boldsymbol{x}, y \sim p_{\text{data}}(\boldsymbol{x}, y)} \mathcal{L}(f(\boldsymbol{x}), y) + \lambda \sum_{y_k} \mathbb{E}_{y, y \neq y_k} \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x}\,|\,y)} \left[\log D_k^*\left(\mathcal{A}_f(\boldsymbol{x}; y_k)\right)\right]$$

where $\mathcal{A}_f(\boldsymbol{x}, y_k)$ is the targeted adversarial attack on the classifier f which transforms the input $\boldsymbol{x}$ to the adversarial target y_k.
- To improve stability of the adversarial mapping during training, we introduce adversarial cycle-consistency constraint:

$$\mathcal{L}_{\text{cycle}}(y_s, y_t) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x}\,|\,y_s)} \left[\left\|\mathcal{A}_f(\mathcal{A}_f(\boldsymbol{x}, y_t), y_s) - \boldsymbol{x}\right\|_2\right] \quad \forall y_s \neq y_t$$

where y_s is the original label of the input and y_t is the adversarial target.
- To attack the model, we use an iterative attack with an update rule:

$$\boldsymbol{r}_i = \frac{\log C - \log p_k(\boldsymbol{x})}{\|\nabla_{\boldsymbol{x}} \log p_k(\boldsymbol{x})\|}$$

where C is the target confidence.

Numerical Experiments

- We train two models and compare our method with Goodfellow et al. 2015 (AT), Miyato et al. 2015 (VAT), and Matyasko et al. 2017 (Margin) defenses.
- We measure robustness ρ as follows $\rho_{\text{adv}}(f) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \frac{\|\mathbf{r}(\mathbf{x})\|}{\|\mathbf{x}\|}$ where $\mathcal{D}$ is a test set.
- To generate adversarial noise, we use Moosavi-Dezfooli et al. 2016 (DeepFool), Carlini et al. 2016, and the proposed attack.

Defense	Error %	DeepFool	Carlini et al.	Our
Reference	1.46	0.131	0.124	0.173
AT	0.90	0.228	0.210	0.299
VAT	0.84	0.244	0.215	0.355
Margin	0.84	0.262	0.230	0.453
Our	1.18	**0.290**	**0.272**	**0.575**

Defense	Error %	DeepFool	Carlini et al.	Our
Reference	0.64	0.157	0.148	0.207
AT	0.55	0.215	0.191	0.286
VAT	0.60	0.225	0.195	0.330
Margin	0.54	0.248	0.225	0.470
Our	0.93	**0.288**	**0.278**	**0.590**

Perceptual Experiments

- To additionally study various defenses, we ask human annotators on MTurk to label adversarial images.

Defense	% Change	% No change
Reference	0.57	98.74
AT	19.02	77.21
VAT	35.08	59.68
Margin	60.47	34.52
Our	87.99	9.86

Defense	% Change	% No change
Reference	2.54	96.53
AT	19.1	75.94
VAT	26.8	67.73
Margin	81.77	13.15
Our	92.29	6.51

Column 2: shows percent of adversarial images which human annotator label with its adversarial target, so adversarial noise changed the "true" label of the input. Column 3: shows percent of the adversarial images which human annotator label with its original label, so adversarial noise did not change the underlying label of the input.

Sheet of Adversarial Examples