Adversarial ML

常見名詞
Part.1

常見名詞

Threat Model	Adversarial Falsification	False Negative(FN) (Adversarial example)	FGSM、L-BFGS、DeepFool、Uni. perturbations、C&W
		False Positive(FP) (Fooling example)	
	Adversary’s Knowledge	White-Box	FGSM、L-BFGS、DeepFool、Uni. perturbations、C&W
		Black-Box	
	Adversarial Specificity	Targeted	L-BFGS、C&W
		Non-Targeted	FGSM、DeepFool、Uni. perturbations
	Attack Frequency	One-time	FGSM
		Iterative	L-BFGS、DeepFool、Uni. perturbations、C&W

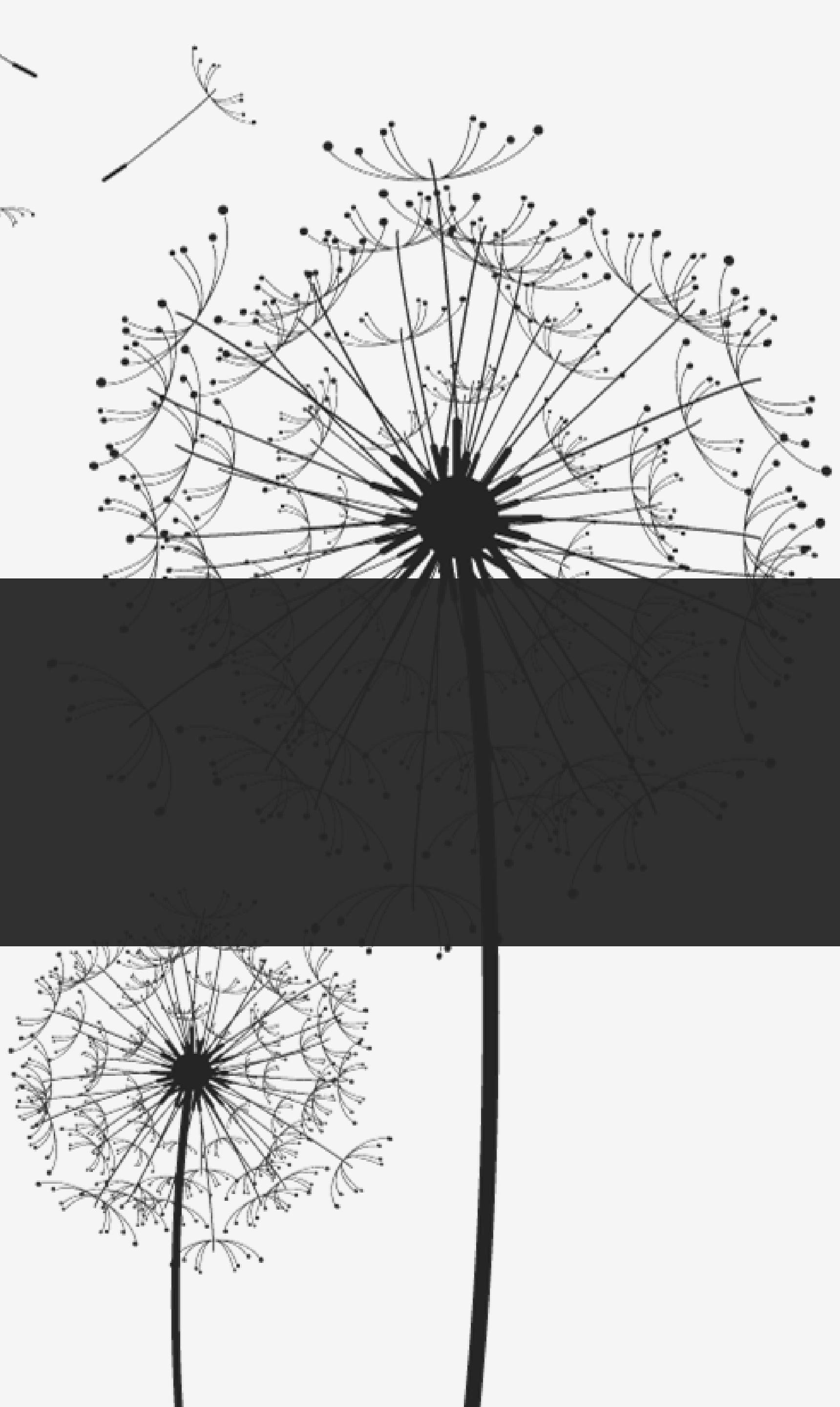
Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems. [Online]. Available: https://arxiv.org/abs/1712.07107

常見名詞

Perturbation	Perturbation Scope	Individual	FGSM、L-BFGS、DeepFool、C&W
		Universal	Uni. perturbations
	Perturbation Limitation	Optimized	L-BFGS、DeepFool、Uni. perturbations、C&W
		Constraint	
		None	FGSM
	Perturbation Measurement (ℓp)	p=0	C&W
		p=1	
		p=2	L-BFGS、Uni. perturbations、DeepFool、C&W
		p=∞	FGSM、DeepFool、Uni. perturbations、C&W

Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems. [Online]. Available: https://arxiv.org/abs/1712.07107

攻撃
Part.2



Box-Constrained L-BFGS

$$\min_{\rho} \; c|\rho| + \mathcal{L}(\mathbf{I}_c + \rho, \ell) \;\; s.t. \;\; \mathbf{I}_c + \rho \in [0, 1]^m$$

- 第一篇提出「Adversarial Example」概念
- Results in the exact solution for a classifier that has a convex loss function.
- Make us understand better the input-to-output mapping represented by the trained network.

C. Szegedy et al. (2013). “Intriguing properties of neural networks.” [Online]. Available: https://arxiv.org/abs/1312.6199

DeepFool

- To find a minimal norm adversarial perturbation for a given image in an iterative manner
- How: 推到分類器的邊界

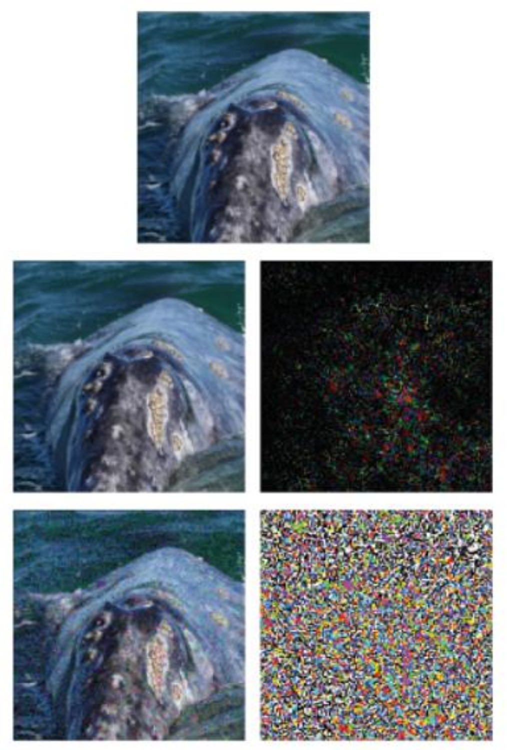


Figure 1: An example of adversarial perturbations. First row: the original image $\boldsymbol{x}$ that is classified as $\hat{k}(\boldsymbol{x})$="whale". Second row: the image $\boldsymbol{x}+\boldsymbol{r}$ classified as $\hat{k}(\boldsymbol{x}+\boldsymbol{r})$="turtle" and the corresponding perturbation $\boldsymbol{r}$ computed by DeepFool. Third row: the image classified as "turtle" and the corresponding perturbation computed by the fast gradient sign method [4]. DeepFool leads to a smaller perturbation.

S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2574–2582 [Online]. Available: https://arxiv.org/abs/1511.04599

DeepFool

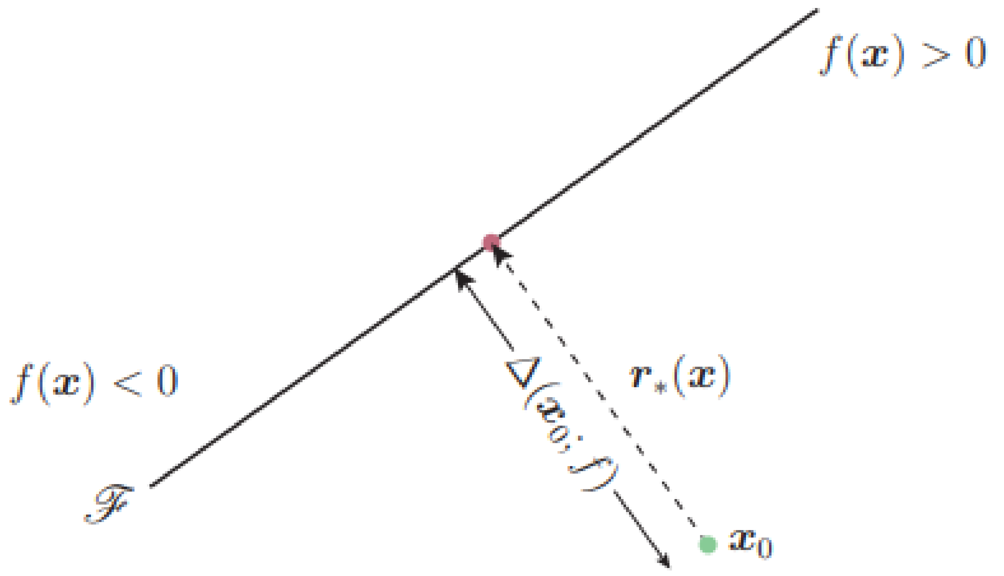


Figure 2: Adversarial examples for a linear binary classifier.

S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2574–2582 [Online]. Available: https://arxiv.org/abs/1511.04599

DeepFool

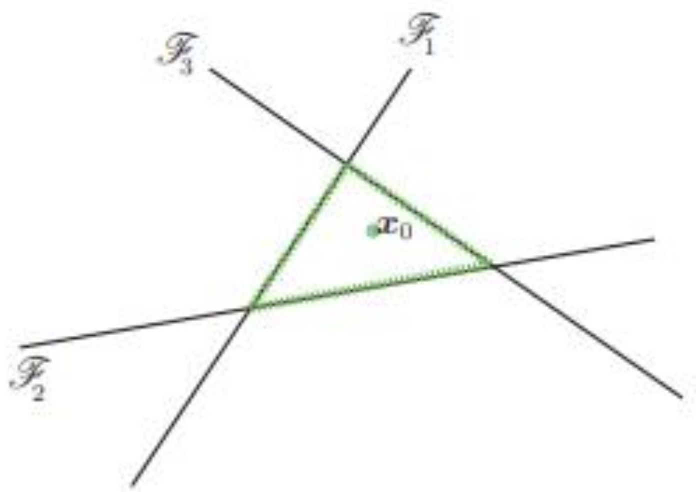


Figure 4: For $\boldsymbol{x}_0$ belonging to class 4, let $\mathscr{F}_k = \{\boldsymbol{x} : f_k(\boldsymbol{x}) - f_4(\boldsymbol{x}) = 0\}$. These hyperplanes are depicted in solid lines and the boundary of P is shown in green dotted line.

S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2574–2582 [Online]. Available: https://arxiv.org/abs/1511.04599

DeepFool

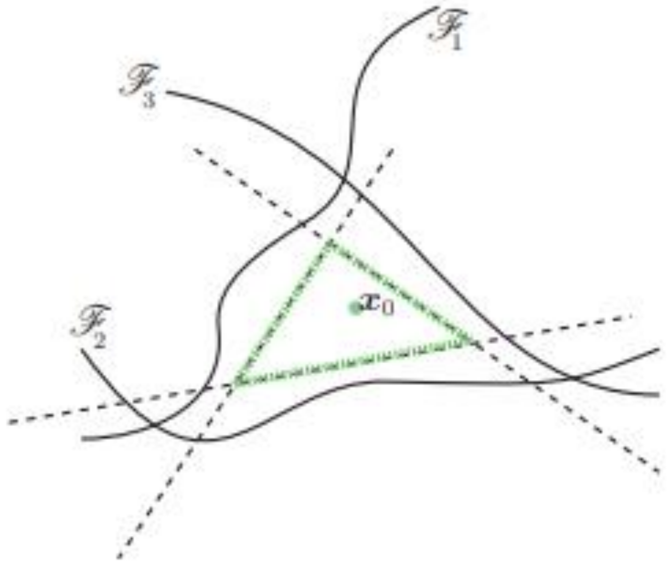


Figure 5: For $\boldsymbol{x}_0$ belonging to class 4, let $\mathscr{F}_k = \{\boldsymbol{x} : f_k(\boldsymbol{x}) - f_4(\boldsymbol{x}) = 0\}$. The linearized zero level sets are shown in dashed lines and the boundary of the polyhedron $\tilde{P}_0$ in green.

S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2574–2582 [Online]. Available: https://arxiv.org/abs/1511.04599

DeepFool

Algorithm 2 DeepFool: multi-class case

input: Image $\boldsymbol{x}$, classifier f.
output: Perturbation $\hat{\boldsymbol{r}}$.

Initialize $\boldsymbol{x}_0 \leftarrow \boldsymbol{x}$, $i \leftarrow 0$.
while $\hat{k}(\boldsymbol{x}_i) = \hat{k}(\boldsymbol{x}_0)$ **do**
    for $k \neq \hat{k}(\boldsymbol{x}_0)$ **do**
        $\boldsymbol{w}'_k \leftarrow \nabla f_k(\boldsymbol{x}_i) - \nabla f_{\hat{k}(\boldsymbol{x}_0)}(\boldsymbol{x}_i)$
        $f'_k \leftarrow f_k(\boldsymbol{x}_i) - f_{\hat{k}(\boldsymbol{x}_0)}(\boldsymbol{x}_i)$
    end for
    $\hat{l} \leftarrow \arg\min_{k \neq \hat{k}(\boldsymbol{x}_0)} \frac{|f'_k|}{\|\boldsymbol{w}'_k\|_2}$
    $\boldsymbol{r}_i \leftarrow \frac{|f'_{\hat{l}}|}{\|\boldsymbol{w}'_{\hat{l}}\|_2^2} \boldsymbol{w}'_{\hat{l}}$
    $\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}_i + \boldsymbol{r}_i$
    $i \leftarrow i + 1$
end while
return $\hat{\boldsymbol{r}} = \sum_i \boldsymbol{r}_i$

S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2574–2582 [Online]. Available: https://arxiv.org/abs/1511.04599

Universal adversarial perturbations

- To find a single pertubation which is able to fool a network on “any” image with high confidence

- How: 漸進的推進到分類的邊界

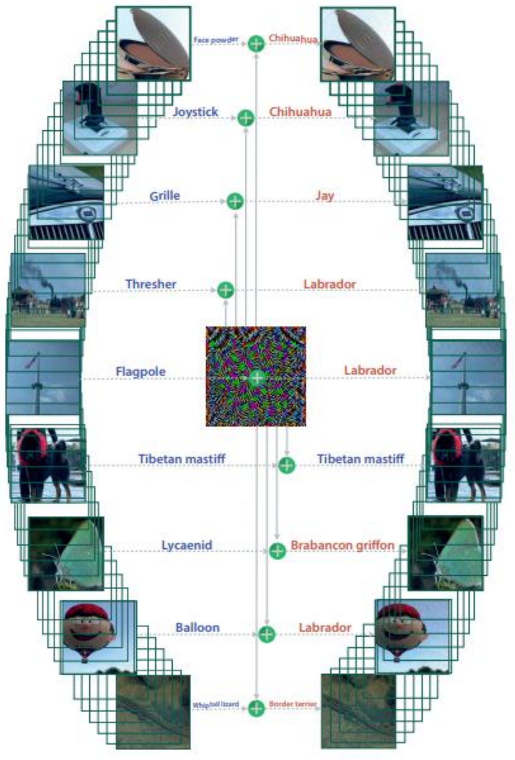


Figure 1: When added to a natural image, a universal perturbation image causes the image to be misclassified by the deep neural network with high probability. *Left images:* Original natural images. The labels are shown on top of each arrow. *Central image:* Universal perturbation. *Right images:* Perturbed images. The estimated labels of the perturbed images are shown on top of each arrow.

S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. [Online]. Available: https://arxiv.org/abs/1610.08401

Universal adversarial perturbations

1. Goal

$$\hat{k}(x+v) \neq \hat{k}(x) \text{ for "most" } x \sim \mu.$$

1. Constraints

- 擾動不能過大

$$\|v\|_p \leq \xi,$$

- 成功率要夠高

$$\mathbb{P}_{x\sim\mu}\left(\hat{k}(x+v) \neq \hat{k}(x)\right) \geq 1-\delta.$$

S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. [Online]. Available: https://arxiv.org/abs/1610.08401

Universal adversarial perturbations

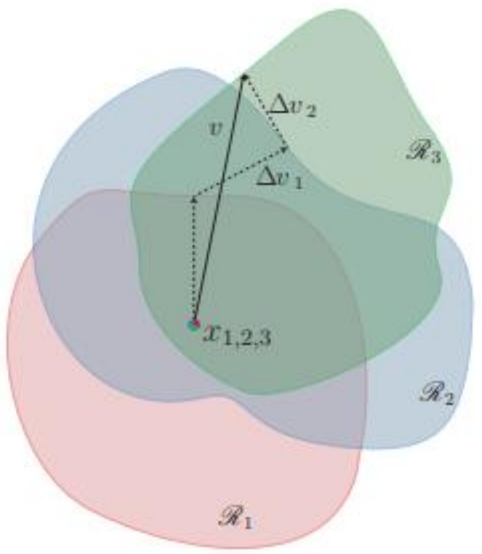


Figure 2: Schematic representation of the proposed algorithm used to compute universal perturbations. In this illustration, data points x_1, x_2 and x_3 are super-imposed, and the classification regions $\mathscr{R}_i$ (i.e., regions of constant estimated label) are shown in different colors. Our algorithm proceeds by aggregating sequentially the minimal perturbations sending the current perturbed points $x_i + v$ outside of the corresponding classification region $\mathscr{R}_i$.

S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. [Online]. Available: https://arxiv.org/abs/1610.08401

Universal adversarial perturbations

3. Algorithm

   a. If v not best pertubation: $$\Delta v_i \leftarrow \arg\min_r \|r\|_2 \text{ s.t. } \hat{k}(x_i + v + r) \neq \hat{k}(x_i).$$

   b. To fit constraint-1: $$\mathcal{P}_{p,\xi}(v) = \arg\min_{v'} \|v - v'\|_2 \text{ subject to } \|v'\|_p \leq \xi.$$

   c. Update v: $$v \leftarrow \mathcal{P}_{p,\xi}(v + \Delta v_i).$$

   d. Stop when: $$\text{Err}(X_v) := \frac{1}{m}\sum_{i=1}^{m} 1_{\hat{k}(x_i+v)\neq\hat{k}(x_i)} \geq 1 - \delta.$$

S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. [Online]. Available: https://arxiv.org/abs/1610.08401

Universal adversarial perturbations

Algorithm 1 Computation of universal perturbations.

input: Data points X, classifier $\hat{k}$, desired ℓ_p norm of the perturbation ξ, desired accuracy on perturbed samples δ.
output: Universal perturbation vector v.
Initialize $v \leftarrow 0$.
while $\mathrm{Err}(X_v) \leq 1 - \delta$ **do**
    for each datapoint $x_i \in X$ **do**
        if $\hat{k}(x_i + v) = \hat{k}(x_i)$ **then**
            Compute the *minimal* perturbation that sends $x_i + v$ to the decision boundary:

$$\Delta v_i \leftarrow \arg\min_r \|r\|_2 \text{ s.t. } \hat{k}(x_i + v + r) \neq \hat{k}(x_i).$$

            Update the perturbation:

$$v \leftarrow \mathcal{P}_{p,\xi}(v + \Delta v_i).$$

        end if
    end for
end while

S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. [Online]. Available: https://arxiv.org/abs/1610.08401

CARLINI AND WAGNER ATTACKS (C&W)

- The objective function to change the label vector

$$
\begin{aligned}
f_1(x') &= -\text{loss}_{F,t}(x') + 1 \\
f_2(x') &= (\max_{i \neq t}(F(x')_i) - F(x')_t)^+ \\
f_3(x') &= \text{softplus}(\max_{i \neq t}(F(x')_i) - F(x')_t) - \log(2) \\
f_4(x') &= (0.5 - F(x')_t)^+ \\
f_5(x') &= -\log(2F(x')_t - 2) \\
f_6(x') &= (\max_{i \neq t}(Z(x')_i) - Z(x')_t)^+ \\
f_7(x') &= \text{softplus}(\max_{i \neq t}(Z(x')_i) - Z(x')_t) - \log(2)
\end{aligned}
$$

N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in Security and Privacy (S&P), 2017 IEEE Symposium on. IEEE, 2017, pp. 39–57. [Online]. Available: https://arxiv.org/abs/1608.04644

CARLINI AND WAGNER ATTACKS (C&W)

- An aderverial method to attack defensive distillation(防禦性蒸餾)
- 攻擊防禦性蒸餾模型實際上很簡單，不考慮這些其他的類向量值，**只考慮需要超過的類向量(目標類)和自身的類向量值即可**，甚至可以只關注增加自身的類向量
  - Ex:

    [-674.3225 , -371.59705 , -177.78831 , 562.87225 ,-1313.5781 , 998.18207 , -886.97107 , -511.58194 ,-126.719666, -43.129272]

N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in Security and Privacy (S&P), 2017 IEEE Symposium on. IEEE, 2017, pp. 39–57. [Online]. Available: https://arxiv.org/abs/1608.04644

CARLINI AND WAGNER ATTACKS (C&W)

- Box Constraint

$$0 \leq x_i + \delta_i \leq 1$$

N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in Security and Privacy (S&P), 2017 IEEE Symposium on. IEEE, 2017, pp. 39–57. [Online]. Available: https://arxiv.org/abs/1608.04644

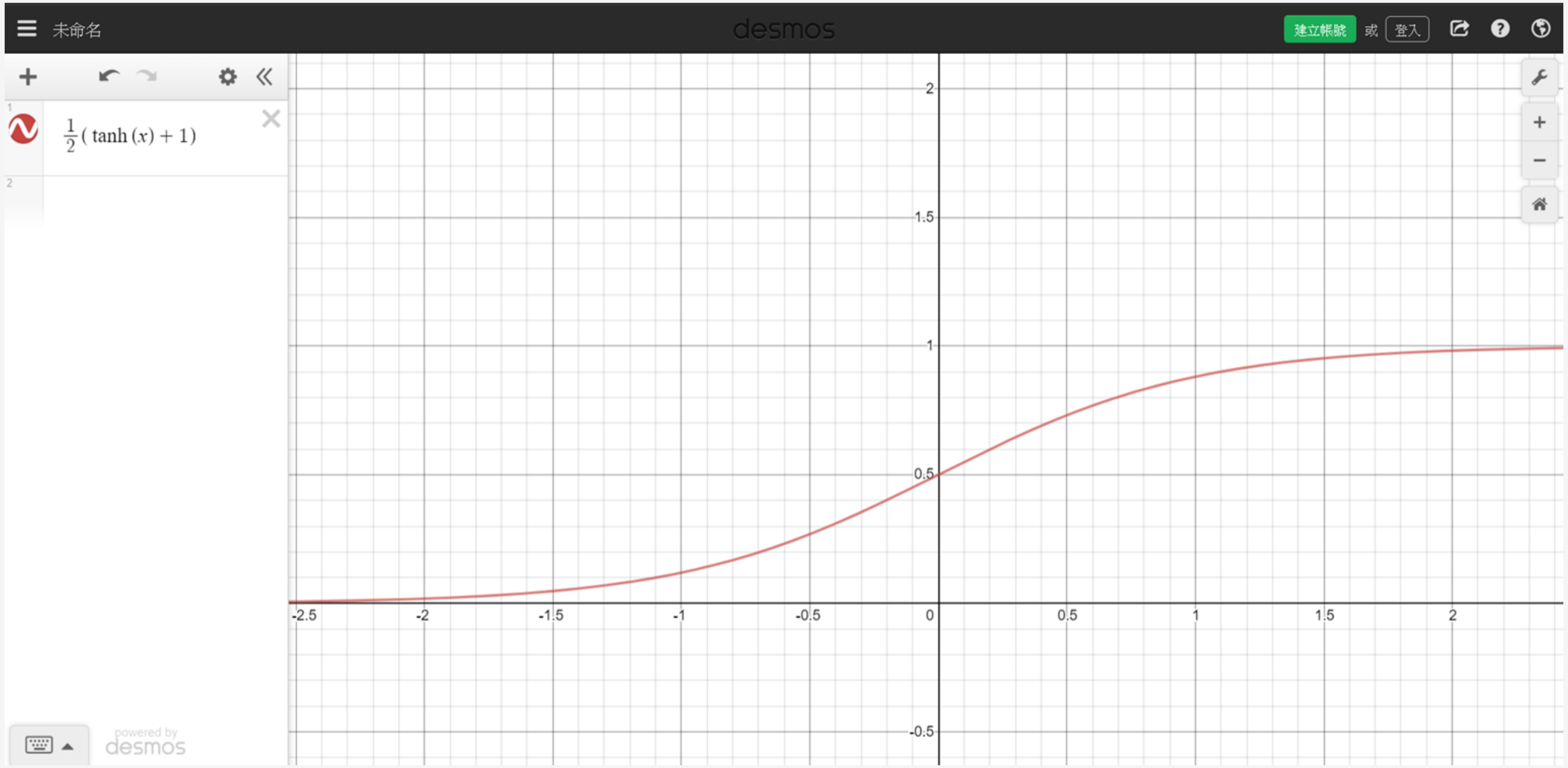
CARLINI AND WAGNER ATTACKS (C&W)

- To solve box constraints
  a. 投影梯度下降法:對於具有復雜更新步驟的梯度下降方法（例如，具有動量的梯度下降），效果不佳
  b. 裁剪梯度下降法:將裁剪直接放入了優化目標，但容易卡在平坦區域，x卡在邊界值動不了
  c. 改變變量:用新的變量w代替原先的x(本篇作者的用法)

$$\delta_i = \frac{1}{2}(\tanh(w_i) + 1) - x_i.$$

N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in Security and Privacy (S&P), 2017 IEEE Symposium on. IEEE, 2017, pp. 39–57. [Online]. Available: https://arxiv.org/abs/1608.04644

CARLINI AND WAGNER ATTACKS (C&W)



N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in Security and Privacy (S&P), 2017 IEEE Symposium on. IEEE, 2017, pp. 39–57. [Online]. Available: https://arxiv.org/abs/1608.04644

CARLINI AND WAGNER ATTACKS (C&W)

- How: L2 attack
  a. Chose the target label t
  b. Our goal is to optimize:

$$\text{minimize} \quad \|\frac{1}{2}(\tanh(w)+1)-x\|_2^2 + c\cdot f(\frac{1}{2}(\tanh(w)+1)$$

$$f(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_t, -\kappa).$$

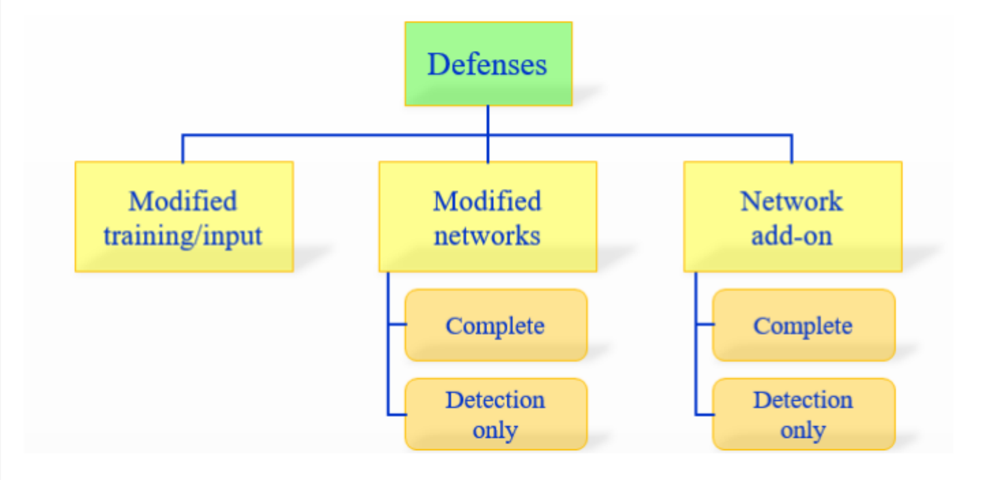
  a. by adjusting c: from 10^-4 to 10^10;
  b. by adjusting k: 錯誤分類發生的置信度

N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in Security and Privacy (S&P), 2017 IEEE Symposium on. IEEE, 2017, pp. 39–57. [Online]. Available: https://arxiv.org/abs/1608.04644

防禦
Part.2

架構

- 三種防禦方法
  a. Training時修改訓練集或testing時修改測試樣本
  b. 更改網路架構-training
  c. 外接其他網路-testing



- 兩種對抗程度
  a. Complete-能辨識回原本label
  b. Detection only-僅辨識是否為攻擊樣本並拒絕分類

Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, 6, 14410-14430. [Online]. Available: https://arxiv.org/abs/1801.00553

Brute-Force Adversarial Training

- Modified Training/Input
- 使用adversarial training的方法，需要來自強大攻擊方法的樣本增加訓練集，使模型正規化，減少overfitting

- Virtual Adversarial Training

- 缺點：依舊能找出新漏洞

Virtual Adversarial Training(VAT)

- 目標函數 $$\frac{1}{N}\sum_{n=1}^{N}\log p(y^{(n)} \mid x^{(n)}, \theta) + \lambda \frac{1}{N}\sum_{n=1}^{N} LDS(x^{(n)}, \theta)$$
- 定義LDS $$LDS(x^{(n)}, \theta) = -\Delta_{KL}(r_{v-adv}^{(n)}, x^{(n)}, \theta)$$
- 定義Rv-adv $$\begin{aligned} \Delta_{KL}(r, x^{(n)}, \theta) &= KL[p(y \mid x^{(n)}, \theta) || p(y \mid x^{(n)} + r, \theta)] \\ r_{v-adv}^{(n)} &= \arg\max_{r}\{\Delta_{KL}(r, x^{(n)}, \theta); ||r|| \leq \epsilon\} \end{aligned}$$

Miyato, T., Maeda, S. I., Koyama, M., & Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE transactions on pattern analysis and machine intelligence, 41(8), 1979-1993. [Online]. Available: https://arxiv.org/abs/1704.03976

Deep Contractive Network

- Modifying the Network
- 借由類似Contractive Auto Encoders的平滑度懲罰項，可防禦L-BGFS

Loss function
T:target
Y:model prediction

$$J_{DCN}(\theta) = \sum_{i=1}^{m} (L(t^{(i)}, y^{(i)}) + \lambda \parallel \frac{\partial y^{(i)}}{\partial x^{(i)}} \parallel_2)$$

Loss function(layer wise)
h:h-th hidden layer

$$J_{DCN}(\theta) = \sum_{i=1}^{m} (L(t^{(i)}, y^{(i)}) + \sum_{j=1}^{H+1} \lambda_j \parallel \frac{\partial h_j^{(i)}}{\partial h_{j-1}^{(i)}} \parallel_2)$$

Gu, S., & Rigazio, L. (2014). Towards deep neural network architectures robust to adversarial examples. arXiv preprint arXiv:1412.5068. [Online]. Available: https://arxiv.org/abs/1412.5068

Gradient Regularization/Masking

- Modifying the Network
- training時減少gradient的變化，或隱藏gradient，有利於對抗基於gradient的攻擊方法

- Masking-based Defense
- 通過在網絡的logit輸出中添加noise，實現了masking based的對於C&W攻擊的防禦。

A Learning and Masking Approach to Secure Learning

- 攻擊方法: ALN
- 防禦方法: DLN、NAC
- DLN
  a. 去噪前後相似
  b. 去噪後要能辨識回原類別
  c. Cat(y)-原類別、D(x)-去噪器、Cp(x)-分類器

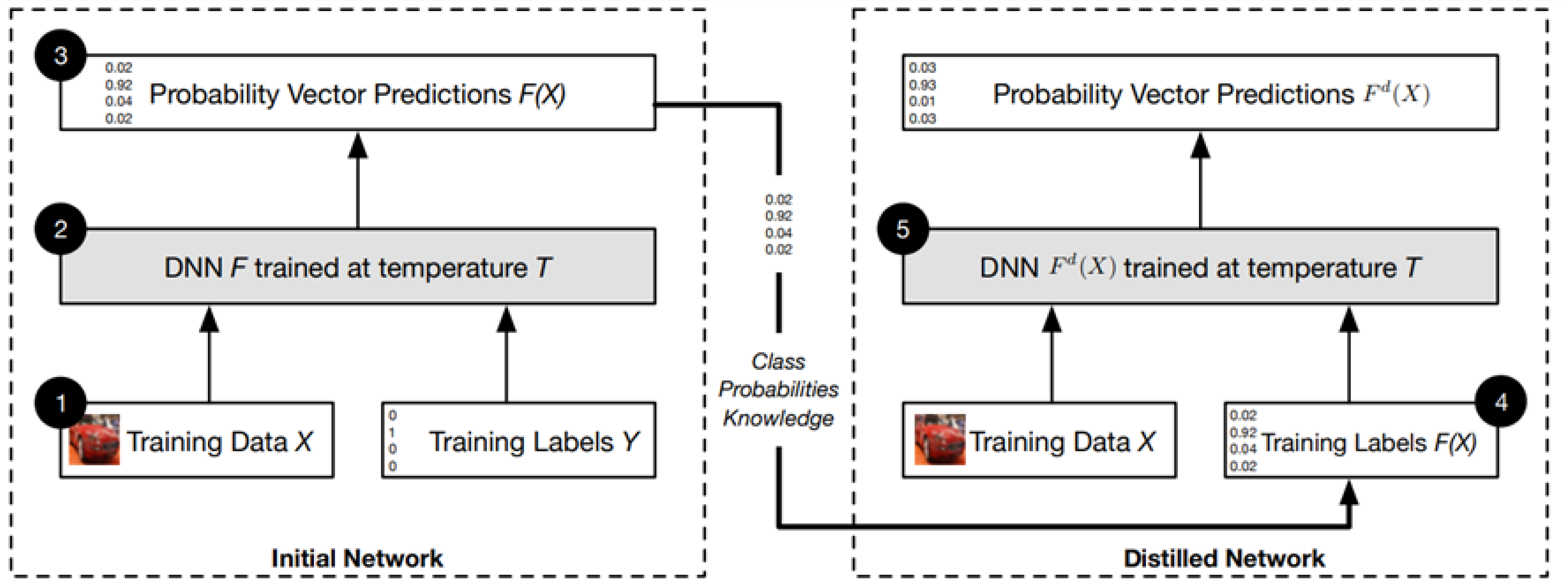
$$\alpha\overline{sim}(x, D(x')) + \overline{opsim}(Cat(y_x), C_p(D(x')))$$

- NAC
  a. 對於低干擾的對抗樣本，它們大多在分類邊界附近，因此可以通過屏蔽分類邊界來愚弄低干擾對抗樣本
  b. 在模型輸出中加入noise

Nguyen, L., Wang, S., & Sinha, A. (2018, October). A learning and masking approach to secure learning. In International Conference on Decision and Game Theory for Security (pp. 453-464). Springer, Cham. [Online]. Available: https://arxiv.org/abs/1709.04447

Defensive Distillation

- Modifying the Network
- Distillation 是指將復雜網絡的知識遷移到簡單網絡上



N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, “Distillation as a defense to adversarial perturbations against deep neural networks,” in Security and Privacy (SP), 2016 IEEE Symposium on. IEEE, 2016, pp. 582–597. [Online]. Available: https://arxiv.org/abs/1712.07107 [Online]. Available: https://arxiv.org/abs/1511.04508

Detector Subnetwork

- Modifying the Network-Detection Only Approaches
- 增加一個子網路辨識是否為擾動樣本，可防禦FGSM、BIM、DeepFool

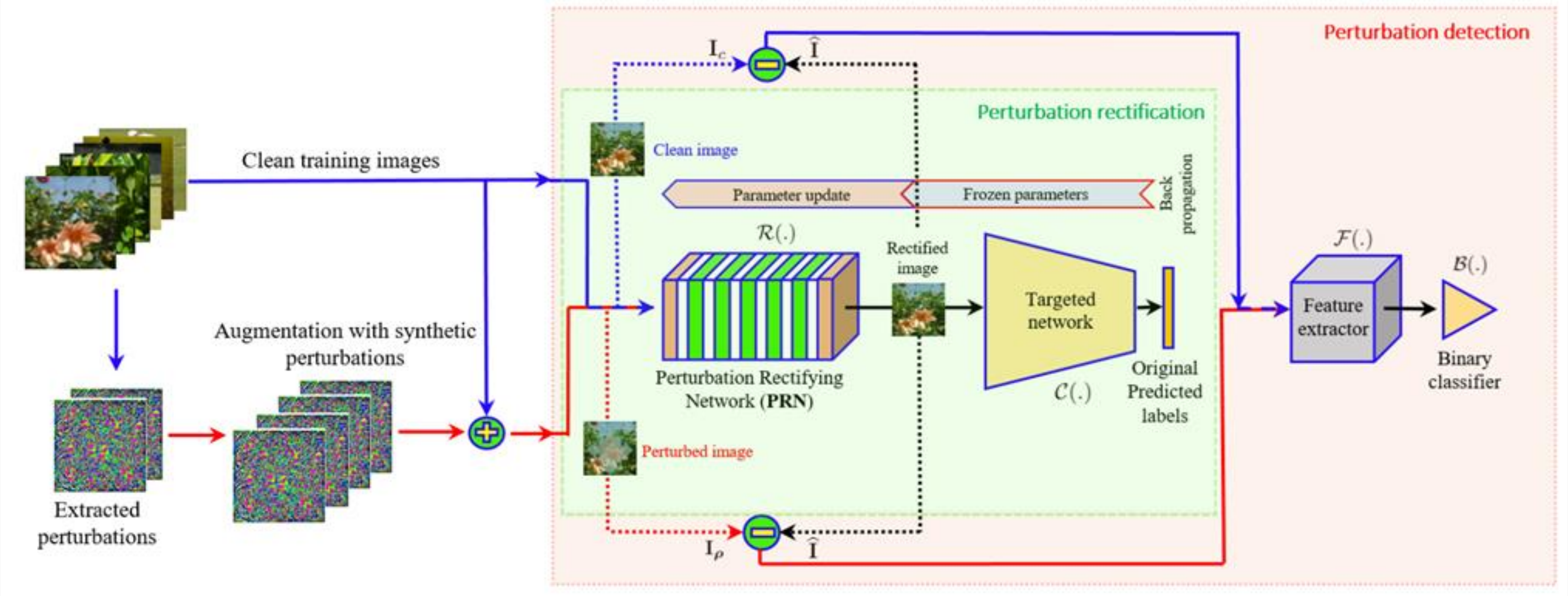
- Additional Class Augmentation
- 方法：增加一個class來辨識攻擊樣本
- 缺點：依舊會被找到漏洞

Lu, J., Issaranon, T., & Forsyth, D. (2017). Safetynet: Detecting and rejecting adversarial examples robustly. In Proceedings of the IEEE International Conference on Computer Vision (pp. 446-454). [Online]. Available: https://arxiv.org/abs/1704.00103

Defense Against Universal Perturbations

- Network Add-ONS
- 增加一個預輸入層-Perturbation Rectifying Network (PRN)

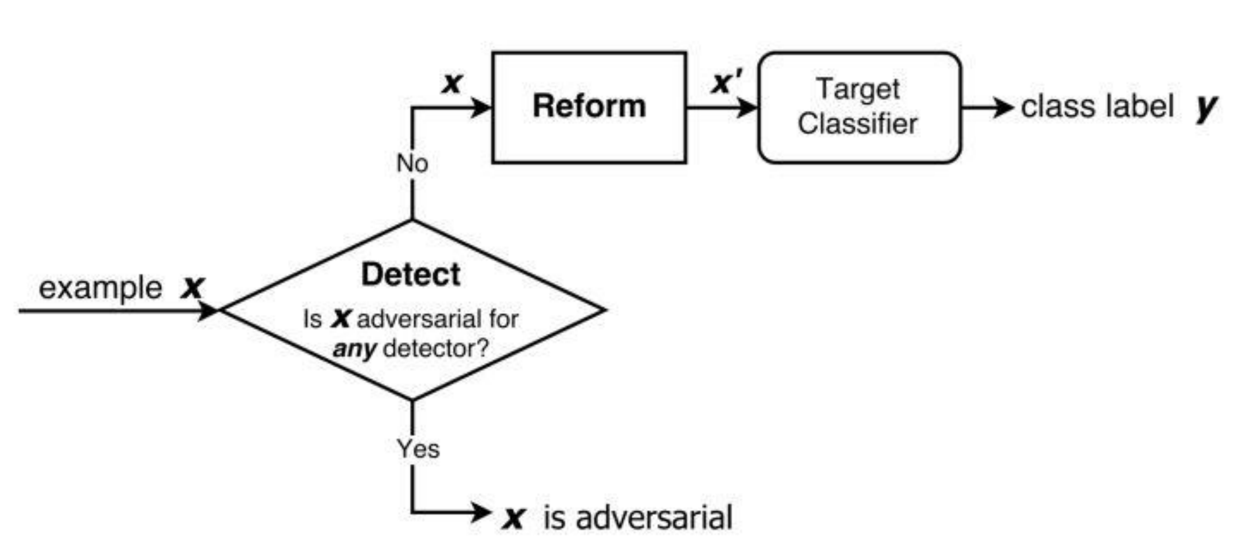
$$\mathcal{J}(\boldsymbol{\theta}_p, \mathbf{b}_p) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(\ell_i^*, \ell_i),$$



Akhtar, N., Liu, J., & Mian, A. (2018). Defense against universal adversarial perturbations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3389-3398). [Online]. Available: https://arxiv.org/abs/1711.05929

MagNet

- Network Add-ONS Detection Only Approaches
- 訓練一個detector來辨識乾淨圖片的manifold，並訓練一個reformer來重構接近manifold邊界的圖片



Meng, D., & Chen, H. (2017, October). Magnet: a two-pronged defense against adversarial examples. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 135-147). ACM. [Online]. Available: https://arxiv.org/abs/1705.09064