Adversarial Examples

Attacks vs. Defenses

Summary of CS294-131 Fa18 8/28/18 Talk

Nicholas Carlini

Organized by Cui Wanqing

Notation:

- $F(x) \rightarrow y$: Classification function F that max input X on some labels Y.
- $F_\theta(x)_L = softmax(F_z(x))$: The probability of L under the distribution outputted by the NN.
- $C(x) = arg\max_L F(x)_L$
- Adversarial Example(对抗样本):

  Image x $\quad C(x) = L$

  Choose a different label T, find x' that $C(x') = T$ and $x' \neq x$

P.S. 在regression中也有类似的现象吗？

Yes，regression中的目标是让“output be maximally wrong”

2014

首次提出对抗样本(Szegedy et al.)：

$$min||x - x'||_2$$

s.t. $C(x) \neq C(x^{'})$

$x^{'}$ valid (for image, $x^{'}$的每个pixel应该是0-255的)

BUT，上述问题不好求解

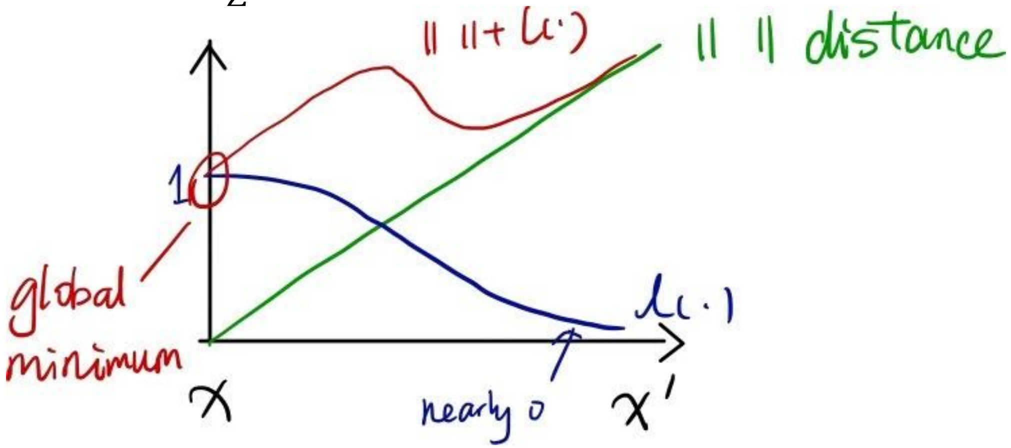
$$\rightarrow min\frac{1}{\alpha}||x - x'||_2 + l(x')$$

其中$l(x') = F(x')_L$，即将x'归为原始label L的confidence

最终会得到x'：x'与x很相似，但x'属于L的信心很低，即x'不再属于原始的label。

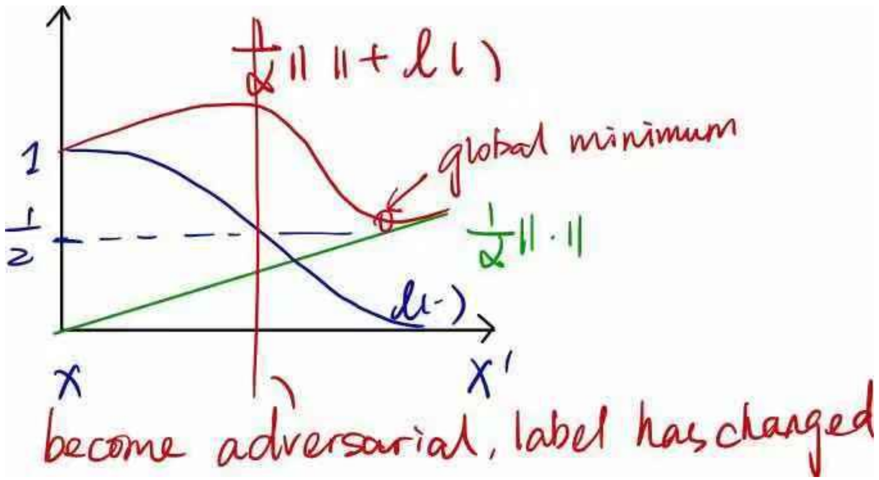
（P.S. 基本假设： 在足够小的L_2限制下，image应该有相同的label）

- $||x-x'||_2+l(x')$图像：



全局最小值并不是想要找的对抗样本

- $\frac{1}{\alpha}||x-x'||_2+l(x')$图像，其中$\alpha$很大：



全局最小值就是想要找的对抗样本，但是有两个问题：

1. 想要找到全局最小需要get over the hill
2. 只需要confidence<1/2就是对抗样本，但全局最小点干扰的太多了，超过了实际需要

2015

FGSM (Goodfellow et al.)：

$$x' = x + \epsilon \cdot sign\left(\nabla_x l\big(F(x)\big)\right)$$

将所有pixel在特定的方向上同时调整ϵ大小

$sign$：对于每个pixel，只关心调整的方向，不关心调整多少

- FGSM是一种生成对抗样本的有效方式
- 反映了NN的决策边界是高度线性的，至少在局部上是这样，所以只需要在某个方向上移动一小步，而不需要做太多花哨的事情就能得到对抗样本

2016

关心对抗样本的原因：

1. Deep learning的核心是“能够做人类做的事情并且do better”，但对对抗样本而言，很明显机器做的还不如人类，所以想要close this gap
2. 安全问题，如自动驾驶
3. From paper *Adversarial Risk and the Dangers of Evaluating Against Weak Attacks*：对抗风险是模型的最坏情况风险的下界

Distillation（蒸馏）

Step 1. train $F(x)$ on (X,Y)

Step 2. 生成$Y^{'} = \{softmax(F_z(x)/T),\ x \in X\}$, T是temperature， T上升，则NN对于他预测结果的信心下降

Step 3. train $G(x)$ on (X,Y')

$F(x)$：teacher，通常比$G(x)$更大，且知道的更多

$G(x)$：student，利用$F(x)$的信息来学习

e.g. $F(x)$学习7→0 0 0 0 0 0 0 1 0 0

$G(x)$学习7→0 .1 .1 0 0 0 0 .8 0 0 → $F(x)$的softmax输出，$F(x)$告诉$G(x)$，7还有点像1和2

因为$G(x)$网络较小，通常情况下直接学习(X,Y)效果没有(X,Y')好

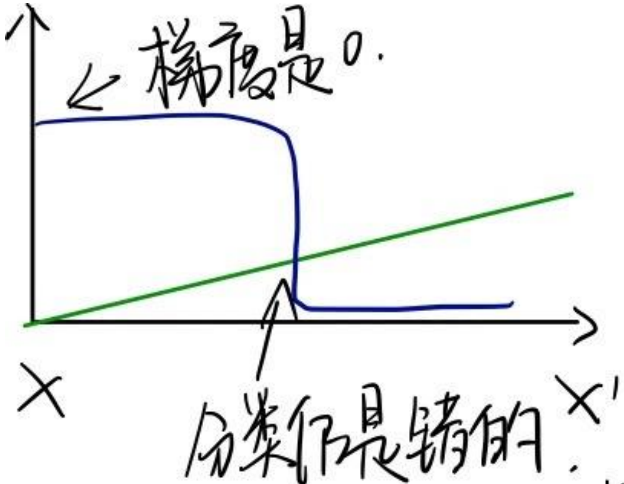
P.S. distillation的初衷是将复杂网络的knowledge移植到简单网络中，以减小模型复杂度和计算复杂度

Distillation as a defence

Slightly different：

Step 2.生成$Y^{'} = \{F_z(x) \cdot T,\ x \in X\}$，$T$是一个较大的数，如：100

train $G_z(x)$ on $\left(X, Y^{'}\right)$: train to match the logits



并未解决对抗样本问题，只是利用梯度无法找到

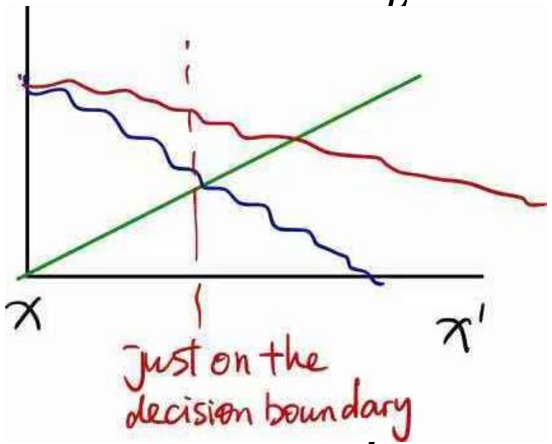
绕过方式：attack by $G_z(x)/T$, $softmax(G_z(x)/T)$ has gradients

2017

- March (C+W)：

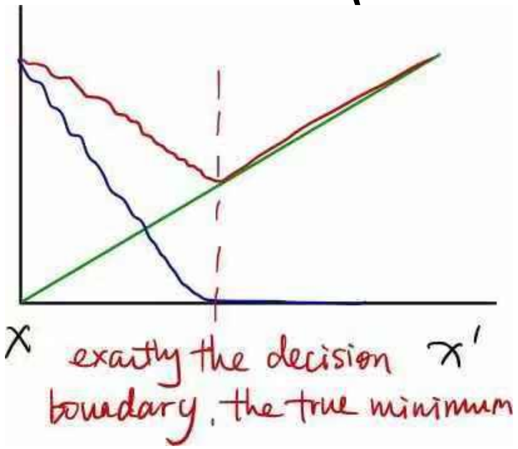
Original： $min||x-x'||_2+\alpha\cdot l(x')$， $l(x')$本质上是softmax

→令$l(x)=F_z(x)_L$， logits是大致线性的



但是并不会及时停下来，会一直沿着梯度方向移动

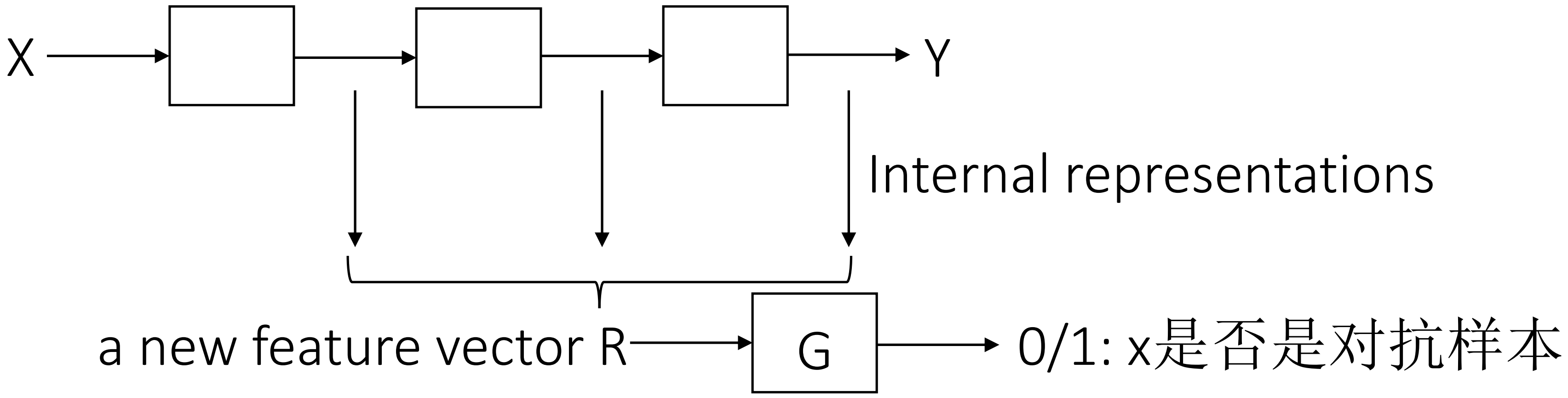
→令$l(x)=\max\left(0, F_z(x)_L-\left(\max_{T\neq L}F_z(x)_T\right)\right)$



- $\max_{T\neq L}F_z(x)_T$:对next most likely label的confidence
- $F_z(x)_L-\left(\max_{T\neq L}F_z(x)_T\right)$：使confidence on true label比false label小
- $\max\left(0, F_z(x)_L-\left(\max_{T\neq L}F_z(x)_T\right)\right)$：一旦达到了决策边界，即false label的信心大于true label，就不会在得到任何好处了

- April (Matsan et al.)

Normal training：



Classifier：ask G 输入是否为对抗样本？

if YES，just reject it

if NO，run the classifier

BUT，对于有些情况，reject是不允许的，如自动驾驶，对所有的路标都要处理

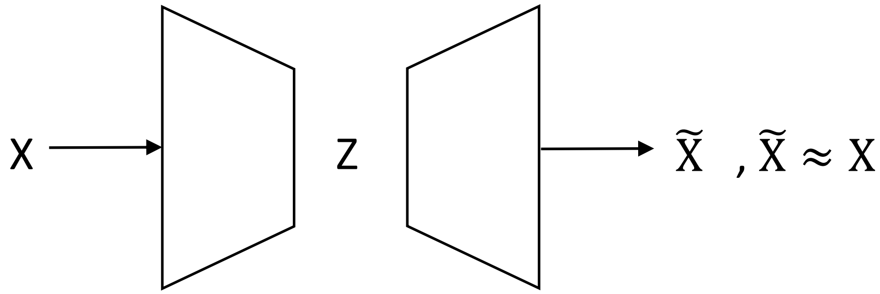
- August

➢ C+W

$$min||x-x'||_2+\alpha\cdot l(x^{'})+\beta\cdot l_0(x^{'})$$

➢ Meng et al.

AutoEncoder:



First：check does $\|AE(x)-x\|<\varepsilon$?

P.S. 因为AE是在clean data上训练的，所以对clean data成立，不会有太大的负面影响

Then：if YES, $F(x)$

if NO, $F\big(AE(x)\big)$

此外，AE是私有的，可以每天训一个新的

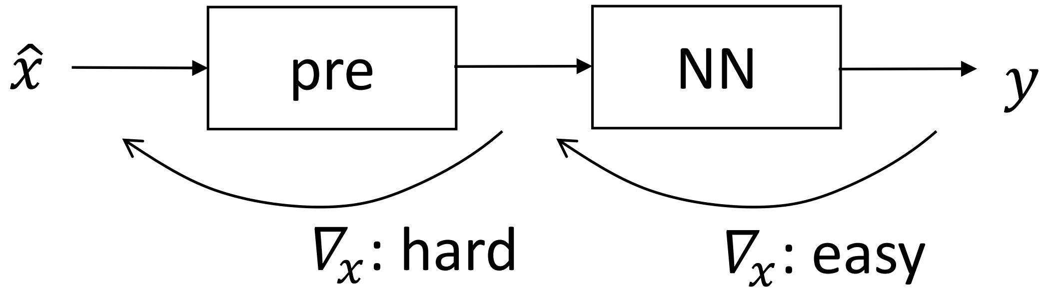
绕过方式：$min||x-x'||_2+\alpha\cdot l(x^{'})+\beta\cdot E_i l_{D_i}(x^{'})$

其中$E_i l_{D_i}(x^{'})$是作为攻击方，用与防御方相同的训练方式训n个AEs，然后求他们的期望

2018

- March：

Defense：



所以认为解决了对抗样本的问题

e.g. Gou et al. : JPEG压缩

Xie et al. : “Quilting”

Buckman et al. : “Thermometer encoding”

BUT，因为$pre(\hat{x}) \approx x$

所以$\nabla_x F\big(pre(x)\big)\big|_{x=\hat{x}} \approx \nabla_x F(x)\big|_{x=pre(\hat{x})}$，虽然是近似，但是迭代多轮足以找到对抗样本（*Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples*）

- April：

Madry et al.

Original：

$$\arg\min_{\theta} l\big(F(x)\big)$$

Now：

$$\arg\min_{\theta}[\max_{\delta} l(F(x+\delta))]$$

BUT，只对训练时使用的approach有用，若换了新的生成对抗样本的模式就不work了，但是，至今为止，基于梯度的这种生成方式是最强的。

Anyway，Nicholas Carlini认为这个方法（在一定程度上）是真正有效的。

Summary of Papers

Organized by Li Jianing

Task Definition

- Adversarial example:

  Given an image x and classifier $f(\cdot)$, an adversarial example (Szegedy et al., 2013) x' satisfies two properties: $\mathcal{D}(x, x')$ is small for some distance metric $\mathcal{D}$, and $c(x') \neq c^*(x)$. That is, for images, x and x' appear visually similar but x' is classified incorrectly.

- Attack: generate adversarial examples to confuse the classifier

- Defend: make the classifier robust to adversarial examples

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

Anish Athalye, Nicholas Carlini,David Wagner

Contributions

- Propose attacks strategies for obfuscated-gradient based defenses
    - Shattered gradients
    - Stochastic Gradients
    - Exploding & Vanishing Gradients

Obfuscated gradient cases

- Shattered gradients
  - The gradient is not available
  - E.g. Use non-differentiable preprocessing
- Strategy: backward pass differentiable approximation (BPDA)
  - For a non-differentiable layer f(x), find a differentiable g(x) to approximate
  - Use g(x) instead of f(x) on the backward pass only

Obfuscated gradient cases

- Stochastic Gradients
  - The gradient is randomized
  - E.g. Use randomized transformation
- Strategy: Expectation over transformation (EOT)
  - Optimize the expectation over the transformations instead of a single transformations

Obfuscated gradient cases

- Exploding & Vanishing Gradients
  - The back-propagated gradient is either exploding or vanishing
  - E.g. Use optimization loop to transform the input to a new input
- Strategy: Reparameterization
  - For the loop g(x), find a differentiable h(z) s.t. g(h(z))=h(z) for all z
  - Use h(z) instead of x, then gradients can be computed through f(h(z))

Adversarial Risk and the Dangers of Evaluating Against Weak Attacks

Jonathan Uesato, Brendan O’Donoghue, Aaron van den Oord, Pushmeet Kohli

Contributions

- Mathematical formularization of adversarial attacks and defenses
  - Worst-case risk
  - Adversarial risk
  - Surrogate adversarial risk
  - Obscurity

$$\sup_{(x,y)\in \mathrm{supp}(D)} \ell(m_\theta(x), y)$$

$$L(\theta) = \mathop{\mathbb{E}}_{(x,y)\sim D}\left[\sup_{x'\in N_\epsilon(x)} \ell(m_\theta(x'), y)\right]$$

$$\hat{L}(\theta, f) = \mathop{\mathbb{E}}_{(x,y)\sim D} \ell(m_\theta(f(\theta, x, y)), y)$$

$$\mathrm{Obscurity}(\theta, f) = L(\theta) - \hat{L}(\theta, f)$$

Contributions

- Empirically show that existing defenses fail to be truly adversarial robust

Dataset	Defense strategy	Original Evaluation	Adversarial Accuracy Bound	Obscurity Bound
CIFAR-10 ($\epsilon = 8$)	PixelDefend (Song et al., 2017)	75%	<10%	>65%
	Adversarial Training (Madry et al., 2017)	47%	<47%	>0%
ImageNet ($\epsilon = 2$)	Non-differentiability (Guo et al., 2017)	15%	0%	15%
	Stochasticity (Xie et al., 2017)	36%	<1%	>35%
	High-level Guided Denoiser (Liao et al., 2017)	75%	0%	75%

Proposed attack strategies

- Projected gradient descent (PGD)
  - Gradient-based method
  - Iterative update with Euclidean projection

$$x^+ = \Pi_{N_\epsilon(x_0)}(x + \alpha \nabla_x J_\theta^{\text{adv}}(x))$$

- Simultaneous perturbation stochastic approximation (SPSA)
  - Gradient-free method
  - Estimate gradient with average directional difference
- Transfer-based attacks
  - Use a surrogate model to mimic the unknown model

Discussion