حملات مقابله‌ای و خصمانه و مقایسه راه‌های مقابله با آن

استاد راهنما: دکتر محمدباقر شمس الهی

استاد درس: دکتر ترانه اقلیدس

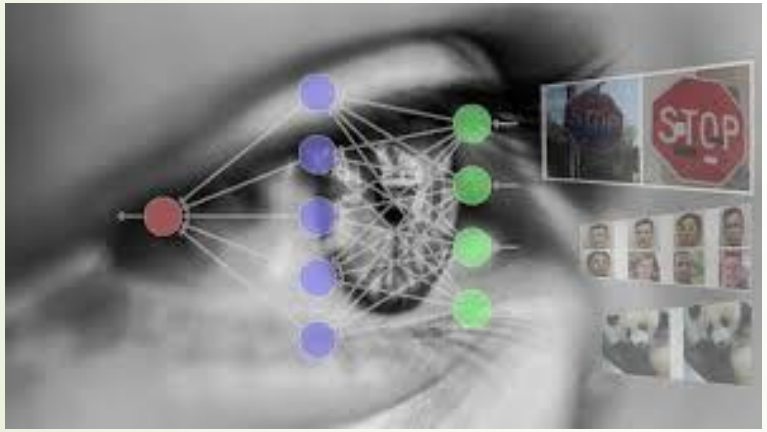
ثنا امین ناجی

پاییز ۱۴۰۱

دانشگاه صنعتی شریف

رئوس مطالب

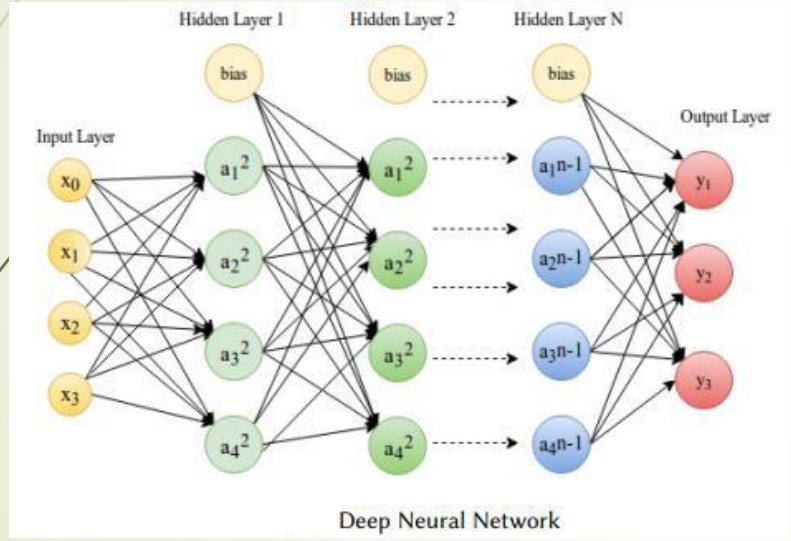
- تعریف شبکه یادگیری عمیق
- تعریف حمله متخاصم و مقابله‌ای
- بررسی مثال‌های حملات و مقابله با آن
- نتیجه گیری و زمینه پژوهشی
- مراجع

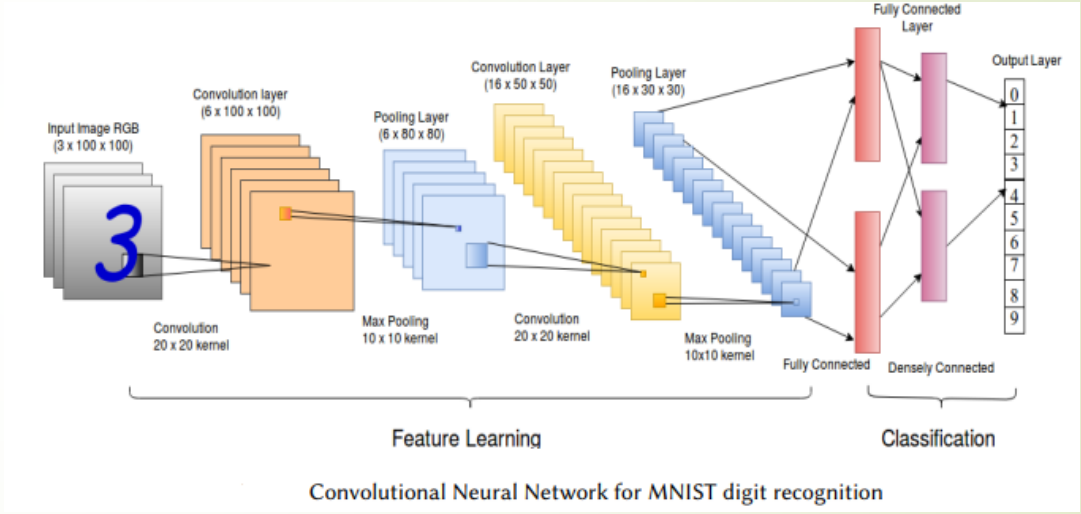


شبکه یادگیری عمیق

- شبکه چند لایه پردازشی برای یادگیری و مدل سازی تجارب و وقایع دنیای واقعی.
- جایگزین مناسب برای شبکه های یادگیری معمول که توان محاسبه بالایی ندارند.
- دو نوع مختلف:
  - DNN
  - CNN

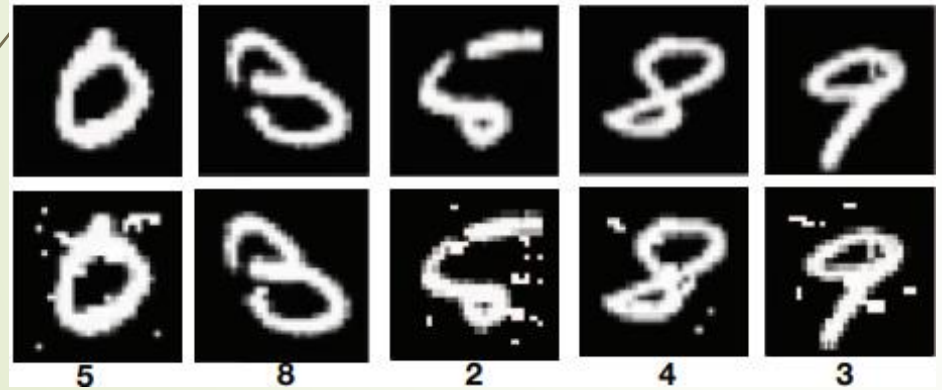
شبکه یادگیری عمیق





حمله متخاصم و مقابله‌ای

- دستکاری وزودی‌های مجاز یک سیستم یادگیری عمیق که با چشم انسان قابل تشخیص نبوده اما باعث دستیابی به خروجی اشتباه میشود.

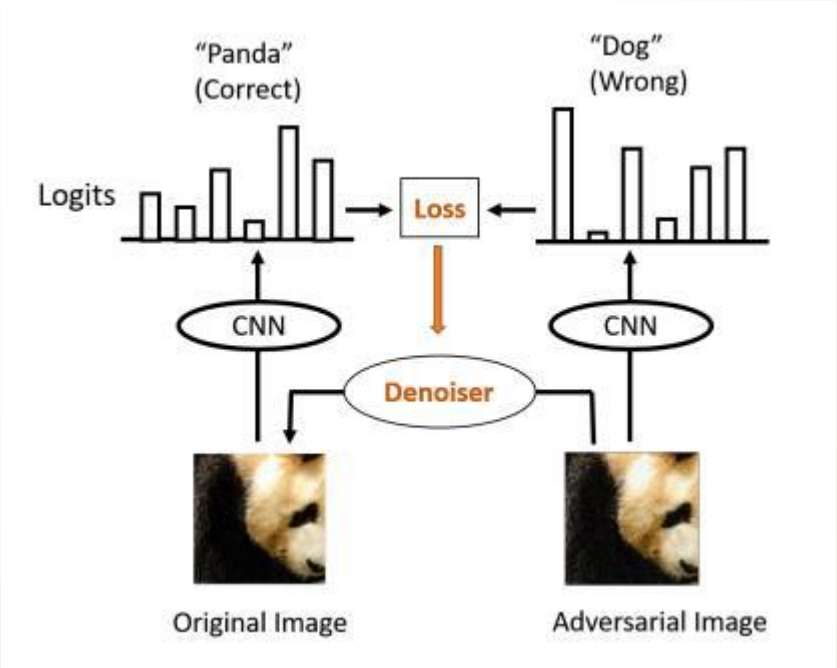


بررسی مثال‌های حملات و مقابله با آن

- گام‌های بررسی:
  - توصیف کلی دیتاست و شبکه یادگیری استقاده شده برای طبقه بندی
  - توصیف شیوه یا شیوه‌های تولید حمله
  - تحلیل نوع دفاع و مقاوم سازی استفاده شده
  - مقابسه و تحلیل نتایج پیاده سازی

مثال اول: Denoiser

- از آنجا که عمده حملات از طریق اضافه کردن یک نویز ناچیز به ورودی‌های اولیه تولید میشوند در نتیجه یک راه مقابله با آنها استفاده از denoiser است:



مثال اول: Denoiser

- Pixel guided denoiser(PGD)

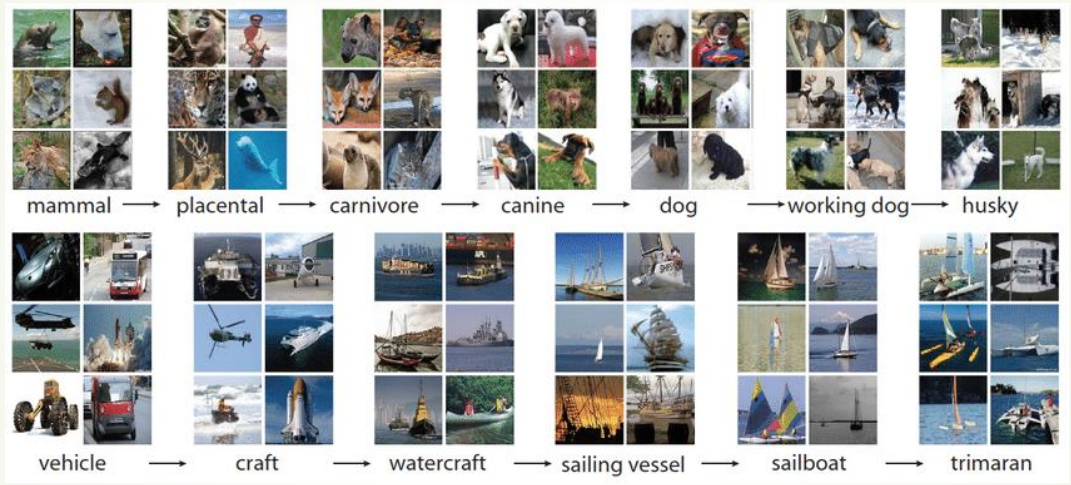
$$L = \left|\left|x - \hat{x}\right|\right|$$

- (-): small perturbation is progressively amplified by deep neural networks →wrong prediction. Even if the denoiser can significantly suppress the pixel-level noise, the remaining noise may still distort the target model.
- High-level representation guided denoiser(HGD)

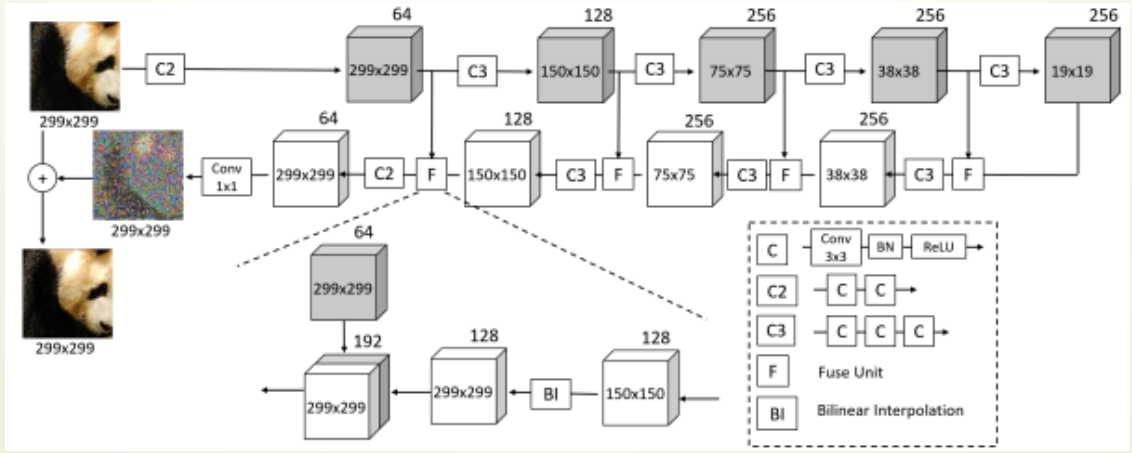
$$L = ||f_l(\hat{x}) - f_l(x)||$$

مثال اول: Denoiser

- دیتاست:
  - ImageNet



- شبکه:



مثال اول: Denoiser

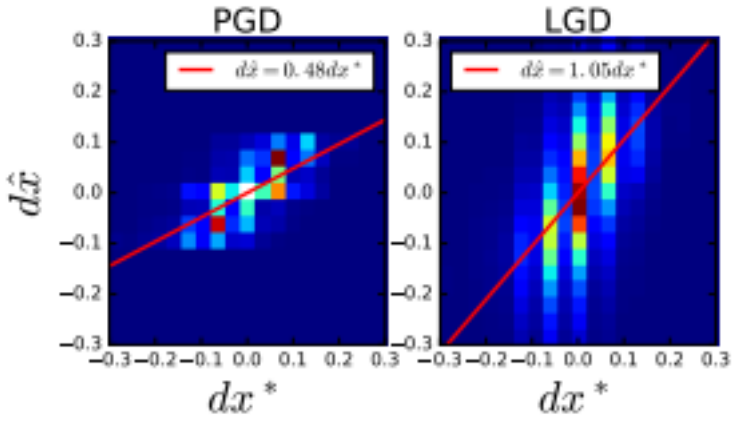
- نتایج:

The transferability of HGD to different model. Resnet is used as the target model.

Denoiser for Resnet	Clean	WhiteTestSet		BlackTestSet	
		$\epsilon = 4$	$\epsilon = 16$	$\epsilon = 4$	$\epsilon = 16$
NA	**78.5%**	63.3%	38.4%	67.8%	48.6%
IncV3 guided LGD	77.4%	75.8%	71.7%	76.1%	72.7%
Resnet guided LGD	78.4%	**76.1%**	**72.9%**	**76.5%**	**74.6%**

The transferability of HGD to different classes. The 1000 ImageNet classes are separated in training and test test.

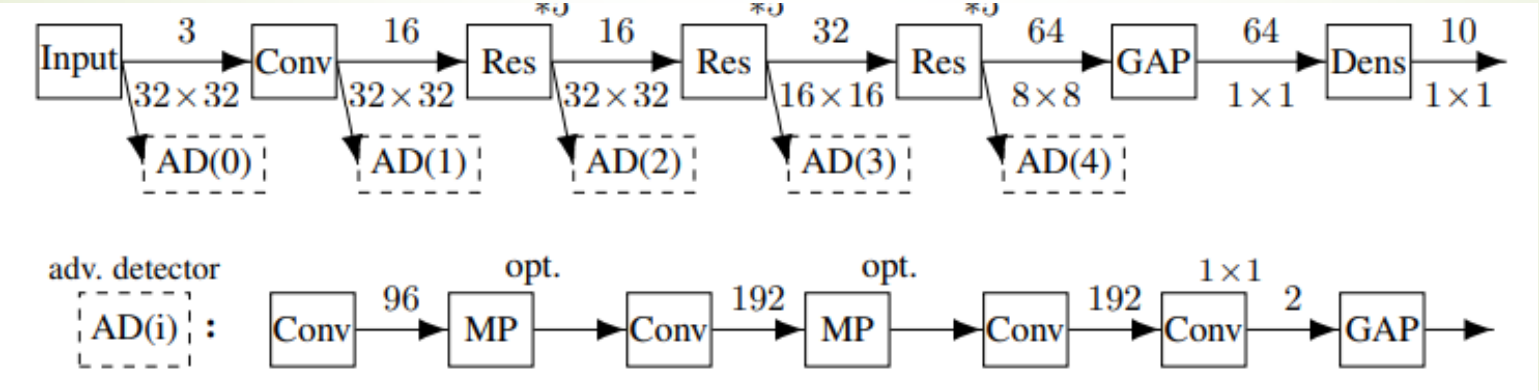
Defense	Clean	WhiteTestSet		BlackTestSet	
		$\epsilon = 4$	$\epsilon = 16$	$\epsilon = 4$	$\epsilon = 16$
NA	**76.6%**	15.4%	15.3%	61.5%	41.7%
LGD	76.3%	**73.9%**	**65.7%**	**74.8%**	**72.2%**



The relationship between dx^* and $d\hat{x}$ in PGD and HGD.

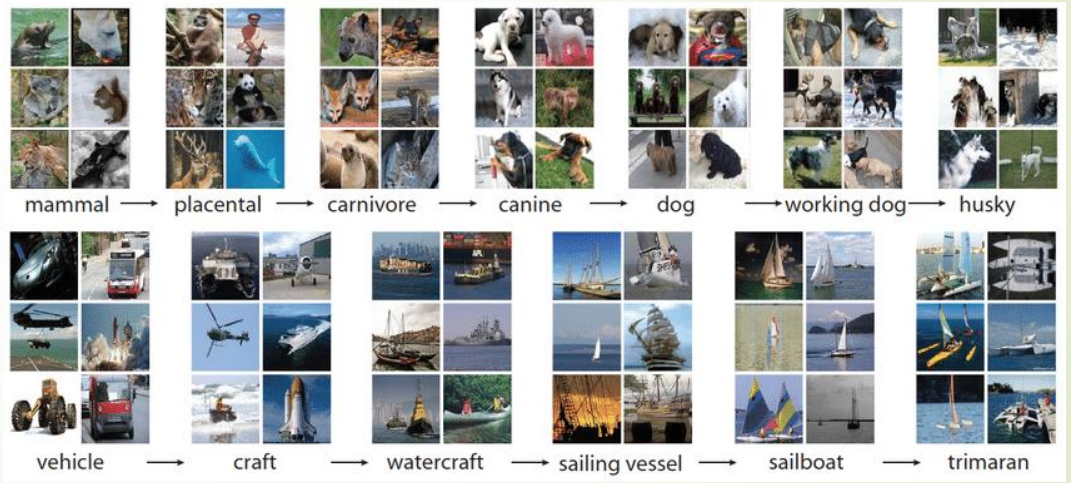
مثال دوم: detector subnetwork

- شبکه اصلی عمیق را با زیر شبکه های به نسبت کوچک تقویت میکنیم، این زیرشبکه‌ها به صورت شاخه‌ای از برخی لایه‌های شبکه اصلی می آبند.
- خروجی زیر شبکه تشخیصی $P_{adv} \in [0,1]$ است.
- دو نوع ورودی:
  - ثابت
  - پویا



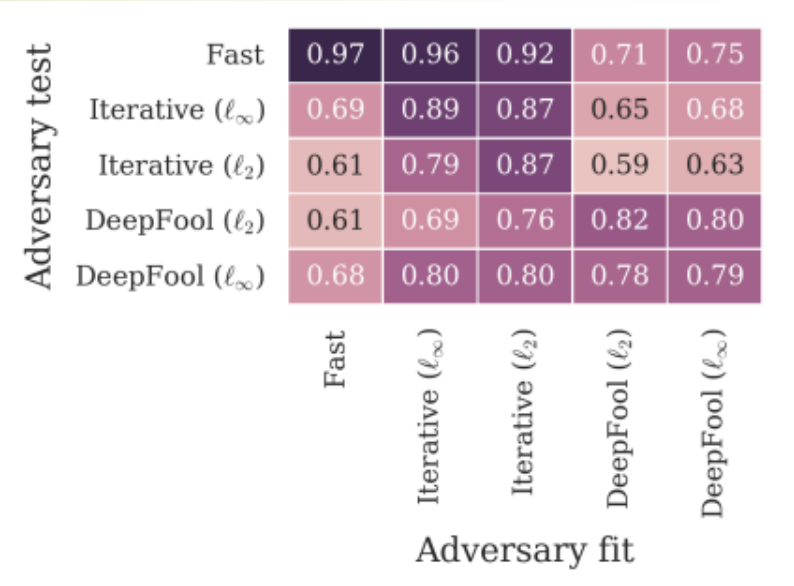
مثال دوم: detector subnetwork

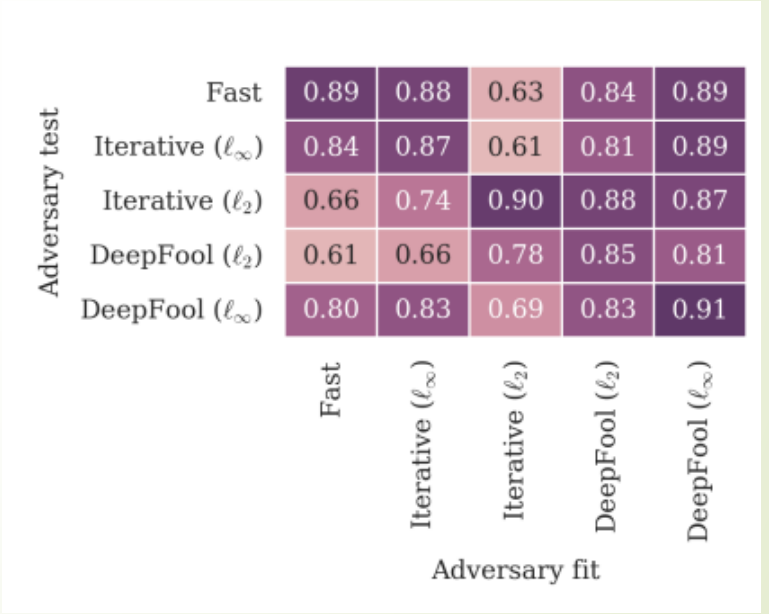
- دیتاست:
  - 32-layer Residual Network <-- CIFAR10
  - pretrained VGG16 <-- 10-CLASS IMAGENET



مثال دوم: detector subnetwork

- نتایج:





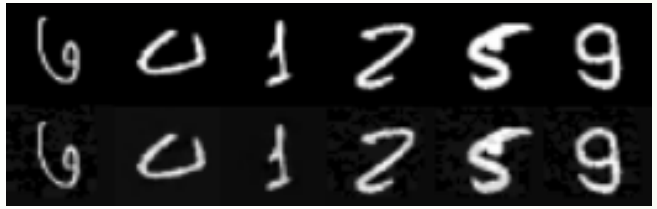
مثال سوم: Robust Optimization

- هدف کلی: پایدار کردن پاسخ های سیستم به ورودی های نزدیکتر به هم. (برای یک همسایگی نزدیک به ورودی های معتبر.)
- فرمول کلی برای تابع هزینه:

$$\min_{\theta} \tilde{J}(\theta,x,y) = \min_{\theta} \sum_{i=1}^{m} \max_{\tilde{x}_i \epsilon U_i} J(\theta,\tilde{x}_i,y_i)$$

مثال سوم: Robust Optimization

- دیتاست ۱: MNIST

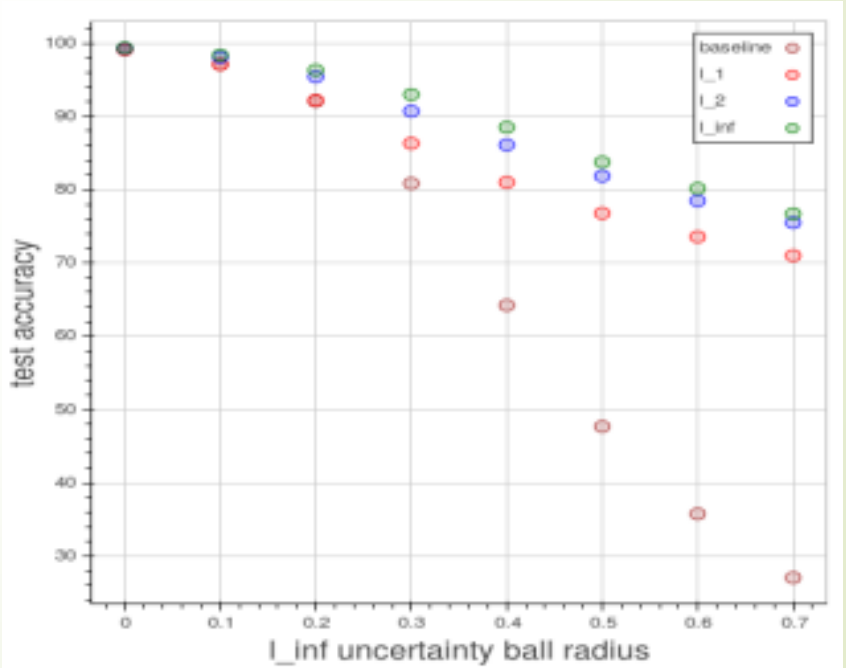


- شبکه:
  - two convolutional layers (containing 32 and 64 5×5 filters), max pooling (3×3 and 2×2) after every convolutional layer, and two fully connected layers (of sizes 200 and 10) on top.

مثال سوم: Robust Optimization

- نتایج دیتاست ۱

Net	MNIST test set	$\mathcal{A}_{\text{mnist}}$
Baseline	99.09%	0%
Robust ℓ_1	99.16%	33.83%
Robust ℓ_2	99.28%	76.55%
Robust ℓ_∞	99.33%	79.96%

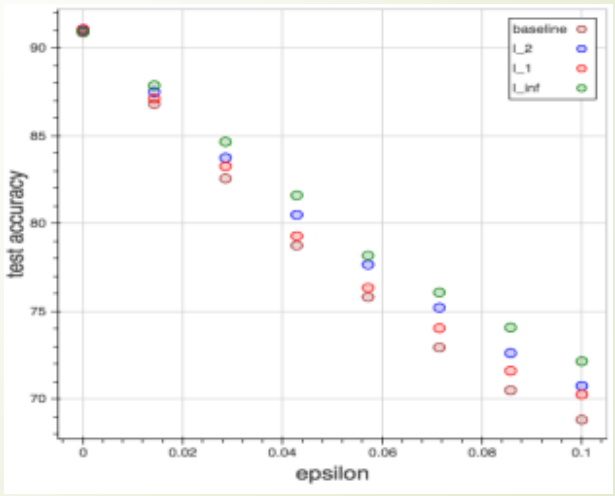


مثال سوم: Robust Optimization

- دیتاست2: CIFAR-10

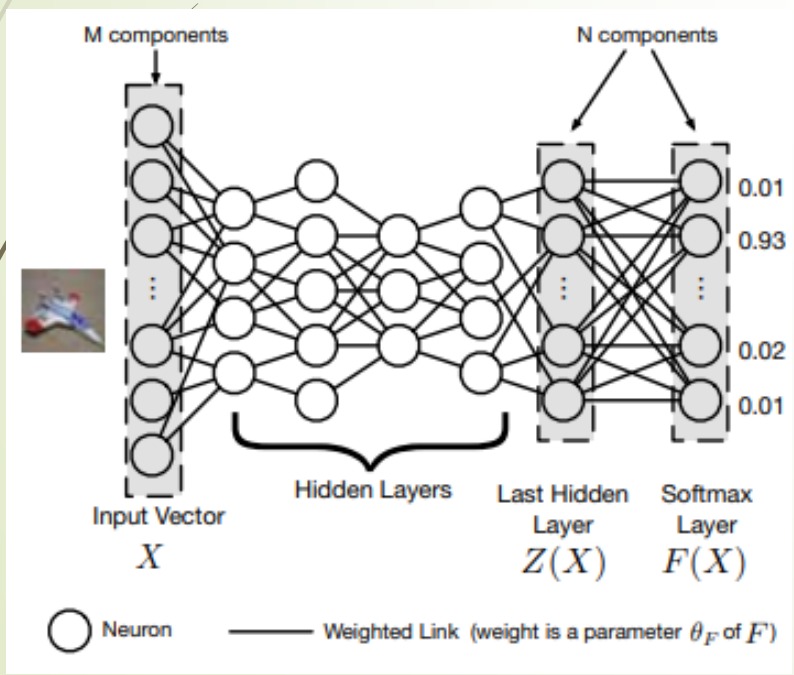
- شبکه:
  - e VGG net, publicly available online
- نتایج:

Net	CIFAR-10 test set	$\mathcal{A}_{\text{cifar10}}$
Baseline	90.79%	0%
Robust ℓ_1	91.11%	56.31%
Robust ℓ_2	91.04%	59.92%
Robust ℓ_∞	91.36%	65.01%



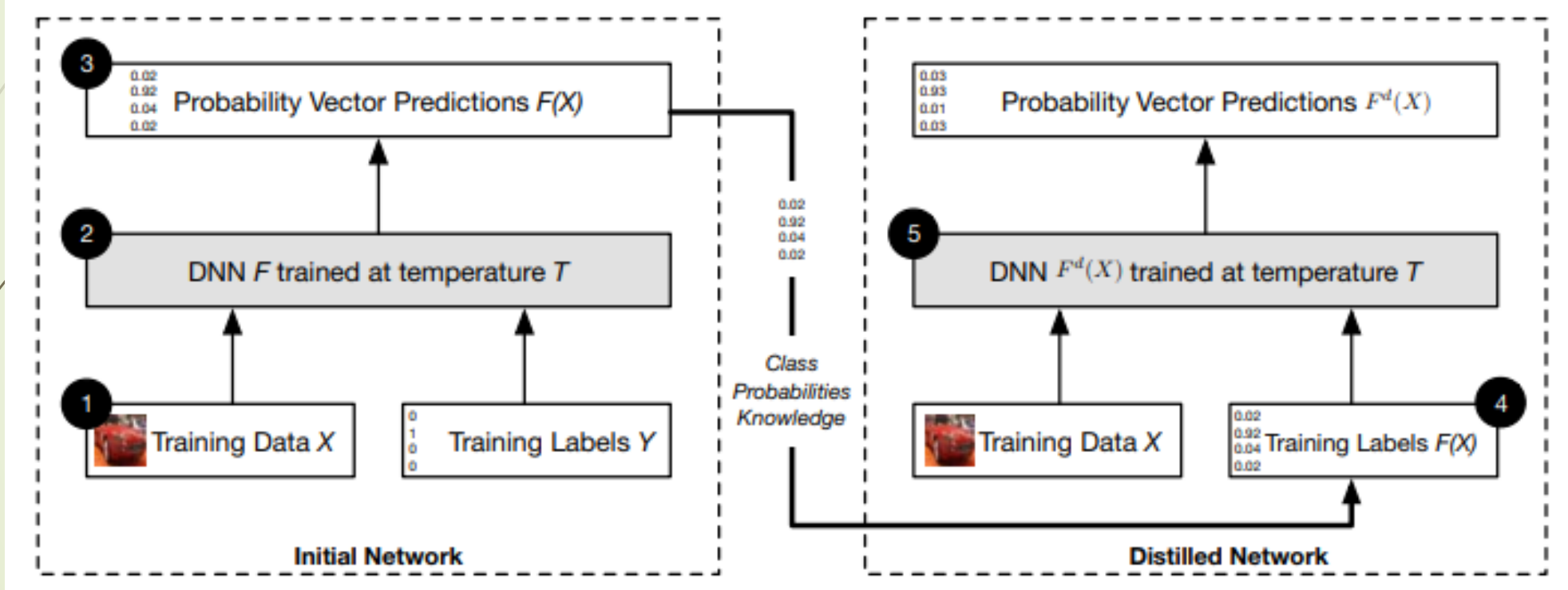
مثال چهارم: Distillation

- تقطیر: به ظور کلی نوعی روش برای آموزش شبکه است. هدف آن طراحی یک شبگه ثانویه با تعداد کمتری پارامتر و بار محاسباتی پایین تر با استفاده از اطلاعات سیستم اولیه است.



$$F(X) = \left[\frac{e^{z_i \cdot (x)/T}}{\sum_{1=0}^{w-1} e^{z_l(x)/T}} \right]$$

مثال چهارم: Distillation



مثال چهارم: Distillation

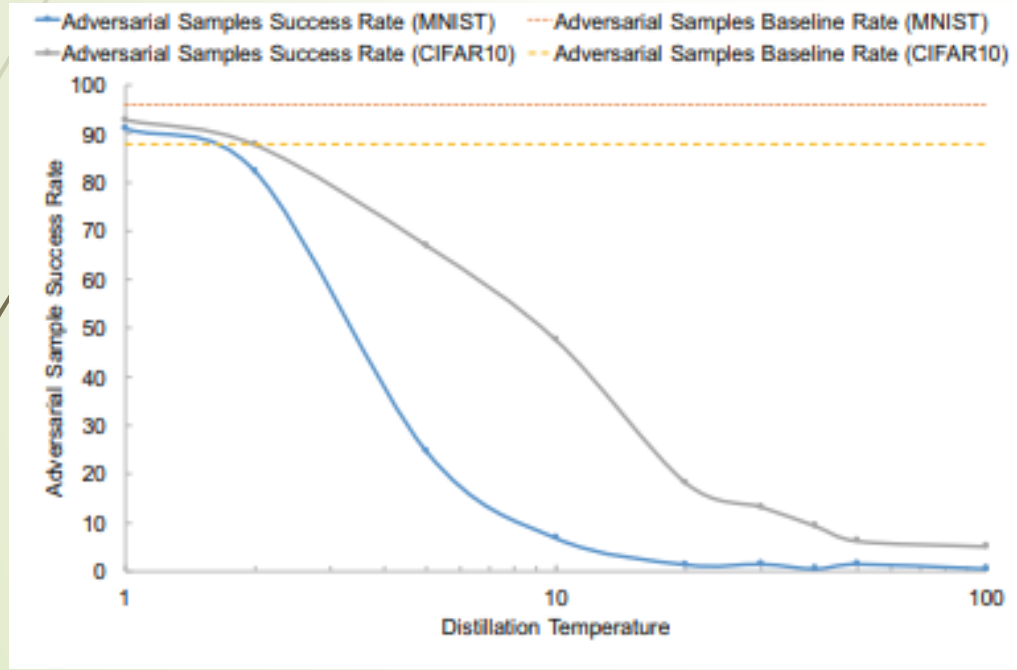
- دیتاست و شبکه:
  - MINST
  - CIFAR10
- شبکه ها:

Layer Type	MNIST Architecture	CIFAR10 Architecture
Relu Convolutional	32 filters (3x3)	64 filters (3x3)
Relu Convolutional	32 filters (3x3)	64 filters (3x3)
Max Pooling	2x2	2x2
Relu Convolutional	64 filters (3x3)	128 filters (3x3)
Relu Convolutional	64 filters (3x3)	128 filters (3x3)
Max Pooling	2x2	2x2
Relu Fully Connect.	200 units	256 units
Relu Fully Connect.	200 units	256 units
Softmax	10 units	10 units

Parameter	MNIST Architecture	CIFAR10 Architecture
Learning Rate	0.1	0.01 (decay 0.5)
Momentum	0.5	0.9 (decay 0.5)
Decay Delay	-	10 epochs
Dropout Rate (Fully Connected Layers)	0.5	0.5
Batch Size	128	128
Epochs	50	50

مثال چهارم: Distillation

- نتایج:



Distillation Temperature	MNIST Adversarial Samples Success Rate (%)	CIFAR10 Adversarial Samples Success Rate (%)
1	91	92.78
2	82.23	87.67
5	24.67	67
10	6.78	47.56
20	1.34	18.23
30	1.44	13.23
40	0.45	9.34
50	1.45	6.23
100	0.45	5.11
No distillation	95.89	87.89

نتیجه گیری و زمینه پژوهشی

- شیوه‌های متنوع دفاع
- کارایی هر یک با توجه به دیتاست‌های متفاوت
- بررسی و پیاده سازی هریک از روش‌ها
- یافتن معیار جامع و کلی برای مقایسه روش‌های مختلف

مراجع

1. Jan Hendrik Metzen, Tim Genewein, Volker Fischer, Bastian Bischoff, "On detecting adversarial perturbations", 2017
2. Uri Shaham, Yutaro Yamada, Sahand Negahban, " Increasing Local Stability of Neural Nets through Robust Optimization ", 2016
3. Yan Zhou, Murat Kantarcioglu, Bowei Xi, "A survey of game theoretic approach for adversarial machine learning", 2018
4. Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu† , Jun Zhu, "Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser", 2018
5. Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami, "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks", 2016
6. Dongyu Meng, Hao Chen, "a Two-Pronged Defense against Adversarial Examples", 2017
7. Negin Entezari, Saba A. Al-Sayouri, Amirali Darvishzadeh, Evangelos E. Papalexakis, "All You Need Is Low (Rank): Defending Against Adversarial Attacks on Graphs", 2020

سپاس از توجه شما