Countering Adversarial Images Using Input Transformations – ICLR (2018)

Chuan Guo (Cornell University) , Mayank Rana & Moustapha Cisse & Laurens van der Maaten (Facebook AI Research)

Presented by:

Shubham Koundinya

UNIVERSITY OF WATERLOO

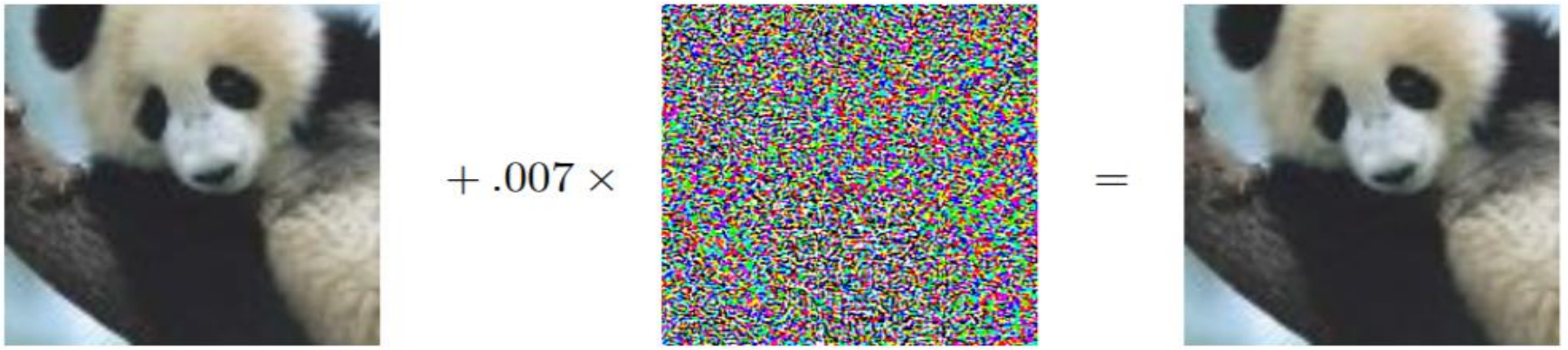
OUTLINE

- Motivation
- Terminology
- Previous Work
- Overview
- Problem Statement
- Adversarial Attacks
- Defenses /Image Transformations
- Experiments
- Conclusions
- Critiques

UNIVERSITY OF WATERLOO

MOTIVATION

- As , the use of Machine Learning has increased robustness has become a critical feature to guarantee the reliability of deployed Machine Learning Systems.
- However, Machine Learning systems have been shown to be fooled by small, carefully designed adversarial perturbations.



UNIVERSITY OF WATERLOO

Goodfellow et al [17]. By carefully adding small perturbation to the original image, GoogLeNet's classification is changed from Panda to Gibbon

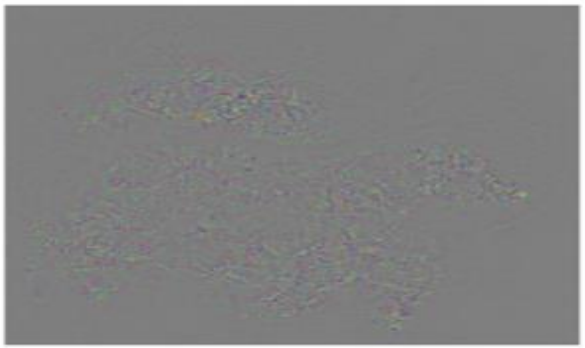
MOTIVATION

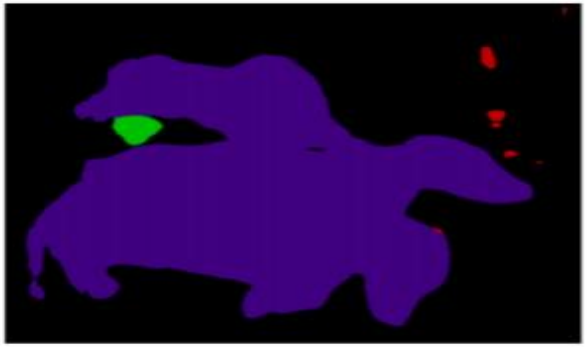
Adversarial Attacks not Limited to Image Classification Systems. Other domains e.g:

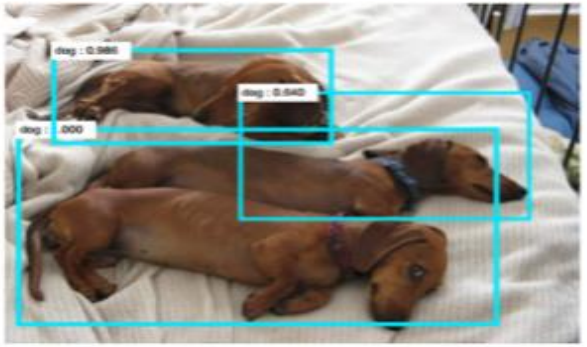
- Speech recognition systems (Cisse et al., 2017)[12]
- Robot vision (Melis et al., 2017)[13]
- Malware Classification (Grosse et. Al [22]
- Image Segmentation and Object Detection (Xie et al) [20]
- Image Captioning (Chen et al) [18].

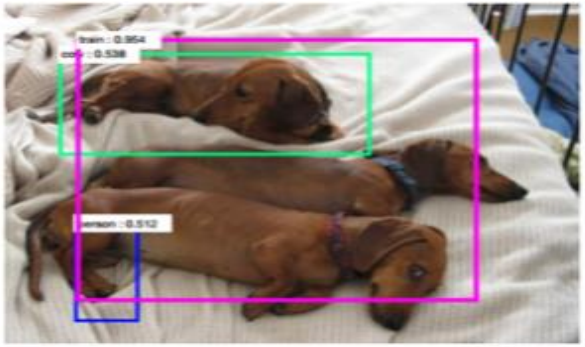
UNIVERSITY OF WATERLOO

MOTIVATION – Segmentation and Object Detection









Xie et al [20] Adversarial example for semantic segmentation and object detection.

Left Colum: Original Image (Top Row) , with Normal Segmentation (in middle) and Object detection in the bottom

Right Colum: After Adversarial Perturbation added (Top Row), Both segmentation (middle) and detection (bottom) , predictions are wrong.

MOTIVATION – Image Captioning

Original Top-3 inferred captions:

1. A red stop sign sitting on the side of a road.
2. A stop sign on the corner of a street.
3. A red stop sign sitting on the side of a street.

Adversarial Top-3 captions:

1. A brown teddy bear laying on top of a bed.
2. A brown teddy bear sitting on top of a bed.
3. A large brown teddy bear laying on top of a bed.

Original Top-3 inferred captions:

1. A man holding a tennis racquet on a tennis court.
2. A man holding a tennis racquet on top of a tennis court.
3. A man holding a tennis racquet on a court.

Adversarial Top-3 captions:

1. A woman brushing her teeth in a bathroom.
2. A woman brushing her teeth in the bathroom.
3. A woman brushing her teeth in front of a bathroom mirror.

- Chen et al [18]. Adversarial Attack When applied to Image Captioning Systems

Terminology

- **Adversarial Example** – Modified Version of Original Image that is intentionally perturbed.
- **Adversarial Perturbation** – Carefully added noise to the Clean Image, to fool the Classifier.
- **Black Box Attack** - Adversary does not have direct access to the Model.
- **Gray Box Attack** - Adversary has access to the model architecture and parameters , but is unaware of the defense strategy being used.
- **Non Targeted Adversarial Attack** – The goal of the attack is to modify a source image in a way such that the image will be classified incorrectly by the network
- **Targeted Adversarial Attack** - The goal of the attack is to modify a source image in way such that image will be classified as a target class by the network.
- **Defense** -A defense is a strategy that aims make the prediction on an adversarial example h(x') equal to the prediction on the corresponding clean example h(x) , where h() is the classifier.

UNIVERSITY OF WATERLOO

Previous Work

- Graese, et al. [3] - input transformation such as shifting, blurring and noise can render the majority of the adversarial examples as non-adversarial.
- Xu et al.[5] demonstrated, how feature squeezing methods, such as reducing the color bit depth of each pixel and spatial smoothing, defends against attacks.
- Dziugaite et al [6], studied the effect of JPG compression on adversarial images
- Adversarial Training
- Ensemble Adversarial Training [2]

Previous Work

- Adversarial Training – Increase robustness of model by injecting adversarial examples into Training Set.

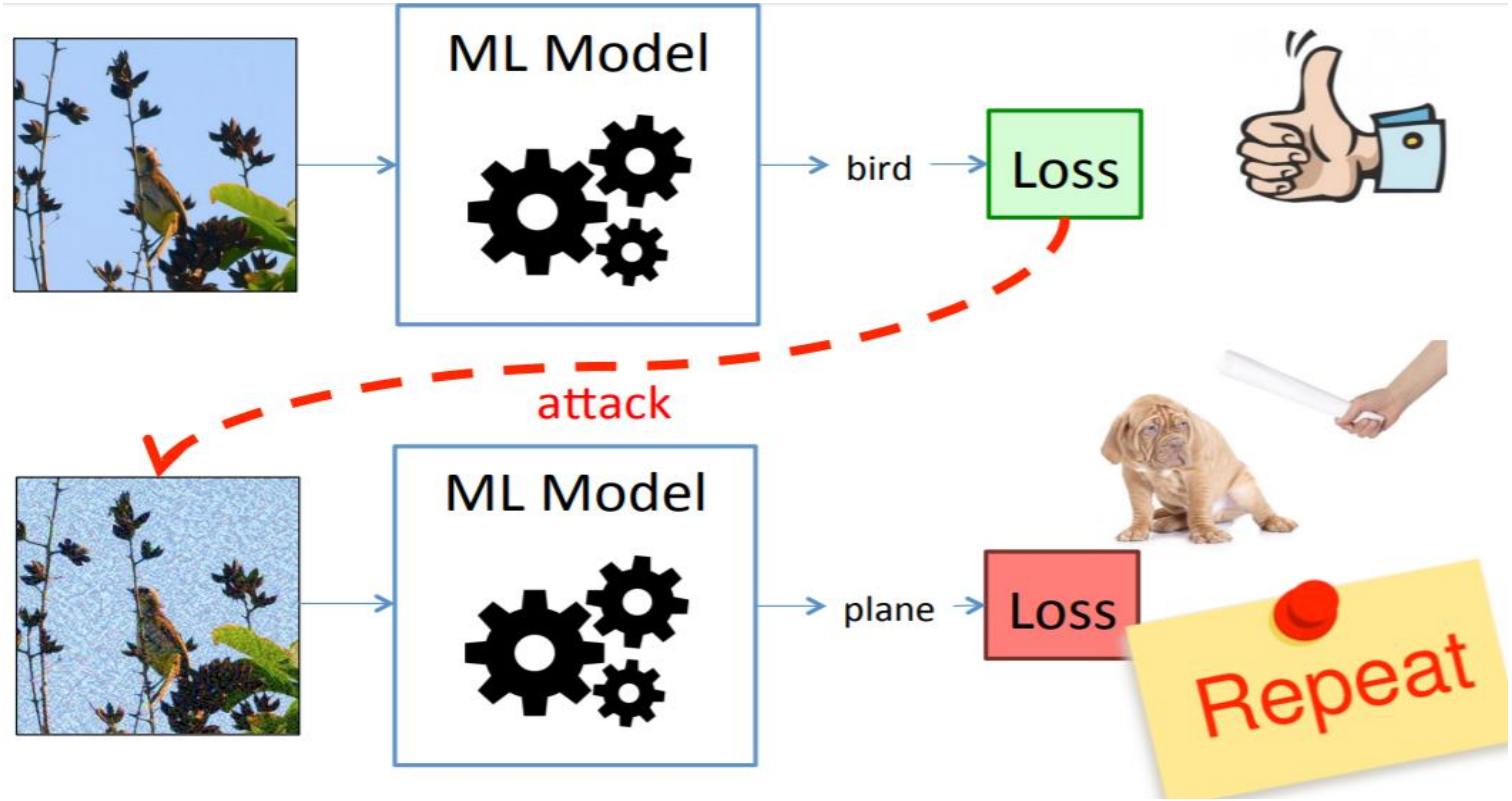


Image Credits: Florian Tramer[21]

Previous Work

- Ensemble Adversarial Training (Tramer et al)[2] – Augment training data with adversarial examples transferred from other models. Inception-Resnet-v2, trained on adversarial examples generated by FGSM against Inception-Resnet-v2 and Inception-v3 models

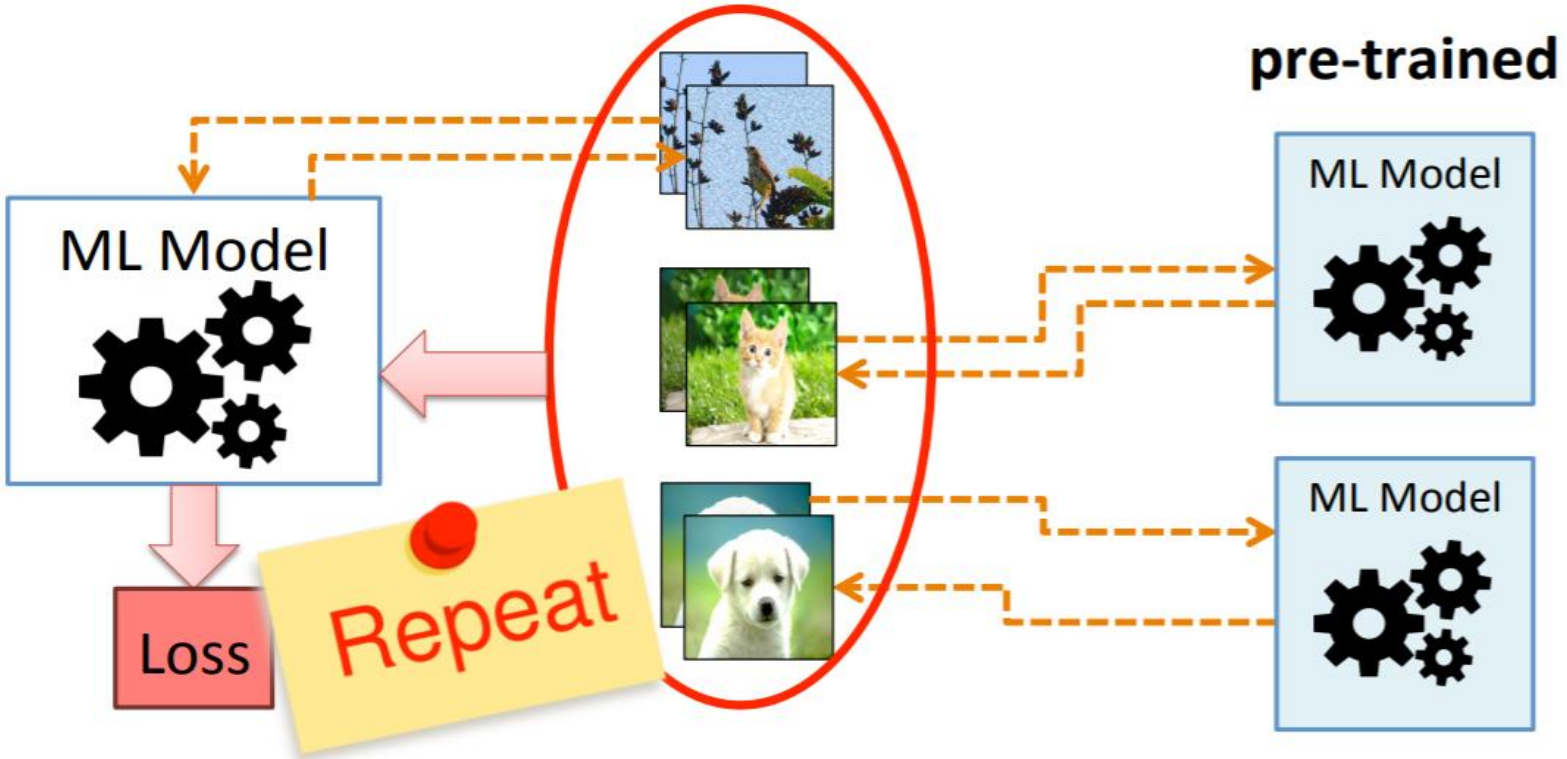


Image Credits: Florian Tramer[21]

Overview

- The paper studies strategies that defend against adversarial-example attacks on image-classification systems by transforming the images before feeding them to a Convolutional Network Classifier.

- **Following image transformations as a means for protecting against adversarial attacks have been studied** :

  1. Image Cropping and Re-scaling (Graese et al, 2016).
  2. Bit Depth Reduction (Xu et. al, 2017)
  3. JPEG Compression (Dziugaite et al, 2016)
  4. Total Variance Minimization (Rudin et al, 1992)
  5. Image Quilting (Efros & Freeman, 2001).

- **Image Transformations have been studied against Adversarial Attacks**:

  1. Fast Gradient Sign Method (FGSM) (Goodfelow et. al., 2015)
  2. Iterative FGSM (IFGSM)
  3. DeepFool
  4. Carlini-Wagner's L2 Attack

Problem Statement

- The paper studies defenses against non targeted adversarial examples for Image Recognition Systems.
- **Adversarial Example :** Given a classifier, h(.), a non targeted Adversarial example of x is x′ such that - $h(x) \neq h(x')$, and d(x, x′) ≤ p , for some dissimilarity d(.,.) commonly used d , Chebyshev or Euclidean.
- **Adversarial Attack :** From a set of N clean images, $[x_1, \ldots, x_n]$, an adversarial attack aims to generate $[x'1, \ldots, x'_n]$, such that x'_n is an adversary of x_n.
- **Success Rate of Attack:** Proportion of predictions altered by the attack.

$$\frac{1}{N}\sum_{n=1}^{N} \mathbb{1}[h(\mathbf{x}_n) \neq h(\mathbf{x}'_n)]$$

- **Notion of L2 Normalized L2 dissimilarity :**

$$\frac{1}{N}\sum_{n=1}^{N} \frac{\|\mathbf{x}_n - \mathbf{x}'_n\|_2}{\|\mathbf{x}_n\|_2}$$

- A strong adversarial attack has a high success rate whilst its normalized L2-dissimilarity is low.

Adversarial Attacks

Below four attacks have been considered in the paper:

- Fast Gradient Sign Method (FGSM; Goodfellow et al. (2015)) [17]
- Iterative Fast Gradient Sign Method (I-FGSM; Kurakin et al. (2016b))[14]
- DeepFool (Moosavi-Dezfooliet al., 2016) [15]
- Carlini Wagner's L2 attack [16]

Fast Gradient Sign Method (FGSM)

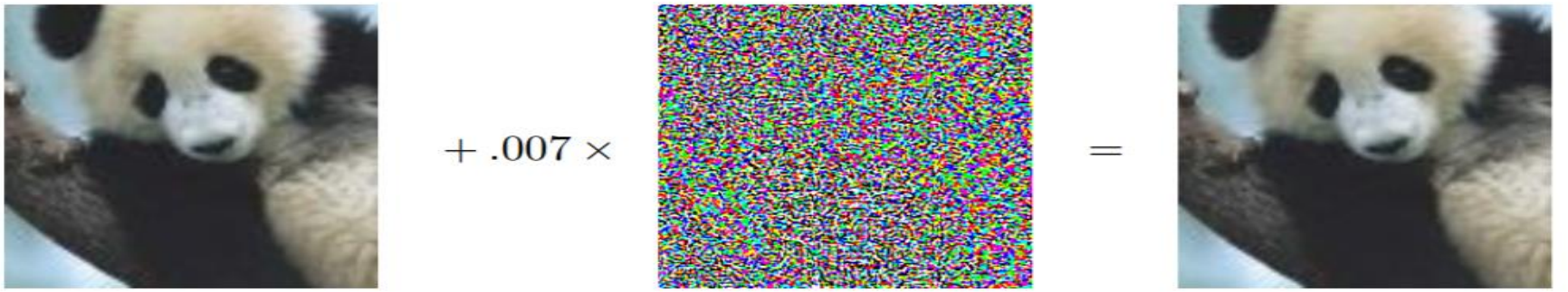
- Fast Gradient Sign Method (FGSM; Goodfellow et al. (2015)) [17]:

Source input x, True Label y , let l be the differentiable loss function, used to train the classifier $h(.)$, Then adversarial example x' is :

$$x' = x + \epsilon \cdot sign(\nabla_x l(x, y))$$

for some $\epsilon > 0$ which controls the perturbation magnitude.

- Keeping the parameters of the model constant, optimize parameters of image to increase the loss.



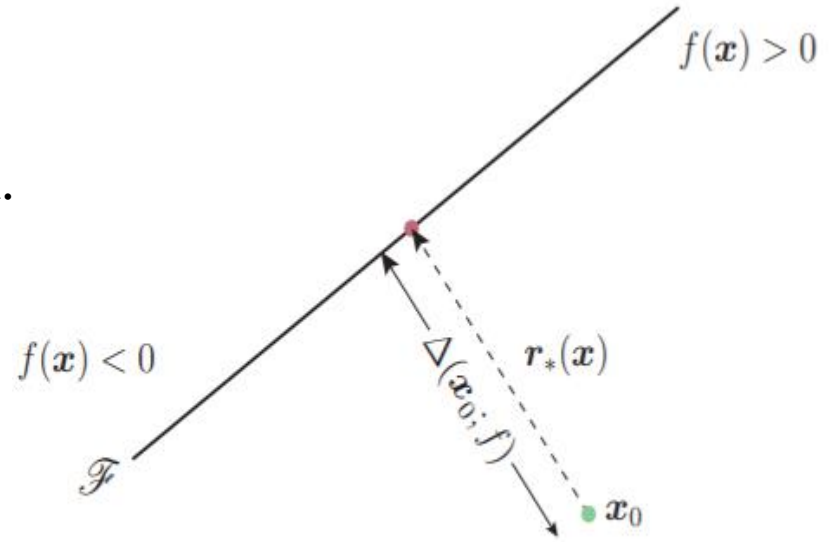
Iterative Fast Gradient Sign Method (IFGSM)

- Iterative Fast Gradient Sign Method (I-FGSM; Kurakin et al. (2016b))[14]

  Iteratively apply FGSM, for certain number of iterations.

DeepFool (Moosavi-Dezfooliet al., 2016) [15]

- Proved to be a stronger attack.
- Finds minimal perturbations needed to misclassify.
- Given a classifier f(x) and sample x_0.
- Project x_0, orthogonally on f(x).
- Keep adding r to x0,until it changes sign.



[15]Adversarial examples for a linear binary classifier.

Carlini Wagner's L2 attack [16]

- Optimization based attack,that directly optimizes for having the minimal distance from the original sample, under the constraint of having the example misclassified by the original problem.
- The untargeted variant of CW-L2 attack forms the following unconstrained optimization problem:

$$\min_{\mathbf{x}'} \left[\|\mathbf{x} - \mathbf{x}'\|_2^2 + \lambda_f \max\left(-\kappa, Z(\mathbf{x}')_{h(\mathbf{x})} - \max\{Z(\mathbf{x}')_k : k \neq h(\mathbf{x})\}\right)\right]$$

$Z(x)$be the operation that computes logit vector(output before the softmax layer)

$Z(x)_k$ = is the logit value corresponding to class k.

$\max\{Z(x')_k : k \neq h(x)\}$ = (Computes the second Largest Logit)

$Z(x')_{h(x)}$- $\max\{Z(x')_k : k \neq h(x)\}$ = Difference Between Largest & Second Largest Logit

K= Denotes a Margin Parameter

Image Transformations

1. Image Cropping and Re-scaling (Graese et al, 2016).
2. Bit Depth Reduction (Xu et. al, 2017)
3. JPEG Compression (Dziugaite et al, 2016)
4. Total Variance Minimization (Rudin et al, 1992)
5. Image Quilting (Efros & Freeman, 2001).

Image Cropping Rescaling

- Idea from (Graese et al, 2016) [3]- Assessing threat of adversarial examples of deep neural networks, where authors experimented on MNIST with FGSM attack.
- Intuition -Image Cropping Rescaling has the effect of altering the spatial positioning of adversarial perturbation, which is very important in making attacks successful.
- Rescale and Crop the images at training time as part of data augmentation and at test time average predictions.

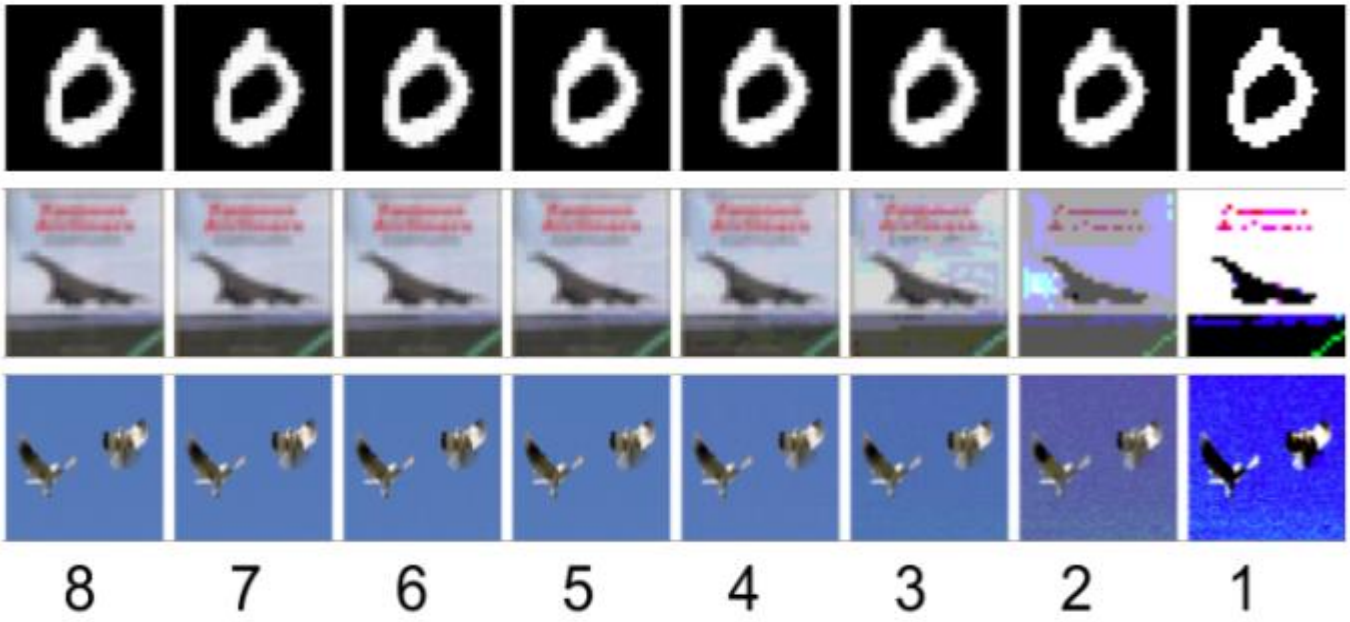
Bit Depth Reduction

- Idea from, (Xu et. al) [5] - Feature squeezing: Detecting adversarial examples in deep neural networks
- Bit Depth - Refers to the number of bits used to indicate the colors of a single pixel.

Gray Scale Image (8 bits per pixel) – $2^8 =$ 256 possible values for each pixel, where 0 is black, and 255 is white.

Colored Images (24 bits per pixel) – $2^{24} \approx$ 16 million possible values for each pixel.

Bit Depth Reduction



Xu et al [5]: Image examples with bit depth reduction. The first column shows images from MNIST, CIFAR-10 and ImageNet, respectively. Other columns show squeezed versions at different color-bit depths, ranging from 8 (original) to 1.

Bit Depth Reduction

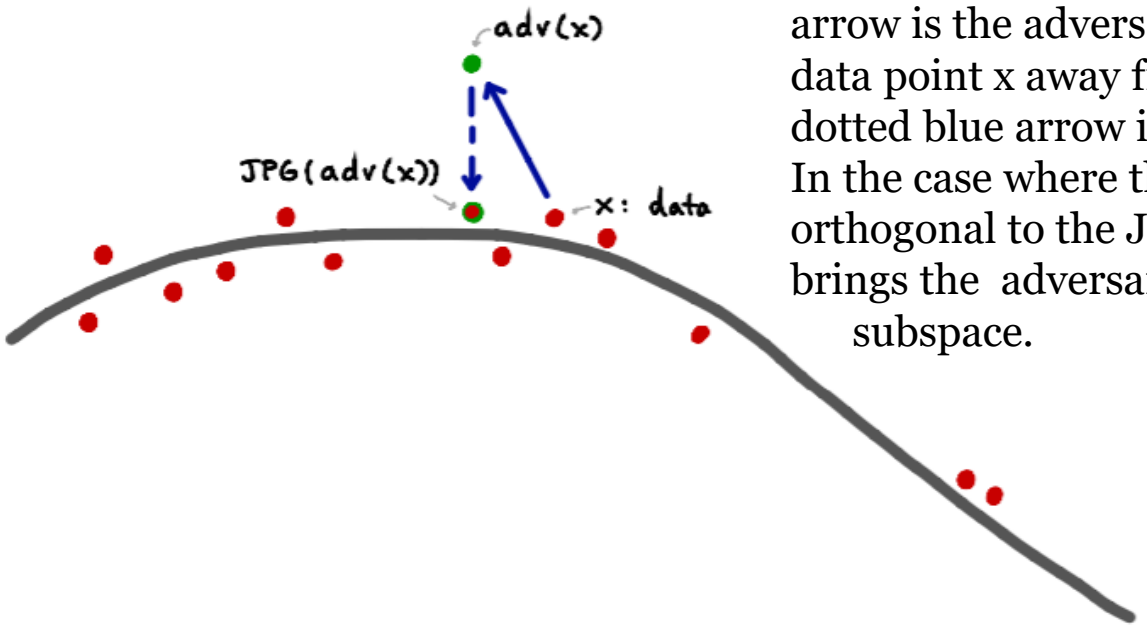
- Intuition : Feature i/p spaces are often unnecessary large and this vast input space provides extensive opportunities for an adversary to construct adversarial examples. Reduce the degrees of freedom available to an adversary by squeezing out unnecessary input features. Xu et al [5]. experimented with MNIST, CIFAR-10 and ImageNet.

- Authors follow the same strategy and reduce the bit depth to 3 , in their experiments.

JPEG Compression

- Idea from (Dziugaite et al.) [6]- A study of the effect of JPG compression on adversarial images. Authors Claim JPEG Compression works for small perturbations and not for large perturbations and they are not sure why JPEG Compression works for small perturbations.

- Hypothesis from above paper:



(Dziugaite et al.) [6]- The red dots represent the data and the grey line the data subspace. The solid blue arrow is the adversarial perturbation that moves the data point x away from the data subspace and the dotted blue arrow is the projection on the subspace. In the case where the perturbation is approximately orthogonal to the JPG subspace, JPG compression brings the adversarial example back to the data subspace.

Total Variance (TV) Minimization

- Total Variance (TV) Minimization – Algorithm from Leonid Rudin, Stanley Osher, and Emad Fatemi-Nonlinear total variation based noise removal algorithms.
- Used in Image Denoising / Signal Denoising.

Total Variance (TV) Minimization [9]

- Randomly select a set of pixels and reconstruct the "simplest", image that is consistent with selected pixels.
- Select a random set of pixels, by sampling a Bernoulli Random Variable $X(i,j,k)$ for each pixel location (i,j,k); maintain a pixel when $X(i,j,k)$=1.Use Total Variation Minimization to construct an image z that is similar to the perturbed input image x for selected set of pixels, where TV_p(z) represents the L_p the total variation of z.

$$\min_{\mathbf{z}} \|(1 - X) \odot (\mathbf{z} - \mathbf{x})\|_2 + \lambda_{\mathrm{TV}} \cdot \mathrm{TV}_p(\mathbf{z})$$

$$\mathrm{TV}_p(\mathbf{z}) = \sum_{k=1}^{K} \left[\sum_{i=2}^{N} \|\mathbf{z}(i,:,k) - \mathbf{z}(i-1,:,k)\|_p + \sum_{j=2}^{N} \|\mathbf{z}(:,j,k) - \mathbf{z}(:,j-1,k)\|_p \right]$$

Image Quilting (Efros & Freeman, 2001) [8].

- Image Quilting – Original idea from Efros & Freeman[8] , for texture synthesis & Texture Transfer.
- Texture Synthesis - While reconstructing , identify matching blocks for a given input block, and randomly pick one of these.

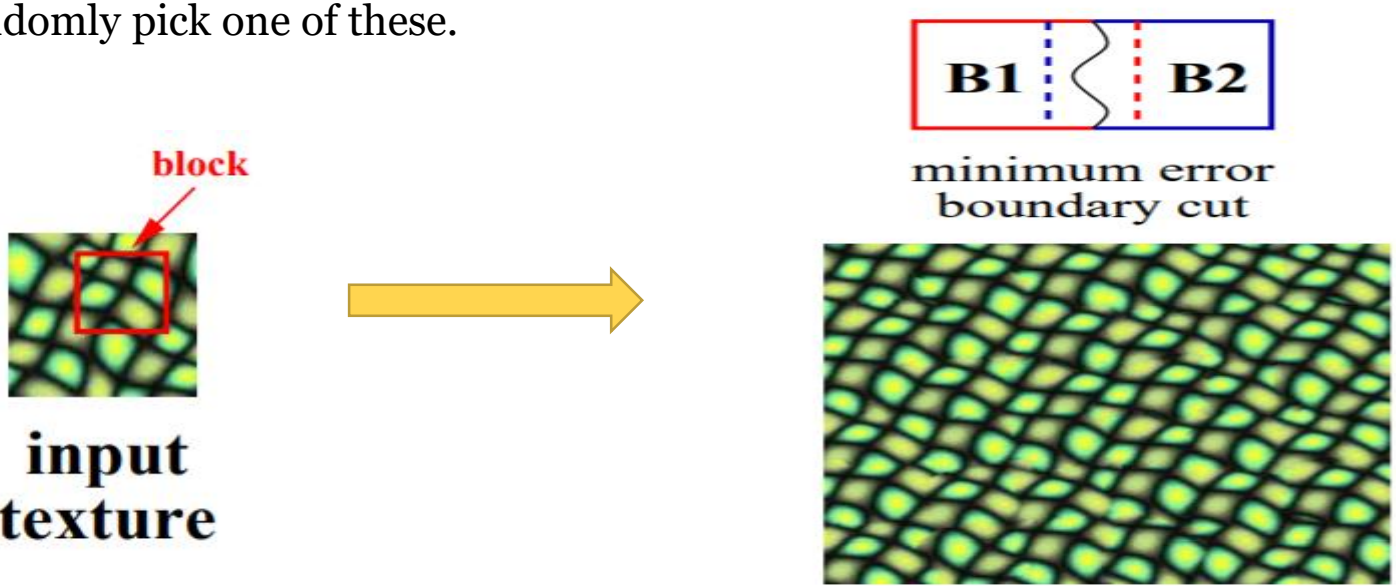


Fig. Square blocks from the input texture are patched together to synthesize a new texture sample. to reduce blockiness the boundary between blocks is computed as a minimum cost path through the error surface at the overlap

Image Quilting (Efros & Freeman, 2001) [8].

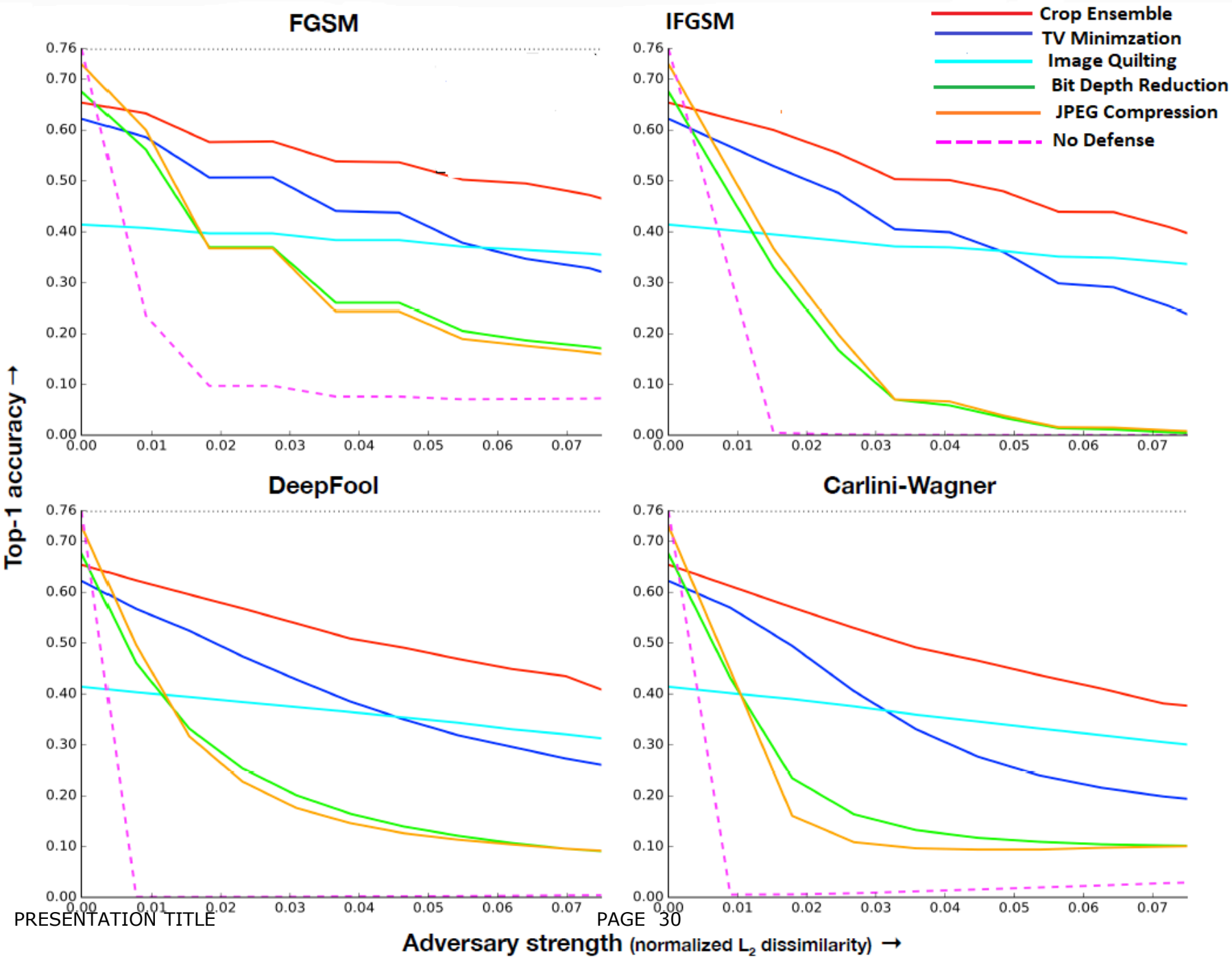
- For paper authors ,Created a database of 1,000,000 images that were selected randomly from ImageNet Training Set.
- Database of Images Contains Clean Images.
- The patches used to create synthesizes images are created by - Finding the K Nearest Neighbors(in pixel space) of corresponding patch from adversarial image in the patch database.
- Intuition of Defense: Resulting image only consists of pixels that were not modified by adversary - the database of real patches is unlikely to contain the structures that appear in adversarial images.

Experiments

- Setup:
  - ImageNet Dataset.
  - Adversarial Images are produced by ,attacking a ResNet 50 model with 4 attacks.
  - Strength of an adversary is measured in terms of its normalized L2-dissimilarity, and classification accuracies are reported as its function
- Five experiments were performed.
  - GrayBox- Image Transformation at Test Time
  - BlackBox - Image Transformation at Training and Test Time
  - Blackbox - Ensembling
  - GrayBox - Image Transformation at Training and Test Time
  - Comparison With Ensemble Adversarial Training

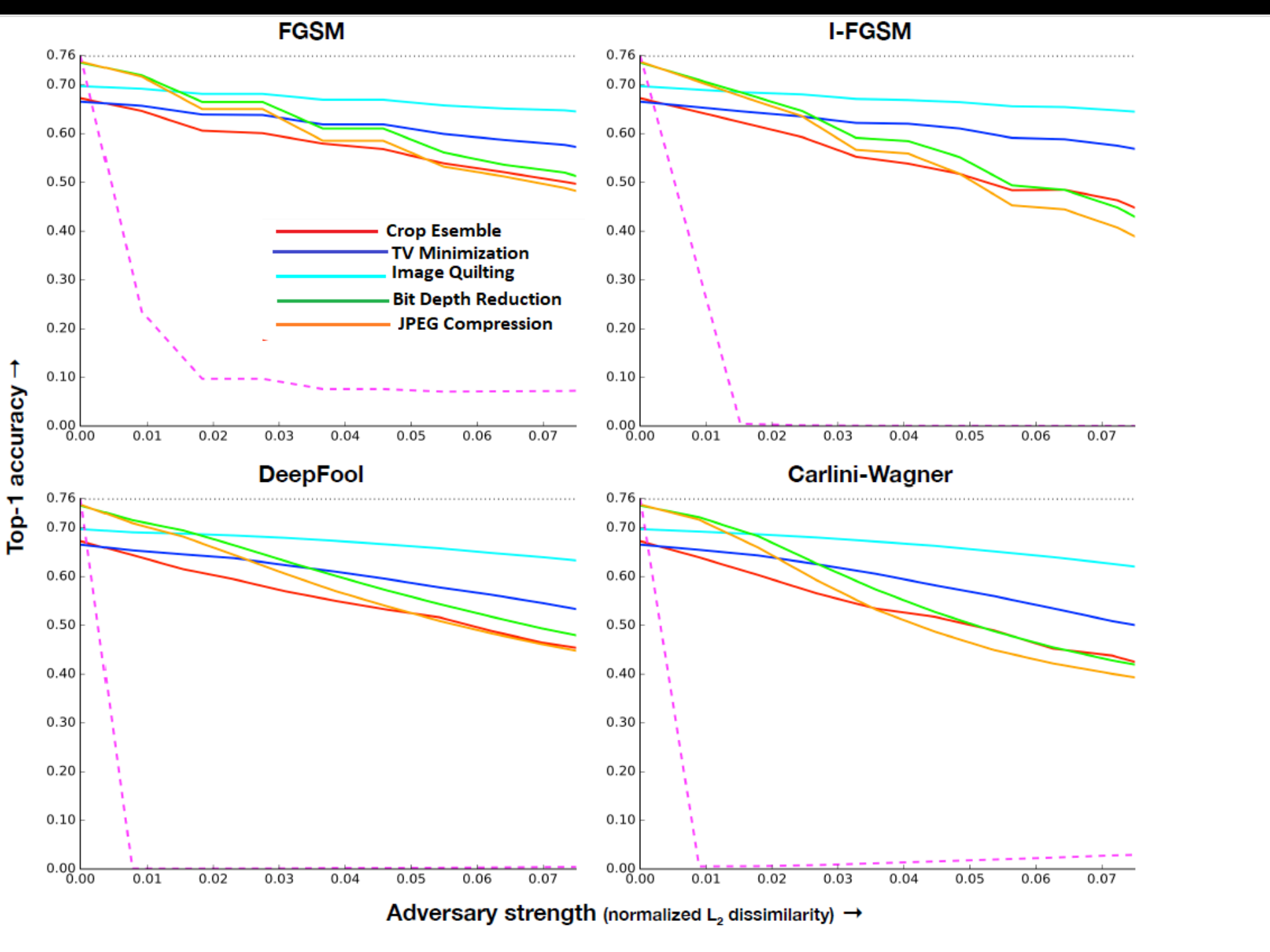
1. GrayBox- Image Transformation at Test Time

- Apply image transformations on adversarial images before feeding them to a ResNet-50 classifier trained to classify Clean images. Few of important results:
  - Proposed Transformations partly eliminate the effects of the attack.
  - Ensembling 30 predictions over different random image crops is very efficient.
  - TV Minimization and Image Quilting , can successfully removes adversarial perturbations
  - Fig on next Slide provides the result in this setting


FGSM
IFGSM
DeepFool
Carlini-Wagner
Crop Ensemble
TV Minimzation
Image Quilting
Bit Depth Reduction
JPEG Compression
No Defense
Top-1 accuracy →
Adversary strength (normalized L_2 dissimilarity) →
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.76
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07

2. BlackBox: Image Transformation at Training and Test Time

- ResNet-50 model trained on transformed (5 image transformation techniques) ImageNet Training Images.
- The same adversarial images from previous experiment were used, which concludes it's a blackbox setting, as adversary cannot use the above model to generate new adversarial images.
- At test time , apply transformations.
- Image Quilting provides promising results – as it successfully defends against 80-90 % of the attacks.
- Fig on next Slide provides the result in this setting


FGSM
I-FGSM
DeepFool
Carlini-Wagner
Crop Esemble
TV Minimization
Image Quilting
Bit Depth Reduction
JPEG Compression
Top-1 accuracy →
Adversary strength (normalized L_2 dissimilarity) →
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.76
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07

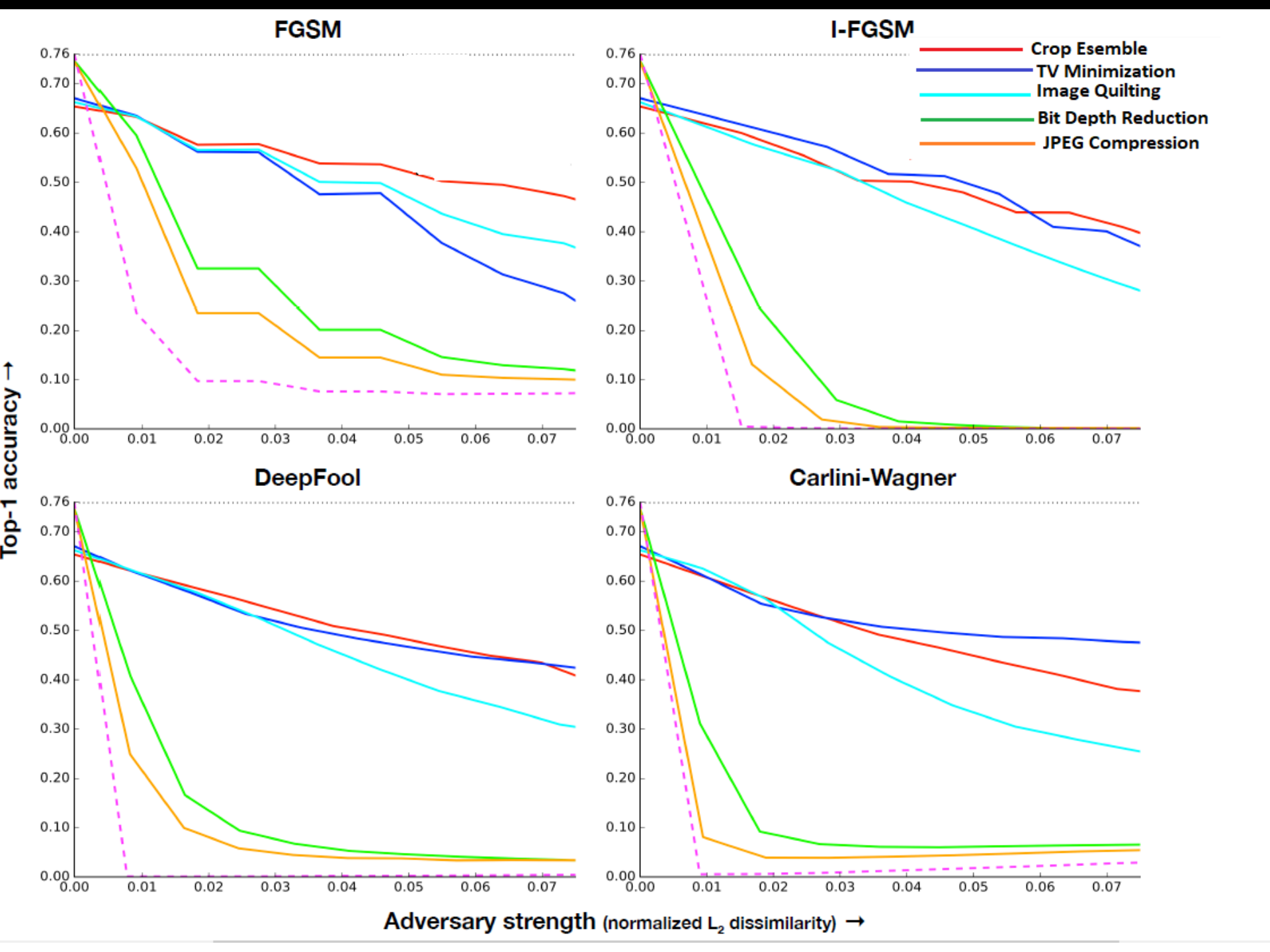
3. BlackBox: Ensembling

- Ensemble Image Transformations and study the attacks on Different Models. Below table highlights different settings. Few of the important results:
  - Adversarial Images generated by attacking ResNet-50.
  - Four architectures were studied- ResNet-50, ResNet-101, DenseNet-169, and Inception-v4.
  - Clean Images Accuracy is 76 %.
  - Attacks can deteriorate the accuracy of the best ensemble by atmost 6 %.

	Quilting				TVM + Quilting				Cropping + TVM + Quilting			
	RN50	RN101	DN169	Iv4	RN50	RN101	DN169	Iv4	RN50	RN101	DN169	Iv4
No Attack	70.07	72.56	70.18	73.01	72.38	74.74	73.10	**75.55**	72.14	74.53	72.92	75.10
FGSM	65.45	68.50	65.96	67.53	65.70	68.77	67.09	69.19	66.65	69.75	67.86	**70.37**
I-FGSM	65.59	68.72	66.16	69.29	65.84	69.10	67.32	71.05	67.03	70.14	68.20	**71.52**
DeepFool	65.20	68.73	65.86	68.70	65.80	69.34	67.40	71.03	67.11	70.49	68.62	**71.47**
CW-L2	64.11	67.72	65.00	68.14	63.99	68.20	66.08	70.13	65.31	69.14	66.96	**70.50**

4. GrayBox: Image Transformation at Training and Test Time

- Similar to experiment 2, but Adversary has access to model to generate Adversarial examples.
- ResNet-50 model trained on transformed (from 5 image transformation techniques) ImageNet Training Images.
- Adversary has access to above trained model to generate Adversarial Examples.
- At test time, apply Five Image Transformation Techniques. Results:
  - Cropping, TV Minimization and Image Quilting defenses classify upto 50 % of images correctly.
  - Fig on next Slide provides the result in this setting


FGSM
I-FGSM
DeepFool
Carlini-Wagner
Crop Esemble
TV Minimization
Image Quilting
Bit Depth Reduction
JPEG Compression
Top-1 accuracy →
Adversary strength (normalized L_2 dissimilarity) →
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.76
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07

5. Comparison With Ensemble Adversarial Training

- The results of the experiment are compared with the state of the art ensemble adversarial training approach proposed by Tramer et al. [2]
- The results show that ensemble adversarial training works better on FGSM attacks (which it uses at training time), but is outperformed by each of the transformation-based defenses all other attacks.

	Cropping	TVM	Quilting	Ensemble Training (Tramèr et al., 2017)
No Attack	65.41	66.29	69.66	**80.3**
FGSM	49.52	31.37	39.55	**69.15**
I-FGSM	**43.89**	40.99	33.22	5.07
DeepFool	**44.92**	44.69	34.54	1.84
CW-L2	41.06	**48.41**	30.51	22.23

Conclusions

- The paper proposed reasonable approaches to countering adversarial images.
- The authors evaluated Total Variance Minimization and Image Quilting and compared it with already proposed ideas like Image Cropping- Rescaling, Bit Depth Reduction, JPEG Compression on the challenging ImageNet dataset
- Future work suggests applying the same techniques to other domains such as speech recognition and image segmentation
- The input transformations can also be studied with ensemble adversarial training by Tramèr et al.[2]

Critiques

- The terminology of Black Box, White Box, and Grey Box attack is not exactly given and sometimes confusing.
- What if Adversary has knowledge about input transformations ?
- Though the authors did a considerable work in showing the effect of four attacks on ImageNet database, much stronger attacks (Madry et al) [7], could have been evaluated.
- Authors claim that the success rate is generally measured as a function of the magnitude of perturbations, performed by the attack using the L2- dissimilarity, but the claim is not supported by any references. None of the previous work has used these metrics.

References

- 1. Chuan Guo , Mayank Rana & Moustapha Ciss´e & Laurens van der Maaten , Countering Adversarial Images Using Input Transformations
- 2. Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, Patrick McDaniel, Ensemble Adversarial Training: Attacks and defenses.
- 3. Abigail Graese, Andras Rozsa, and Terrance E. Boult. Assessing threat of adversarial examples of deep neural networks. CoRR, abs/1610.04256, 2016.
- 4. Qinglong Wang, Wenbo Guo, Kaixuan Zhang, Alexander G. Ororbia II, Xinyu Xing, C. Lee Giles, and Xue Liu. Adversary resistant deep neural networks with an application to malware detection. CoRR, abs/1610.01239, 2016a.
- 5.Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. CoRR, abs/1704.01155, 2017.
- 6. Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel Roy. A study of the effect of JPG compression on adversarial images. CoRR, abs/1608.00853, 2016.
- 7.Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu .Towards Deep Learning Models Resistant to Adversarial Attacks, arXiv:1706.06083v3
- 8.Alexei Efros and William Freeman. Image quilting for texture synthesis and transfer. In Proc. SIGGRAPH, pp. 341–346, 2001.

References

- 9.Leonid Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. Physica D, 60:259–268, 1992.
- 10.Qinglong Wang, Wenbo Guo, Kaixuan Zhang, Alexander G. Ororbia II, Xinyu Xing, C. Lee Giles, and Xue Liu. Learning adversary-resistant deep neural networks. CoRR, abs/1612.01401, 2016b.
- 11. Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. CoRR, abs/1611.02770, 2016.
- 12. Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured prediction models. CoRR, abs/1707.05373, 2017
- 13. Marco Melis, Ambra Demontis, Battista Biggio, Gavin Brown, Giorgio Fumera, and Fabio Roli. Is deep learning safe for robot vision? adversarial examples against the icub humanoid. CoRR,abs/1708.06939, 2017.
- 14. Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. CoRR, abs/1607.02533, 2016b.
- 15. Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In Proc. CVPR, pp. 2574–2582, 2016.
- 16. Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In IEEE Symposium on Security and Privacy, pp. 39–57, 2017.

References

- 17. Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Proc. ICLR, 2015
- 18. Hongge Chen,Huan Zhang, Pin-Yu Chen,Jinfeng Yi and Cho-Jui Hsieh , Attacking Visual Language Grounding with Adversarial Examples: A Case Study on Neural Image Captioning
- 19. Naveed Akhtar and Ajmal Mian : Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey
- 20. C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, Adversarial Examples for Semantic Segmentation and Object Detection, arXiv preprint arXiv:1703.08603, 2017
- 21.http://floriantramer.com/docs/slides/cyberbest17ensemble.pdf
- 22. Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, Patrick McDaniel - Adversarial Perturbations Against Deep Neural Networks for Malware Classification

THANK YOU

Questions ?