Adversarial Robustness

CSIT375/975 AI and Cybersecurity

Dr Wei Zong

SCIT University of Wollongong

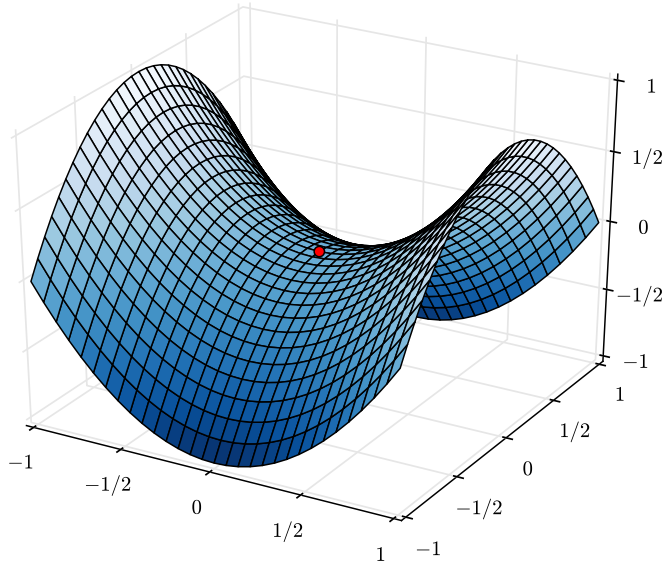
Disclaimer: The presentation materials come from various sources. For further information, check the references section

Outline

- Adversarial training
- Input transformations as defence
- Detect adversarial examples
- Reflections on adversarial examples

Adversarial Training

- Key question:
  - How can we train deep neural networks that are robust to adversarial inputs?
- An optimization view on adversarial robustness
  - Saddle point
    - A saddle point is any location where all gradients of a function vanish, but which is neither a global nor a local minimum
    - A saddle point (in red) on the graph of $z = x^2 - y^2$.
    - It looks like a saddle, which is where this mathematical property got its name.



Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

Adversarial Training

- Adversarial robustness as a saddle point problem:

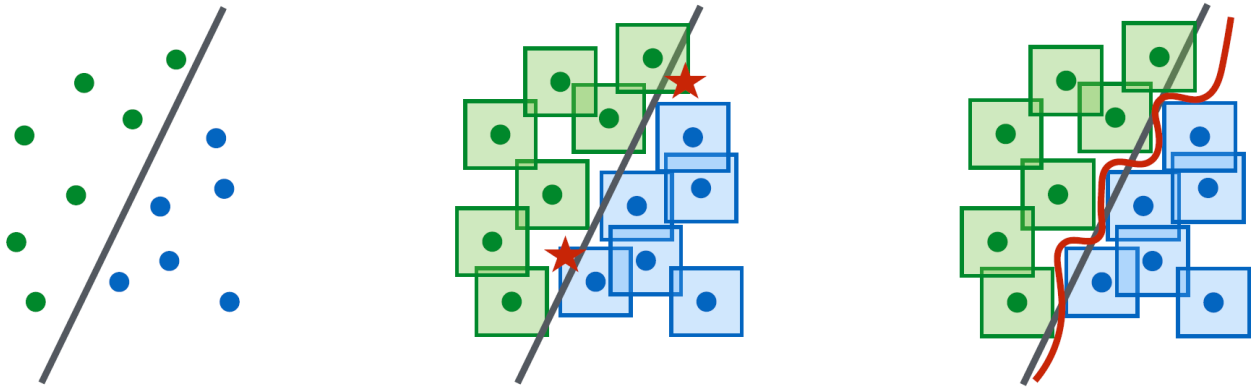
$$minimize_{\theta}\ E_{(x,y)\sim D}[maximize_{\delta \in S}\ \ell(\theta,\ x+\delta, y)]$$

  - Formulations of this type have a long history in robust optimization, going back to 1945.
  - Viewing the saddle point problem as the composition of an inner maximization problem and an outer minimization problem.
    - The inner maximization problem aims to find an adversarial version of a given data point x that achieves a high loss.
      - This is precisely the problem of attacking a given neural network.
    - The outer minimization problem aims to find model parameters so that the "adversarial loss" given by the inner attack problem is minimized.
      - This is precisely the problem of training a robust classifier using adversarial training techniques.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

Adversarial Training

- Classifying examples in a robust way requires a stronger classifier



- A conceptual illustration of standard vs. adversarial decision boundaries.
  - Left: a set of points that can be easily separated with a simple (in this case, linear) decision boundary.
  - Middle: the simple decision boundary does not separate the L_∞-balls (here, squares) around the data points.
    - Hence there are adversarial examples (the red stars) that will be misclassified.
  - Right: separating the L_∞-balls requires a significantly more complicated decision boundary.
    - The resulting classifier is robust to adversarial examples with bounded L_∞-norm perturbations.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

Adversarial Training

- Adversarial training is a simple but effective defense.
  - The adversary of choice is **projected gradient descent (PGD)**.
  - Adversarial training with PGD achieves the best empirical results

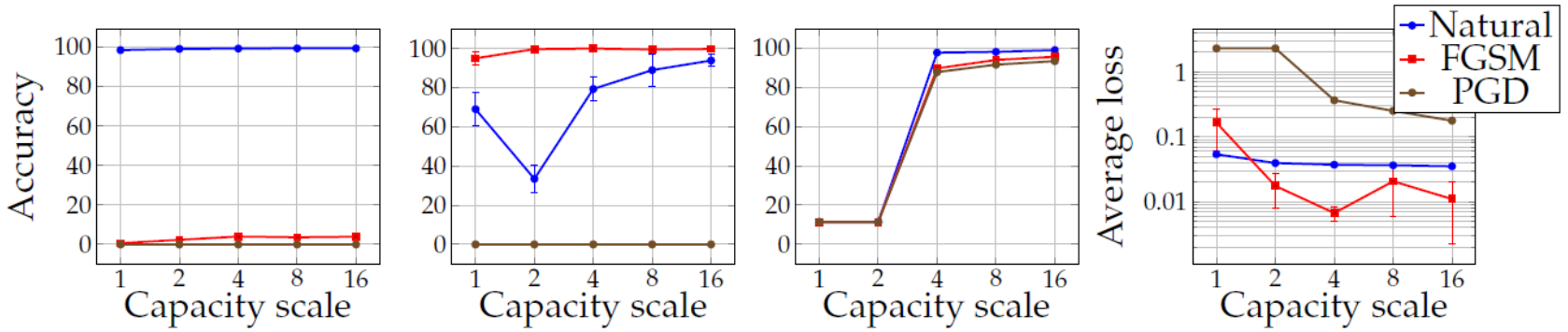
- Example Python code of adversarial training:

```
for batch_idx, (inputs, targets) in enumerate(train_loader):
    inputs, targets = inputs.to(device), targets.to(device)
    optimizer.zero_grad()

    adv = PGD_attack(inputs, targets)
    adv_outputs = net(adv)
    loss = criterion(adv_outputs, targets)

    loss.backward()
    optimizer.step()
```


Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

Adversarial Training

MNIST



CIFAR10

	Simple	Wide
Natural	92.7%	95.2%
FGSM	27.5%	32.7%
PGD	0.8%	3.5%

(a) Standard training

	Simple	Wide
Natural	87.4%	90.3%
FGSM	90.9%	95.1%
PGD	0.0%	0.0%

(b) FGSM training

	Simple	Wide
Natural	79.4%	87.3%
FGSM	51.7%	56.1%
PGD	43.7%	45.8%

(c) PGD training

	Simple	Wide
Natural	0.00357	0.00371
FGSM	0.0115	0.00557
PGD	1.11	0.0218

(d) Training Loss

- The effect of network capacity on the performance of the network.
  - MNIST and CIFAR10 networks are trained with varying capacity on: (a) natural examples, (b) with FGSM-made adversarial examples, (c) with PGD-made adversarial examples.
  - The first three plots/tables of each dataset show how the standard and adversarial accuracy changes with respect to capacity for each training regime.
  - The final plot/table show the value of the cross-entropy loss on the adversarial examples the networks were trained on.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

Adversarial Training

- Performance of the adversarially trained network against different attacks
  - White-box attacks with PGD for a different number of iterations and restarts, denoted by source **A**.
  - Black-box attacks from an independently trained copy of the network, denoted **A'**.
  - Black-box attacks from a version of the same network trained only on natural examples, denoted $\mathbf{A}_{nat}$.
  - Black-box attacks from a different convolution architecture, denoted **B**.

- The most successful attacks are in **bold**.
  - Adversarial training lowers model accuracy on natural (clean) data.
  - Robustness of more complex models is more challenging.

CIFAR10

Method	Steps	Source	Accuracy
Natural	-	-	87.3%
FGSM	-	A	56.1%
PGD	7	A	50.0%
PGD	20	A	**45.8%**
CW	30	A	46.8%
FGSM	-	A′	67.0%
PGD	7	A′	**64.2%**
CW	30	A′	78.7%
FGSM	-	A_{nat}	85.6%
PGD	7	A_{nat}	86.0%

MNIST

Method	Steps	Restarts	Source	Accuracy
Natural	-	-	-	98.8%
FGSM	-	-	A	95.6%
PGD	40	1	A	93.2%
PGD	100	1	A	91.8%
PGD	40	20	A	90.4%
PGD	100	20	A	**89.3%**
Targeted	40	1	A	92.7%
CW	40	1	A	94.0%
CW+	40	1	A	93.9%
FGSM	-	-	A′	96.8%
PGD	40	1	A′	96.0%
PGD	100	20	A′	**95.7%**
CW	40	1	A′	97.0%
CW+	40	1	A′	96.4%
FGSM	-	-	B	**95.4%**
PGD	40	1	B	96.4%
CW+	-	-	B	95.7%

Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

Adversarial Training

- Limitation: the naïve adversarial training is slow

```
for batch_idx, (inputs, targets) in enumerate(train_loader):
    inputs, targets = inputs.to(device), targets.to(device)
    optimizer.zero_grad()

    adv = PGD_attack(inputs, targets)
    adv_outputs = net(adv)
    loss = criterion(adv_outputs, targets)

    loss.backward()
    optimizer.step()
```

Backward propagation needs to run multiple times to generate AEs.

Backward propagation must be called at least once to optimize parameters. Can we exploit this call to generate AEs?

Shafahi, A., Najibi, M., Ghiasi, M.A., Xu, Z., Dickerson, J., Studer, C., Davis, L.S., Taylor, G. and Goldstein, T., 2019. Adversarial training for free!. Advances in neural information processing systems, 32.

Adversarial Training

- Achieve adversarial robustness for free
  - Key idea: the gradient of the loss with respect to the input is computed on the same backward pass to train parameters.

Algorithm 1 "Free" Adversarial Training (Free-m)

Require: Training samples X, perturbation bound ϵ, learning rate τ, hop steps m

Initialize θ
$\delta \leftarrow 0$
for epoch $= 1 \ldots N_{ep}/m$ **do**
  for minibatch $B \subset X$ **do**
    for i $= 1 \ldots m$ **do**
      Update θ with stochastic gradient descent
        $g_\theta \leftarrow \mathbb{E}_{(x,y)\in B}[\nabla_\theta\, l(x+\delta, y, \theta)]$
        $g_{adv} \leftarrow \nabla_x\, l(x+\delta, y, \theta)]$
        $\theta \leftarrow \theta - \tau g_\theta$
      Use gradients calculated for the minimization step to update δ
        $\delta \leftarrow \delta + \epsilon \cdot \text{sign}(g_{adv})$
        $\delta \leftarrow \text{clip}(\delta, -\epsilon, \epsilon)$
    end for
  end for
end for

Keep the same number of iterations.

Mini-batch replay enables multiple adversarial updates like PGD

Train parameters

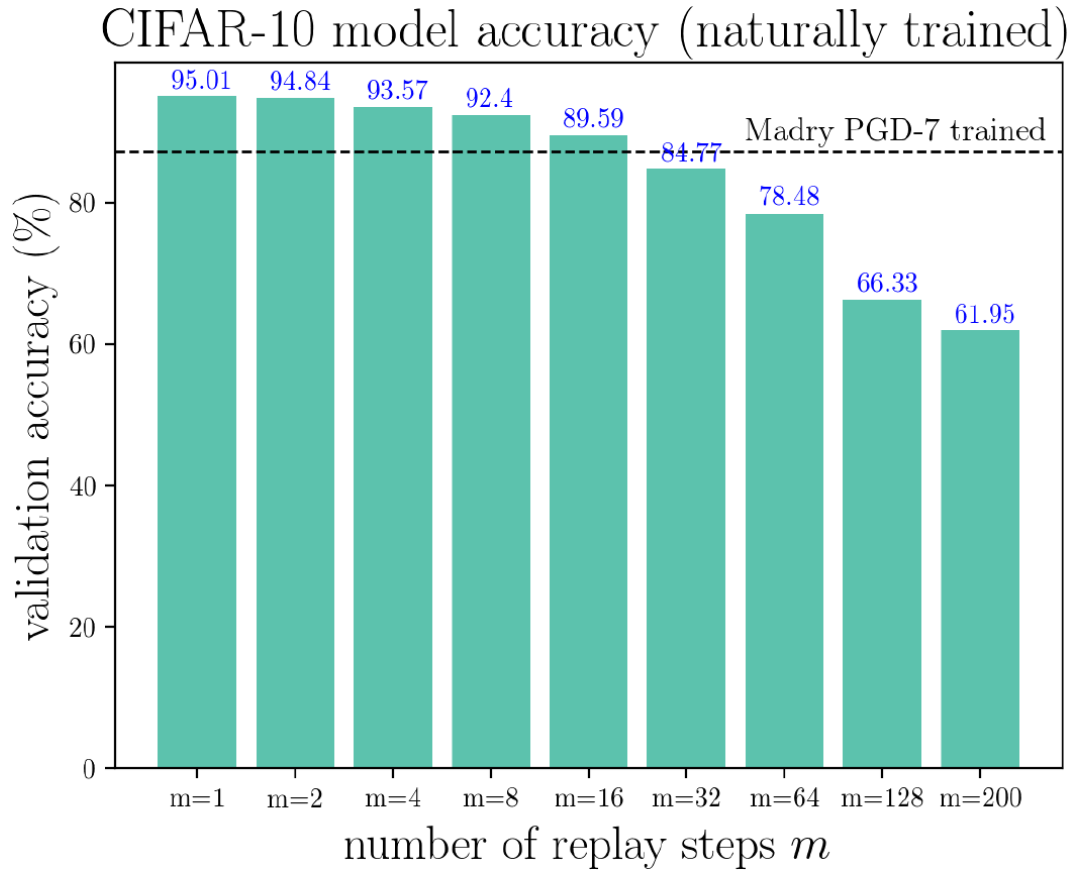
Get all the gradients by Calling loss.backward() only once

No need to call loss.backward() again to generate AEs

Shafahi, A., Najibi, M., Ghiasi, M.A., Xu, Z., Dickerson, J., Studer, C., Davis, L.S., Taylor, G. and Goldstein, T., 2019. Adversarial training for free!. Advances in neural information processing systems, 32.

Adversarial Training

- Achieve adversarial robustness for free
  - Mini-batch replay can hurt performance
    - Models may “forget” learned knowledge as it takes more iterations to see the same input.
    - If m = total number of epochs, the same input will be seen again after mini-batch replay.



Shafahi, A., Najibi, M., Ghiasi, M.A., Xu, Z., Dickerson, J., Studer, C., Davis, L.S., Taylor, G. and Goldstein, T., 2019. Adversarial training for free!. Advances in neural information processing systems, 32.

Adversarial Training

- Achieve adversarial robustness for free
  - Validation accuracy and robustness of CIFAR-10 models trained with various methods.

Training	Evaluated Against					Train Time (min)
	Nat. Images	PGD-20	PGD-100	CW-100	10 restart PGD-20	
Natural	**95.01%**	0.00%	0.00%	0.00%	0.00%	**780**
Free $m = 2$	91.45%	33.92%	33.20%	34.57%	33.41%	816
Free $m = 4$	87.83%	41.15%	40.35%	41.96%	40.73%	800
Free $m = 8$	85.96%	**46.82%**	**46.19%**	**46.60%**	**46.33%**	785
Free $m = 10$	83.94%	46.31%	45.79%	45.86%	45.94%	785
7-PGD trained	87.25%	45.84%	45.29%	46.52%	45.53%	5418

  - Achieve similar adversarial robustness compared to naïve adversarial training with 7-step PGD.
  - Negligibly increase in training time.
    - Naïve adversarial training is almost 7 times slower.

Shafahi, A., Najibi, M., Ghiasi, M.A., Xu, Z., Dickerson, J., Studer, C., Davis, L.S., Taylor, G. and Goldstein, T., 2019. Adversarial training for free!. Advances in neural information processing systems, 32.

Adversarial Training

Model	Def.	FGSM	C&W.	stAdv
A	Adv.	4.3%	4.6%	**32.62%**
	Ens.	1.6%	4.2%	**48.07%**
	PGD	4.4%	2.96%	**48.38%**
B	Adv.	6.0%	4.5%	**50.17%**
	Ens.	2.7%	3.18%	**46.14%**
	PGD	9.0%	3.0%	**49.82%**
C	Adv.	3.22%	0.86%	**30.44%**
	Ens.	1.45%	0.98%	**28.82%**
	PGD	2.1%	0.98%	**28.13%**

Model	Def.	FGSM	C&W.	stAdv
ResNet32	Adv.	13.10%	11.9%	**43.36%**
	Ens.	10.00%	10.3%	**36.89%**
	PGD	22.8%	21.4%	**49.19%**
wide ResNet34	Adv.	5.04%	7.61%	**31.66%**
	Ens.	4.65%	8.43%	**29.56%**
	PGD	14.9%	13.90%	**31.6%**

- Limitation: not robust against **novel attacks**.
  - Spatially transformed adversarial examples (stAdv) does not constrain perturbations using p-norm.
  - The table above shows the attack success rate of adversarial examples generated by stAdv against 3 different models under adversarial training on MNIST, and against ResNet32 and wide ResNet34 on CIFAR-10.
  - We observe that the defense can achieve high performance (less than 10% attack success rate) against FGSM and C&W attacks
  - These defense methods only achieve low defense performance on stAdv, which improve the attack success rate to more than 30%.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

Input Transformations as Defence

- Idea
  - Transform the inputs before feeding them to a model .
    - This is done at both training and testing time.
    - Adversarial perturbations have a particular structure to fool the target model.
      - The structure is calculated via optimization.
    - The alterations undo the effects of the adversarial attack.

- Goal
  - Remove the adversarial perturbations from input images.
  - Maintain sufficient information in input images to correctly classify them.

Guo, C., Rana, M., Cisse, M. and Van Der Maaten, L., 2017. Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117.

Input Transformations as Defence

- Transformations
  - Ensemble cropping-rescaling
    - Average predictions over random image crops, e.g., 30 crops.
  - Bit-depth reduction
    - Perform a simple type of quantization that can removes small (adversarial) variations in pixel values from an image;
    - Reduce images to 3 bits in experiments.
  - JPEG compression
    - Perform compression at quality level 75 (out of 100)

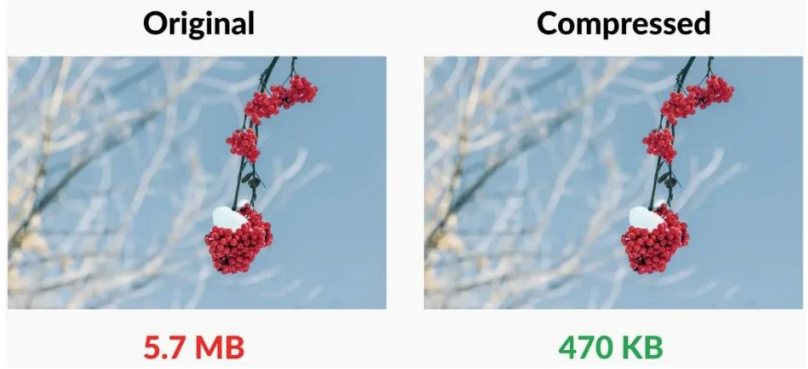
Cropping-rescaling

https://pytorch.org/vision/master/auto_examples/transforms/plot_transforms_illustrations.html#sphx-glr-auto-examples-transforms-plot-transforms-illustrations-py

Bit-depth Reduction

https://blenderartists.org/t/reducing-the-number-of-colors-color-depth/571154

JPEG Compression

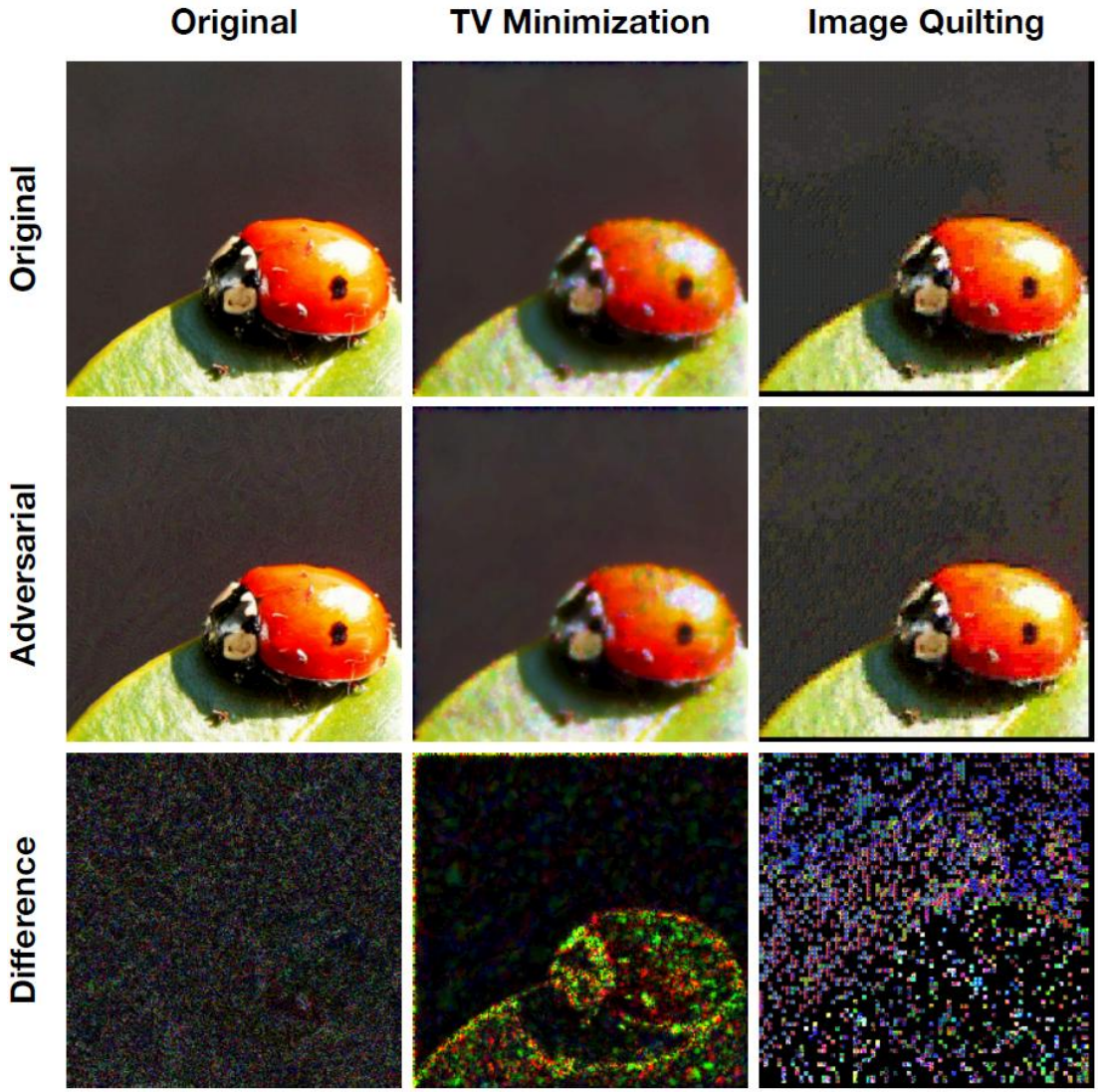


https://clubeestech.com/the-science-behind-image-compression-how-it-works/

Guo, C., Rana, M., Cisse, M. and Van Der Maaten, L., 2017. Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117.

Input Transformations as Defence

- Transformations
  - Total variation minimization
    - Remove small perturbations in the image.
  - Image quilting
    - synthesizes images by piecing together small patches that are taken from a database of image patches
      - Extracting 1,000,000 random patches from the ImageNet training set.
      - Constructing a patch database that only contains patches from “clean” images (without adversarial perturbations)
      - The patches are selected by finding the K nearest neighbors (in pixel space) of the corresponding patch from input image in the patch database, and picking one of these neighbors uniformly at random.



Difference images were multiplied by a constant scaling factor to increase visibility.

Guo, C., Rana, M., Cisse, M. and Van Der Maaten, L., 2017. Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117.

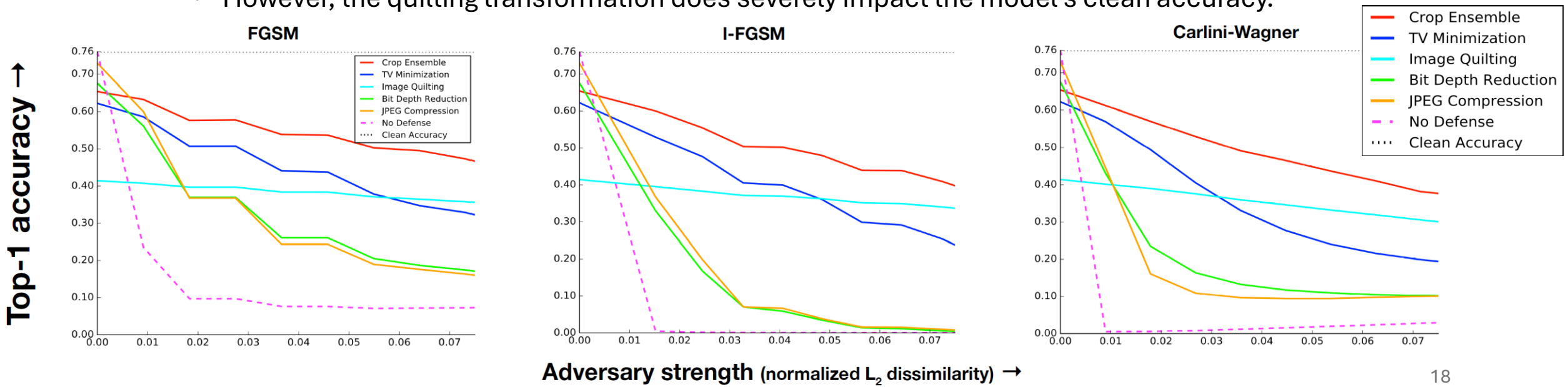
Input Transformations as Defence

- Experimental setup
  - Target model: ResNet-50
    - Trained on ImageNet: over 1 million training images.
  - Attacks:
    - Fast gradient sign method (FGSM)
    - Iterative fast gradient sign method (I-FGSM)
      - Iteratively apply the FGSM update
      - Similar to projected gradient descent (PGD). The difference is that I-FGSM does not project perturbations within a p-norm.
    - CW L2 attack
      - We experiment with this strong attack in a lab.
  - Measuring attack strength
    - Normalized L2-dissimilarity: $\frac{1}{N}\sum_{n=1}^{N}\frac{\|\mathbf{x}_n - \mathbf{x}'_n\|_2}{\|\mathbf{x}_n\|_2}$
      - x_n denotes clean image
      - x'_n denotes attack
      - Relative magnitude of perturbations compared to clean images.

Guo, C., Rana, M., Cisse, M. and Van Der Maaten, L., 2017. Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117.

Input Transformations as Defence

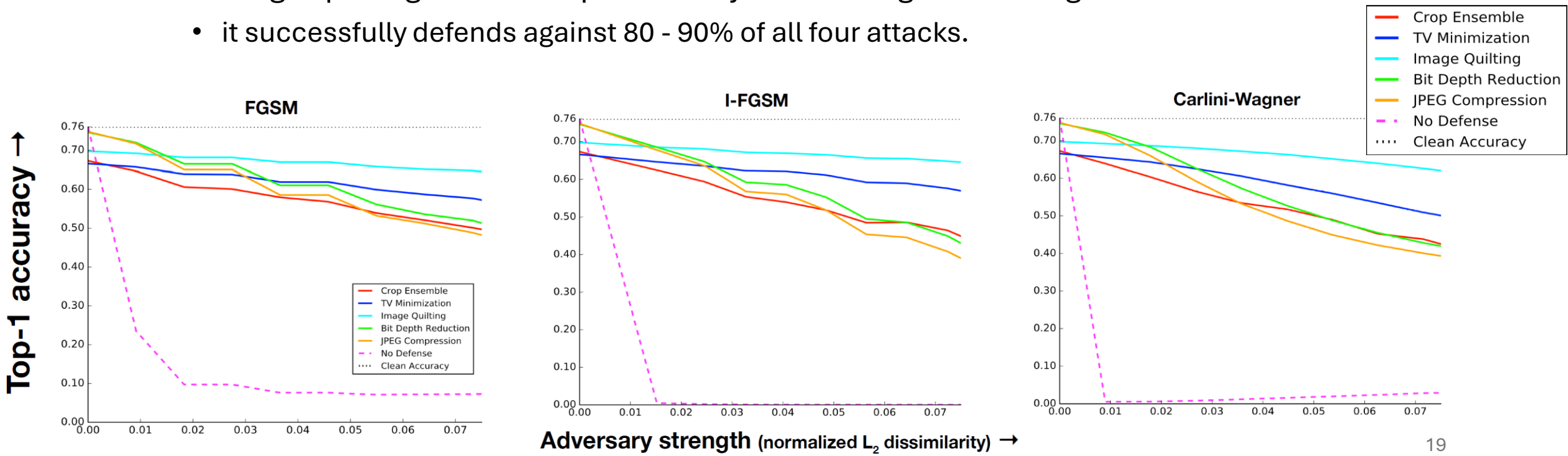
- Gray-box attack (adversaries know everything except for the defence)
  - Gradients w.r.t input can be calculated.
  - Results of transforming adversarial images before using them as input.
    - Ensemble cropping-rescaling is very efficient: maintaining 40 – 60% accuracy.
      - Cropping-rescaling is included in data augmentation during training.
        - The other transformations are not.
      - Adversarial perturbations are susceptible to changes in the location and scale.
    - Accuracy of the image-quilting defense hardly decreases.
      - Slight decrease can be due to the decease in image quality as attacks get stronger.
      - However, the quilting transformation does severely impact the model's clean accuracy.



Guo, C., Rana, M., Cisse, M. and Van Der Maaten, L., 2017. Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117.

Input Transformations as Defence

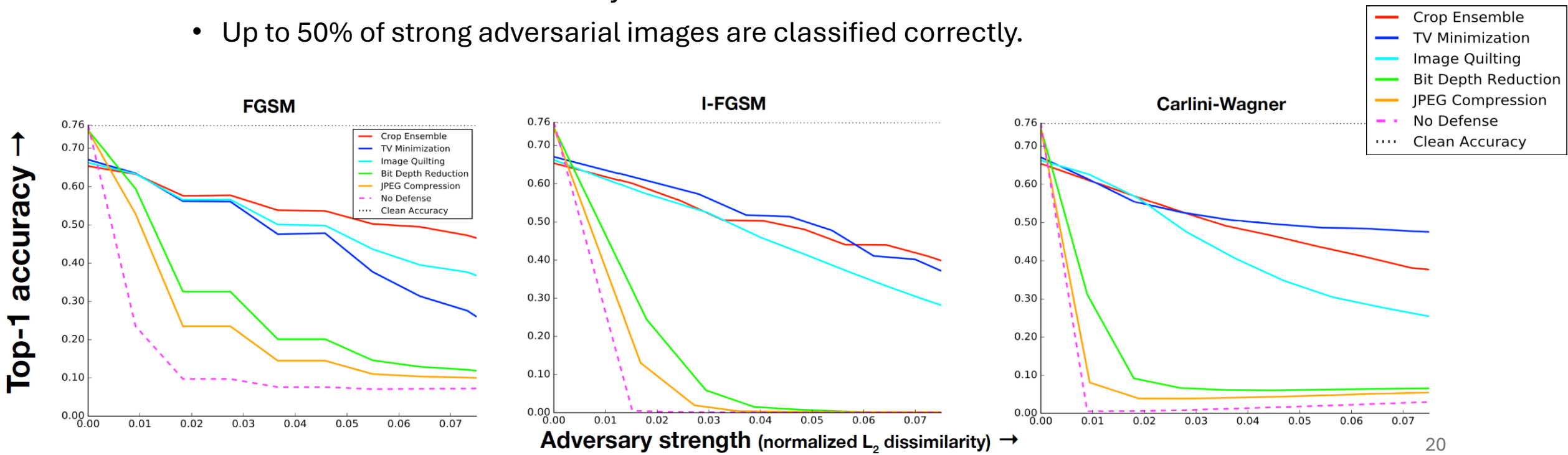
- Black-box attack (cannot calculate gradients & do not know the defence)
  - All the transformations are included as data augmentation during training.
    - Transfer previously generated attacks to black-box models
  - Results
    - Robustness is dramatically improved for all the defense.
    - the image-quilting defense is particularly effective against strong attacks
      - it successfully defends against 80 - 90% of all four attacks.



Guo, C., Rana, M., Cisse, M. and Van Der Maaten, L., 2017. Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117.

Input Transformations as Defence

- Gray-box attack (adversaries know everything except for the defence)
  - All the transformations are included as data augmentation during training.
    - Gradients w.r.t input can be calculated.
  - Results
    - Bit-depth reduction and JPEG compression are weak defenses in such a setting.
    - The other defences are still fairly robust.
      - Up to 50% of strong adversarial images are classified correctly.



Guo, C., Rana, M., Cisse, M. and Van Der Maaten, L., 2017. Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117.

Input Transformations as Defence

- Security through obscurity
  - This defence assumes that adversaries do not know defence deployed.
    - This is an impractical assumption.
      - As soon as the work is published, any potential adversary adjust their attacks accordingly
    - Deterministic transformations are easy to defeat.
    - Stochastic transformations may be robust as each run uses a slightly different function.
      - E.g., crop ensemble and image quilting.

- Can adaptive attacks bypass stochastic pre-processing defence?
  - If answer is true, pre-processing defence is not secure (yet).

Guo, C., Rana, M., Cisse, M. and Van Der Maaten, L., 2017. Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117.

Adaptive attacks against Pre-processing Defence

- Expectation over Transformation (EOT)
  - A standard technique to break pre-processing defence.
    - Use the expectation of the gradient to generate attacks, e.g. PGD:

$$\boldsymbol{x}^{i+1} \leftarrow \boldsymbol{x}^{i} + \alpha \cdot \operatorname{sgn}\left\{ \mathbb{E}_{\boldsymbol{\theta} \sim \Theta}\left[\nabla \mathcal{L}\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}^{i}), y \right) \right] \right\} \approx \boldsymbol{x}^{i} + \alpha \cdot \operatorname{sgn}\left\{ \frac{1}{m} \sum_{j=1}^{m} \nabla \mathcal{L}\left(f_{\boldsymbol{\theta}_j}(\boldsymbol{x}^{i}), y \right) \right\}$$

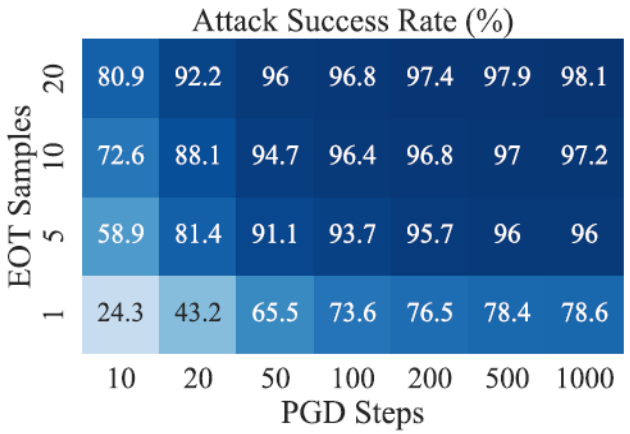
      - θ is the random variable drawn from some randomization space Θ that parameterizes the defense.
      - α is the step size.
      - L is the loss function
      - Each iteration is projected to the L_∞ ball around x^0 of radius ϵ.

    - A similar idea was used to generate physical adversarial examples
      - Include potential physical transformations into the generation process.

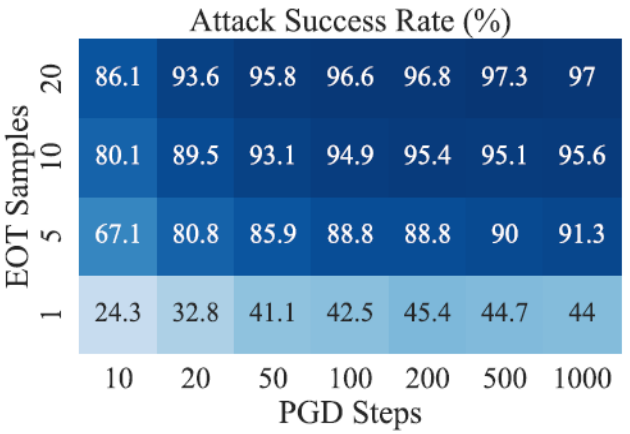
Gao, Y., Shumailov, I., Fawaz, K. and Papernot, N., 2022. On the limitations of stochastic pre-processing defenses. Advances in neural information processing systems, 35, pp.24280-24294.

Adaptive attacks against Pre-processing Defence

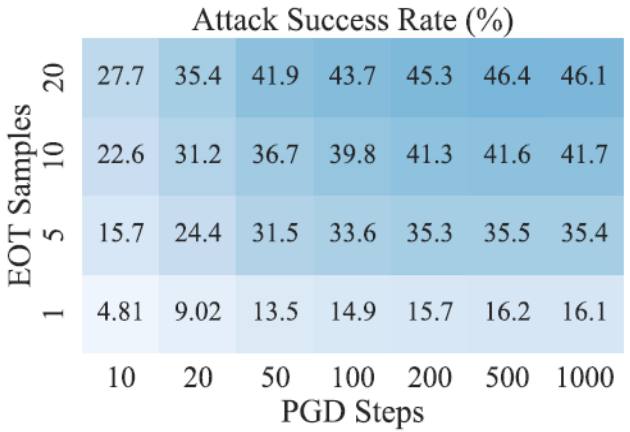
- Defeating defence that adds Gaussian noise to the input image
  - Generating **targeted** attacks using PGD+EOT.
    - Equivalent to “PGD only” if EOT samples=1.
  - Results
    - EOT starts to improve the attack when the defense has a higher level of randomness.
    - For a fixed number of PGD steps, applying EOT significantly improves the attack in most of the settings.



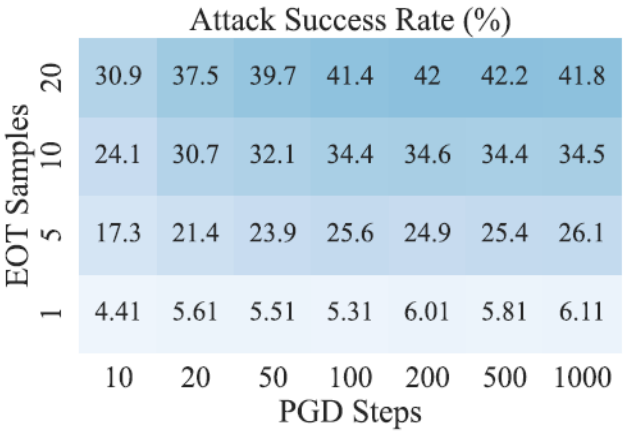
$\sigma = 0.25$ (step size = 2/255)



$\sigma = 0.25$ (step size = 4/255)



$\sigma = 0.50$ (step size = 2/255)



$\sigma = 0.50$ (step size = 4/255)

Gao, Y., Shumailov, I., Fawaz, K. and Papernot, N., 2022. On the limitations of stochastic pre-processing defenses. Advances in neural information processing systems, 35, pp.24280-24294.

Detecting Adversarial Examples

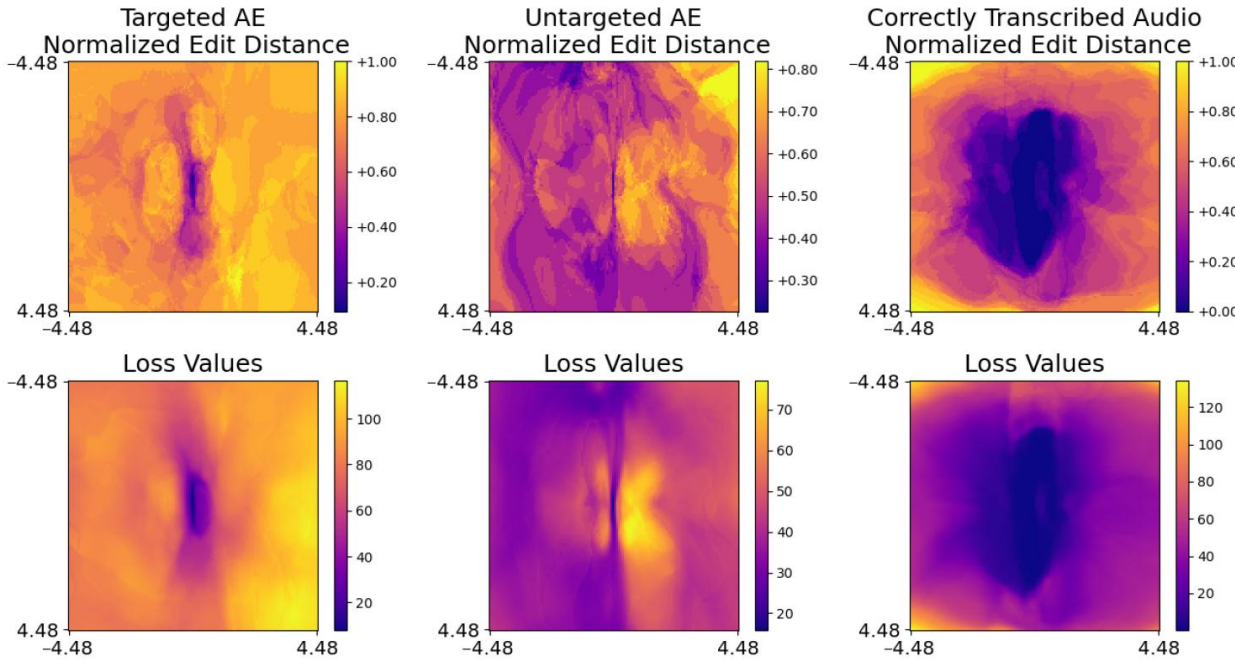
- Detecting adversarial attack is another option
  - In addition to making models adversarially robust or destroying adversarial perturbations.
  - Key is to find properties to distinguish between normal input and attacks.
    - Detect robots that hide in human society.
      - Robots “eat” electricity vs. humans eat food
      - Robots do not have emotions (yet?) vs. humans have
      - Etc.
- Detect attacks for speech-to-text using decision boundary patterns
  - Untargeted adversarial examples
    - Fool a speech-to-text model to output **incorrect** transcripts.
  - Targeted adversarial examples
    - Fool a speech-to-text model to output **predefined** transcripts.
      - “Open the garage door”
      - “Call 12341234”
      - “Turn on the gas cooker”
      - Etc.

Zong, W., Chow, Y.W., Susilo, W., Kim, J. and Le, N.T., 2022. Detecting Audio Adversarial Examples in Automatic Speech Recognition Systems Using Decision Boundary Patterns. *Journal of Imaging*, *8*(12), p.324.

Detecting Adversarial Examples

- The center of the heat maps represents unmodified audio
  - The horizontal axis is the direction of the gradient of the loss function w.r.t. the input audio.
  - The vertical axis is a random perpendicular direction.
- Observations
  - When a **targeted AE** is modified slightly, the resulting loss function value and normalized edit distance change significantly.
    - Normalized edit distance: the change in text divided by length of the original transcribed text.
    - Adversarial perturbations in targeted audio AEs are not robust.
  - Changes in loss function values and normalized edit distances for **correctly transcribed benign audio** are significantly smaller
    - Correctly transcribed benign audio is much more robust against perturbations.
  - Robustness of the **untargeted AE** is in the middle.
    - But still shows a different pattern.

Target model: **DeepSpeech** (open-sourced by Mozilla).



(a) Targeted AE (DS) (b) Untargeted AE (DS) (c) Benign Audio (DS)

Zong, W., Chow, Y.W., Susilo, W., Kim, J. and Le, N.T., 2022. Detecting Audio Adversarial Examples in Automatic Speech Recognition Systems Using Decision Boundary Patterns. *Journal of Imaging*, *8*(12), p.324.

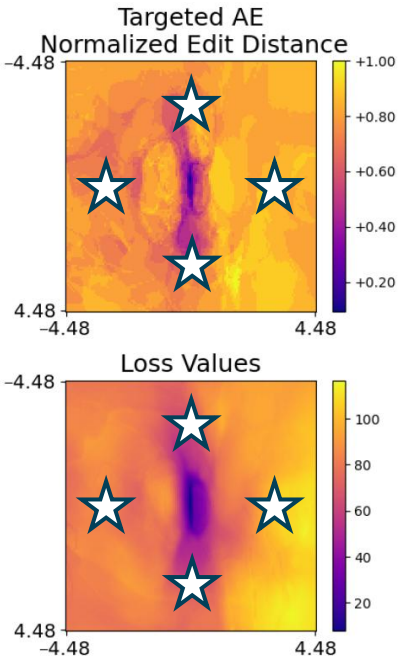
Detecting Adversarial Examples

- Extracting features from the patterns

$$v_{loss} = \begin{bmatrix} \ell_{net}(f(x+\bar{g}),y) - \ell_{net}(f(x),y) \\ \ell_{net}(f(x-\bar{g}),y) - \ell_{net}(f(x),y) \\ \ell_{net}(f(x+\bar{p}),y) - \ell_{net}(f(x),y) \\ \ell_{net}(f(x-\bar{p}),y) - \ell_{net}(f(x),y) \end{bmatrix}$$

$$v_{edit} = \begin{bmatrix} d_{edit}(f(x+\bar{g}),y)/h_{length}(y) \\ d_{edit}(f(x-\bar{g}),y)/h_{length}(y) \\ d_{edit}(f(x+\bar{p}),y)/h_{length}(y) \\ d_{edit}(f(x-\bar{p}),y)/h_{length}(y) \end{bmatrix}$$

$$v_{ft} = \begin{bmatrix} v_{loss} \\ v_{edit} \end{bmatrix}$$

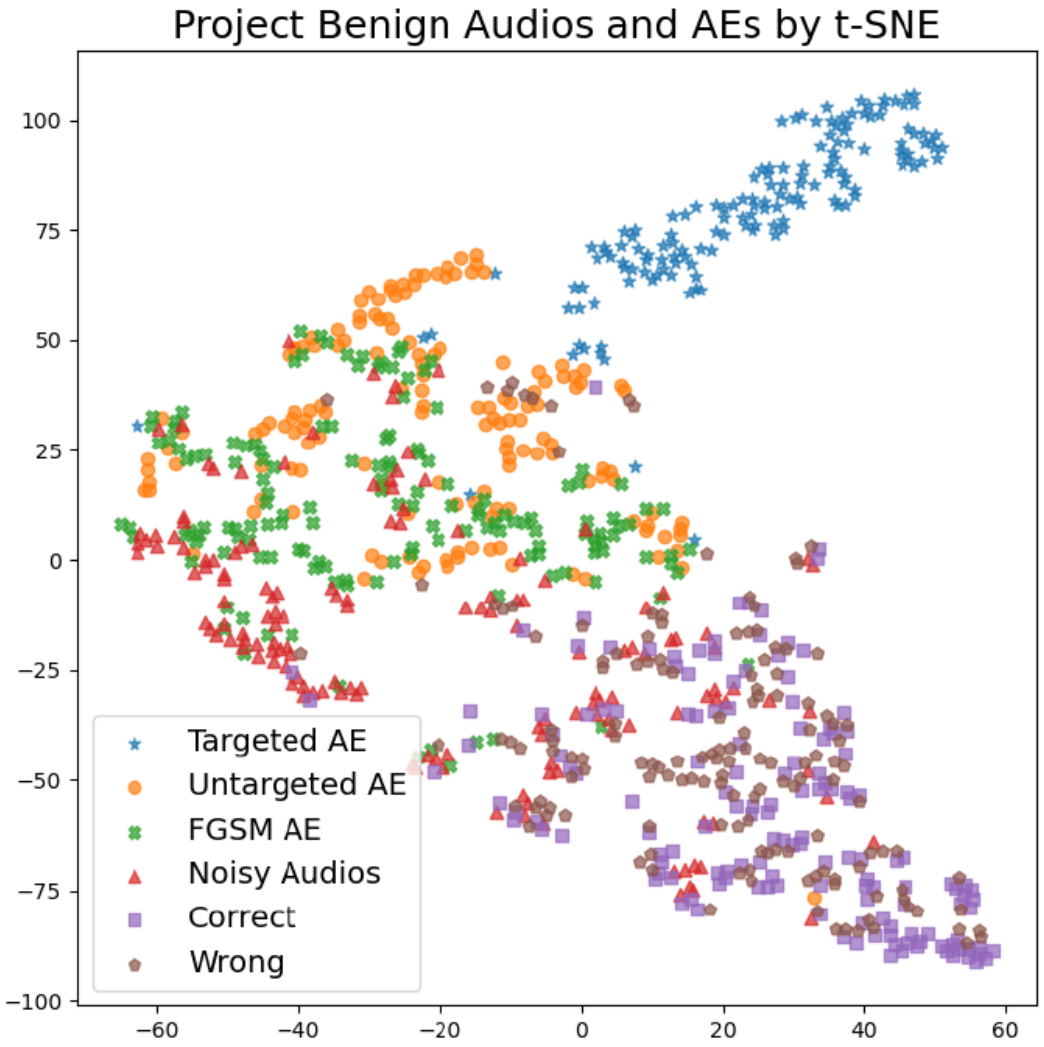


- x is the input audio and y be the transcript of x output by the model.
- $\bar{g}$ is the gradient direction and $\bar{p}$ is a direction perpendicular to $\bar{g}$.

Zong, W., Chow, Y.W., Susilo, W., Kim, J. and Le, N.T., 2022. Detecting Audio Adversarial Examples in Automatic Speech Recognition Systems Using Decision Boundary Patterns. *Journal of Imaging*, *8*(12), p.324.

Detecting Adversarial Examples

- Dimension Reduction
  - Using t-SNE (a dimensionality reduction method) to project features onto 2D space.
  - Three clusters, excluding noisy audio, can be identified
    - **Targeted audio AEs** are clearly grouped in the first cluster.
    - The second cluster mainly contains **correctly and incorrectly transcribed benign audio**.
    - The third cluster consists of **untargeted audio AEs** and **FGSM AEs**, i.e., untargeted attacks.
  - The various audio types are clustered according to their categories.
  - Incorrectly transcribed audio does not overlap with untargeted audio AEs or FGSM AEs.
    - although all of them lead to incorrect transcriptions.
  - Noisy audio is contained in both the second cluster (benign audio) and the third cluster (untargeted attack).



Zong, W., Chow, Y.W., Susilo, W., Kim, J. and Le, N.T., 2022. Detecting Audio Adversarial Examples in Automatic Speech Recognition Systems Using Decision Boundary Patterns. *Journal of Imaging*, *8*(12), p.324.

Detecting Adversarial Examples

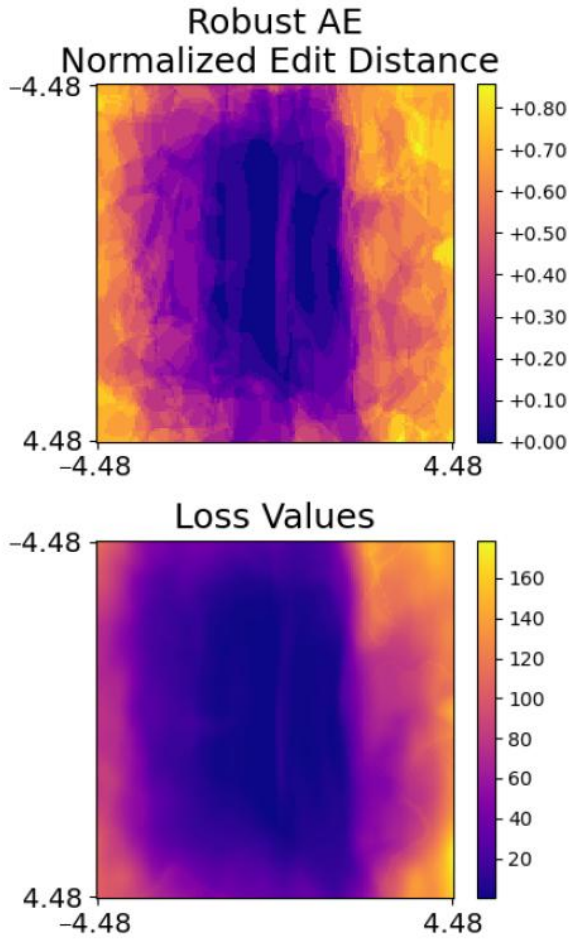
DeepSpeech					
Type	**TP**	**FP**	**TN**	**FN**	**DR**
Targeted AEs	150	-	-	0	100.00%
Untargeted AEs	120	-	-	30	80.00%
FGSM AEs	86	-	-	64	57.33%
Noisy Audio	-	9	141	-	94.00%
Correctly trans.	-	4	146	-	97.33%
Incorrectly trans.	-	6	144	-	96.00%
	Pre	**Rec**	**Acc**		
	94.93%	79.11%	87.44%		

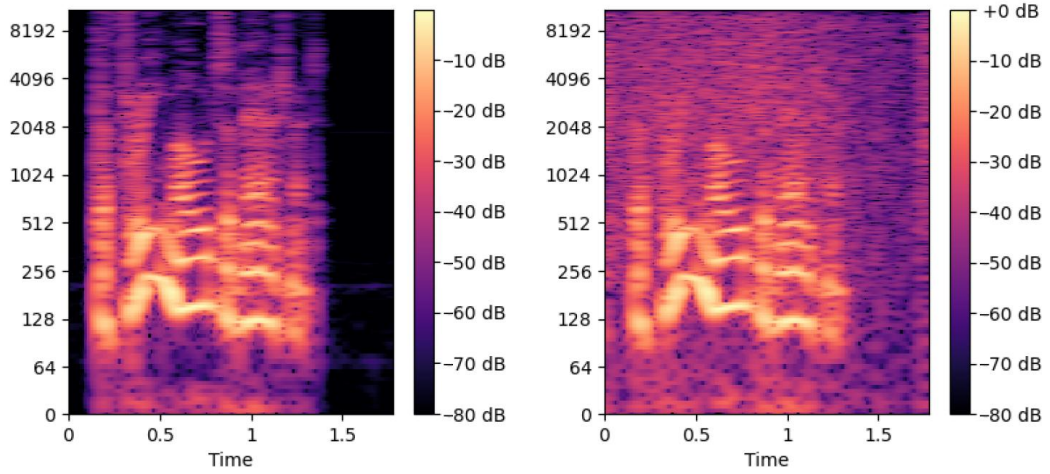
- Train an EllipticEnvelope model implemented by scikit-learn.
  - This simple model detects outliers in a Gaussian-distributed data set.
- Targeted AEs are easily detected at detection rates of **100%**.
  - This is consistent with the observation that targeted AEs can clearly be separated from other audio types in lower-dimensional space.
- The detection rates of FGSM and untargeted AEs are lower.
  - False negatives of untargeted attacks are much less harmful than targeted attacks.

Zong, W., Chow, Y.W., Susilo, W., Kim, J. and Le, N.T., 2022. Detecting Audio Adversarial Examples in Automatic Speech Recognition Systems Using Decision Boundary Patterns. *Journal of Imaging*, *8*(12), p.324.

Detecting Adversarial Examples

- Adaptive adversarial perturbations
  - Can an adversary generate attacks that can bypass this detection?
    - Unfortunately, this is feasible.
    - The figure on the right shows an example.
  - However, the noise introduced is obvious
    - The figure below shows spectrograms of a robust attack.
    - This abnormal noise can be detected by other techniques.
  - Can the perturbations be quiet?
    - Endless battle between attackers and defenders.





(a) Benign audio (b) Robust AE

Zong, W., Chow, Y.W., Susilo, W., Kim, J. and Le, N.T., 2022. Detecting Audio Adversarial Examples in Automatic Speech Recognition Systems Using Decision Boundary Patterns. *Journal of Imaging*, *8*(12), p.324.

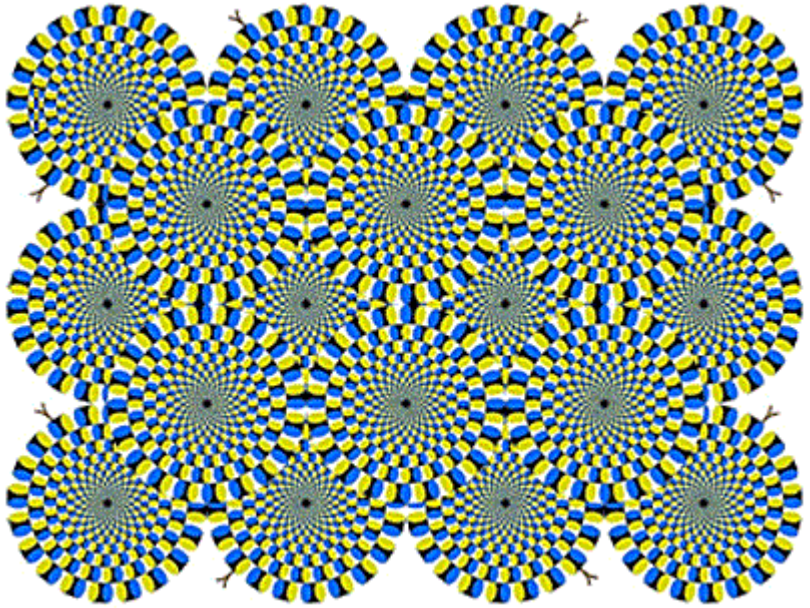
Reflections on Adversarial Examples

- The existence of adversarial examples shows inconsistence between human and machine intelligence.
  - It is acceptable that humans and machines are fooled in the same way.
    - E.g., fooled by optical illusions.
  - It is security issue if humans are not fooled but machines are fooled.

How many animals do you see in the image?

https://www.reddit.com/r/opticalillusions/comments/z2ir1t/the_rabbit_is_right_handedthe_duck_disagreesdo/?rdt=61232

Are the circles moving in the image?

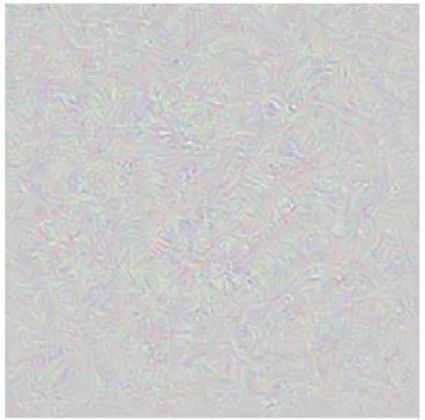


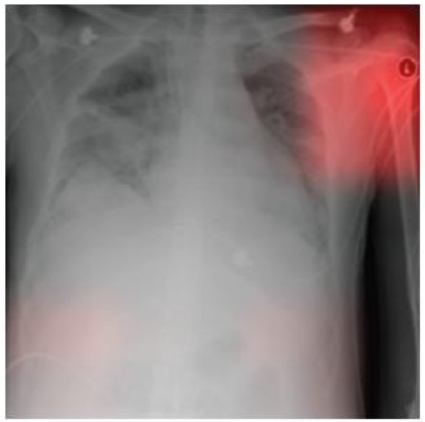
https://en.wikipedia.org/wiki/File:Illusions.jpg

Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M. and Wichmann, F.A., 2020. Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), pp.665-673.

Reflections on Adversarial Examples

- Shortcut learning explains adversarial examples.
  - Clever Hans used shortcut learning.
    - Humans may use shortcut learning as well, e.g., answering multiple choice questions.
  - Machine learning/deep learning models also rely on shortcuts.





Article: Super Bowl 50

Paragraph: "*Peython Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had a jersey number 37 in Champ Bowl XXXIV.*"

Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

Task for DNN	Caption image	Recognise object	Recognise pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognise primary object	Uses features irrecognisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

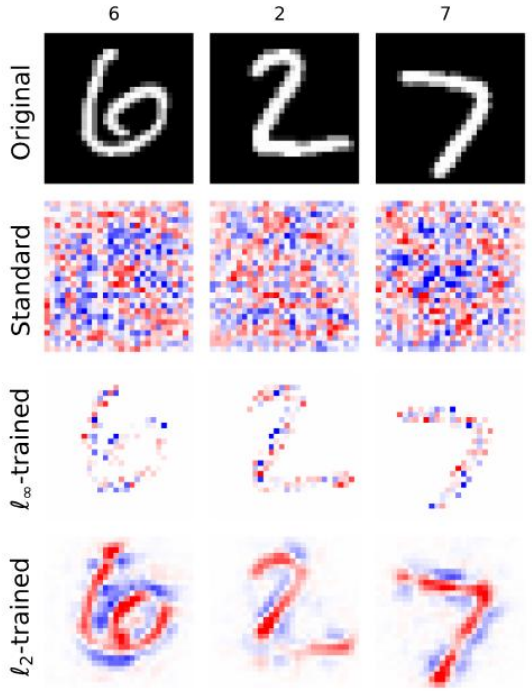
Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M. and Wichmann, F.A., 2020. Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), pp.665-673.

Reflections on Adversarial Examples

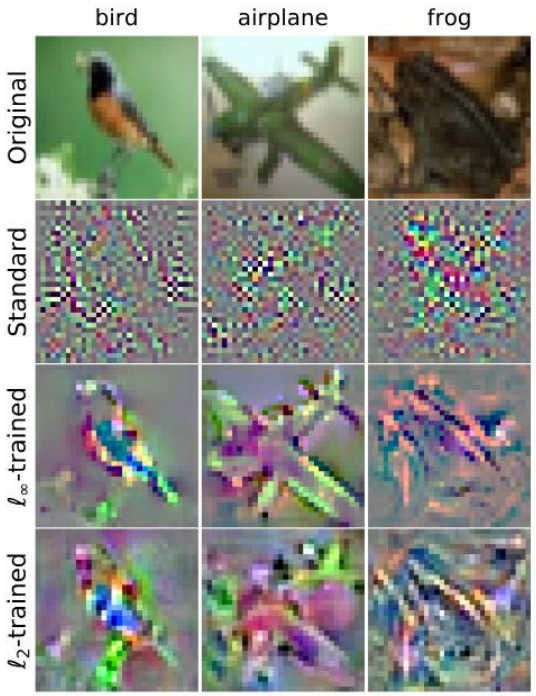
- Eliminating shortcut learning can be a solution to adversarial examples.
  - Existence of adversarial examples implies shortcut learning is a common strategy for models in **ALL** fields.
    - The loss function for training does not consider human perception and learning.
      - Cross-entropy loss for classification tasks.
    - Human perception cannot be mathematically formulated (yet?).
    - If models do not rely on shortcut learning, adversarial examples do not exist.
      - If an attack fools models, it should also fool humans, e.g., optical illusions.
  - Investigating how humans learn knowledge may eventually eliminate shortcut learning in models.
    - Researchers are actively bringing together computer scientists, cognitive scientists, and neuroscientists to work in close collaboration.
    - This new field is dedicated to developing a computationally based understanding of human intelligence and establishing an engineering practice based on that understanding.

Reflections on Adversarial Examples

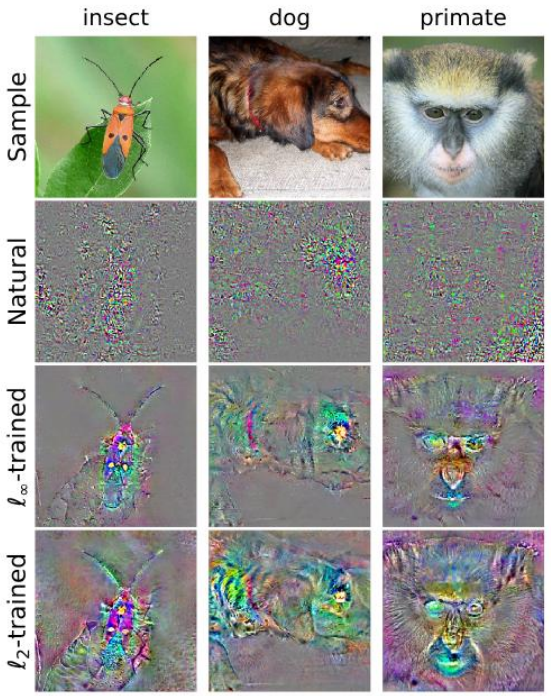
- Eliminate shortcut learning.
  - Empirical evidence shows that adversarial robustness is somehow related to human perception.
    - Visualization of the loss gradient with respect to input pixels for adversarially trained networks.



(a) MNIST



(b) CIFAR-10

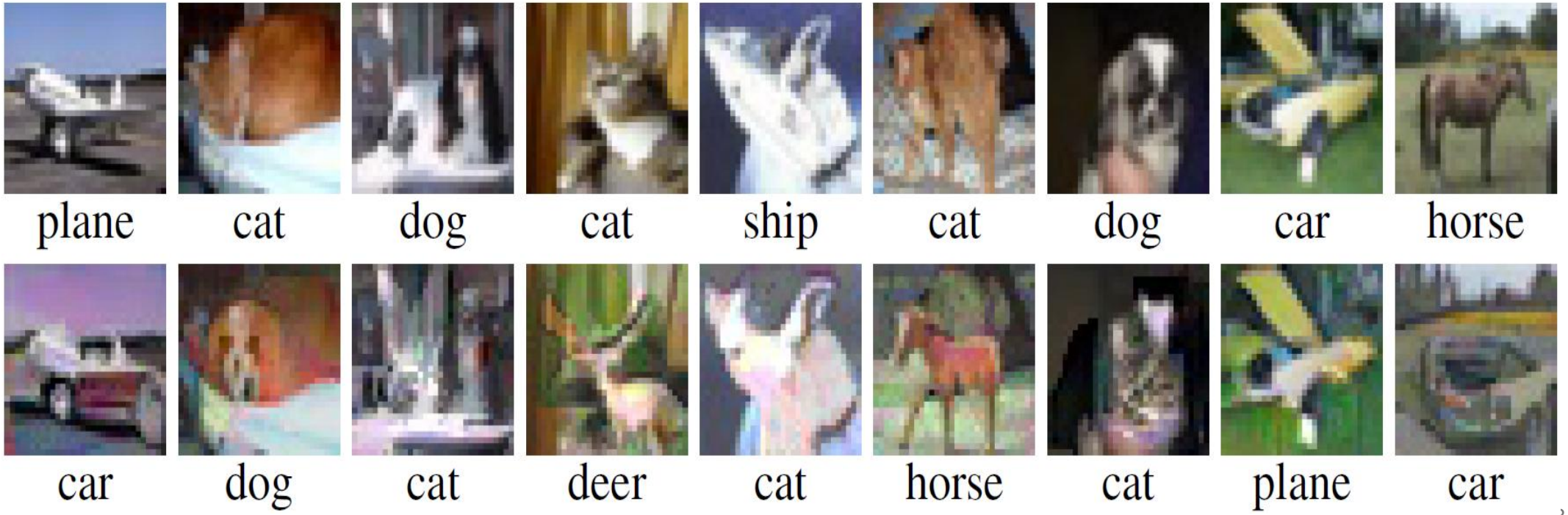


(c) Restricted ImageNet

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A. and Madry, A., 2018. Robustness may be at odds with accuracy. arXiv preprint arXiv:1805.12152.

Reflections on Adversarial Examples

- Eliminate shortcut learning.
  - Empirical evidence shows that adversarial robustness is somehow related to human perception.
    - Attack images built for adversarially trained models look like the class into which they get misclassified.
      - First row: clean images
      - second row: adversarial examples against an adversarially trained model.



Shafahi, A., Najibi, M., Ghiasi, M.A., Xu, Z., Dickerson, J., Studer, C., Davis, L.S., Taylor, G. and Goldstein, T., 2019. Adversarial training for free!. Advances in neural information processing systems, 32.

References

- Goodfellow, I.J., Shlens, J. and Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.
- Zong, W., Chow, Y.W., Susilo, W., Kim, J. and Le, N.T., 2022. Detecting Audio Adversarial Examples in Automatic Speech Recognition Systems Using Decision Boundary Patterns. Journal of Imaging, 8(12), p.324.
- Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M. and Wichmann, F.A., 2020. Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), pp.665-673.
- Tsipras, D., Santurkar, S., Engstrom, L., Turner, A. and Madry, A., 2018. Robustness may be at odds with accuracy. arXiv preprint arXiv:1805.12152.
- Guo, C., Rana, M., Cisse, M. and Van Der Maaten, L., 2017. Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117.
- Gao, Y., Shumailov, I., Fawaz, K. and Papernot, N., 2022. On the limitations of stochastic pre-processing defenses. Advances in neural information processing systems, 35, pp.24280-24294.