Feature Denoising for Improving Adversarial Robustness

Xie, Cihang, et al., Facebook AI Research

Sein Jang

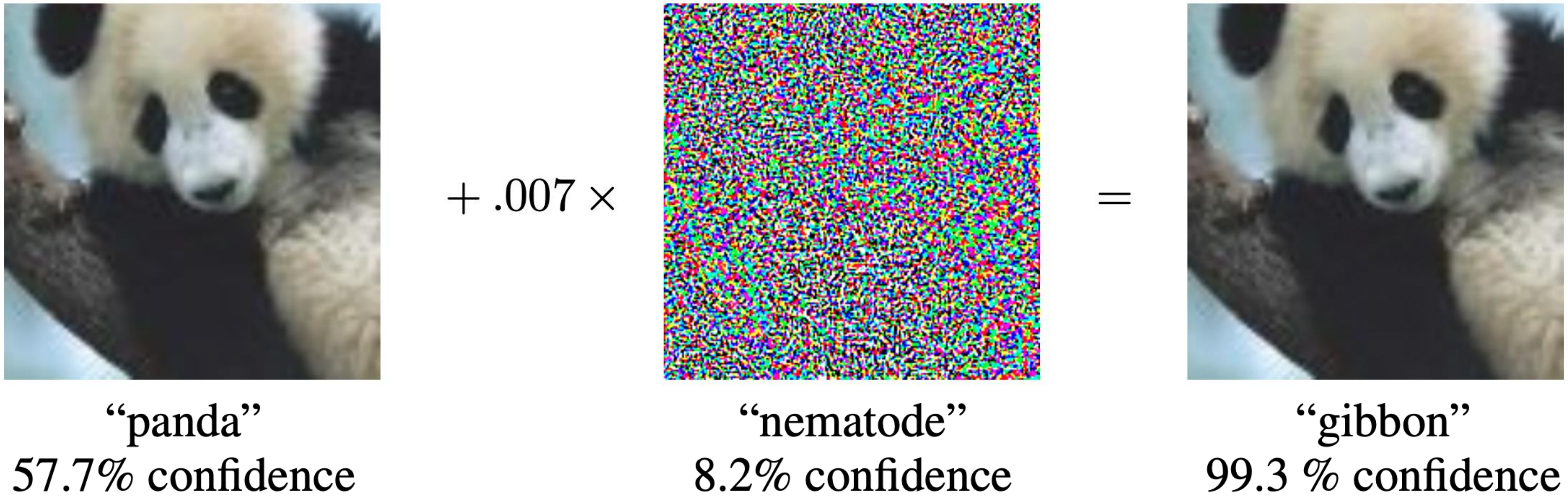
tpdls24@gmail.com

March 27, 2020

Contents

Adversarial Attack

An **adversarial attack** consists of subtly **modifying an original image** in such a way that the **changes are almost undetectable to the human eye**. The modified image is called an adversarial image, and when submitted to a classifier is misclassified, while the original one is correctly classified.



Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples"

Adversarial Examples

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification"

- With a perturbation in the form of only black and white stickers
- Attack a real stop sign, causing targeted misclassification in 84.8% of the captured video frames obtained on a moving vehicle for the target classifier
- These attacks can cause serious problems for autonomous driving systems

Adversarial Examples

Thys, Simen, Wiebe Van Ranst, and Toon Goedemé. "Fooling automated surveillance cameras: adversarial patches to attack person detection."

- Small printable patch can successfully hide a person from a person detector
- An attack that could for instance be used maliciously to circumvent surveillance systems

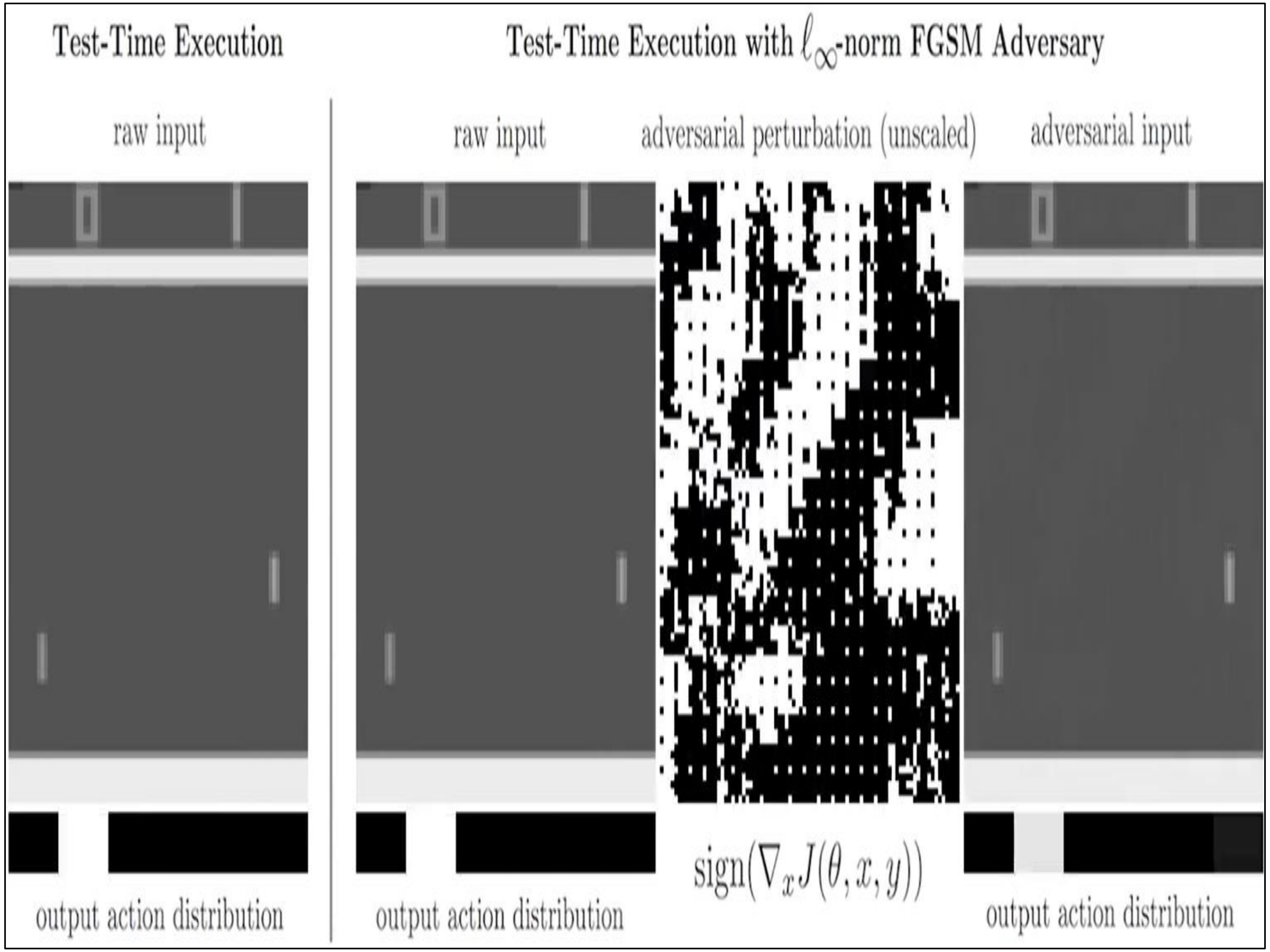
Adversarial Examples

	Confidence(%)	Prediction	Text
Original	97.9	1	I enjoyed this film which I thought was well written and acted , there was plenty of humour and a provoking storyline, a warm and enjoyable experience with an emotional ending.
	99.7	0	I am sorry but this is the worst film I have ever seen in my life. I cannot believe that after making the first one in the series, they were able to get a budget to make another. This is the least scary film I have ever watched and laughed all the way through to the end.
	95.8	1	This is a unique masterpiece made by the best director ever lived in the ussr. He knows the art of film making and can use it very well. If you find this movie, buy or copy it!
GA	50.6	0	I cared this film which I thought was well written and acted, there was plenty of humour and a igniting storyline, a tepid and enjoyable experience with an emotional ending.
	92.7	1	I am sorry but this is the harshest film I have ever seen in my life. I cannot believe that after making the first one in the series, they were able to get a budget to make another. This is the least scary film I have ever watched and laughed all the way through to the end.
	59.0	0	This is a sole masterpiece made by the nicest director permanently lived in the ussr. He knows the art of film making and can use it much well. If you find this movie, buy or copy it!

Wang, Xiaosen, Hao Jin, and Kun He. "Natural language adversarial attacks and defenses in word level."

- Adversarial Attack in NLP

Adversarial Examples

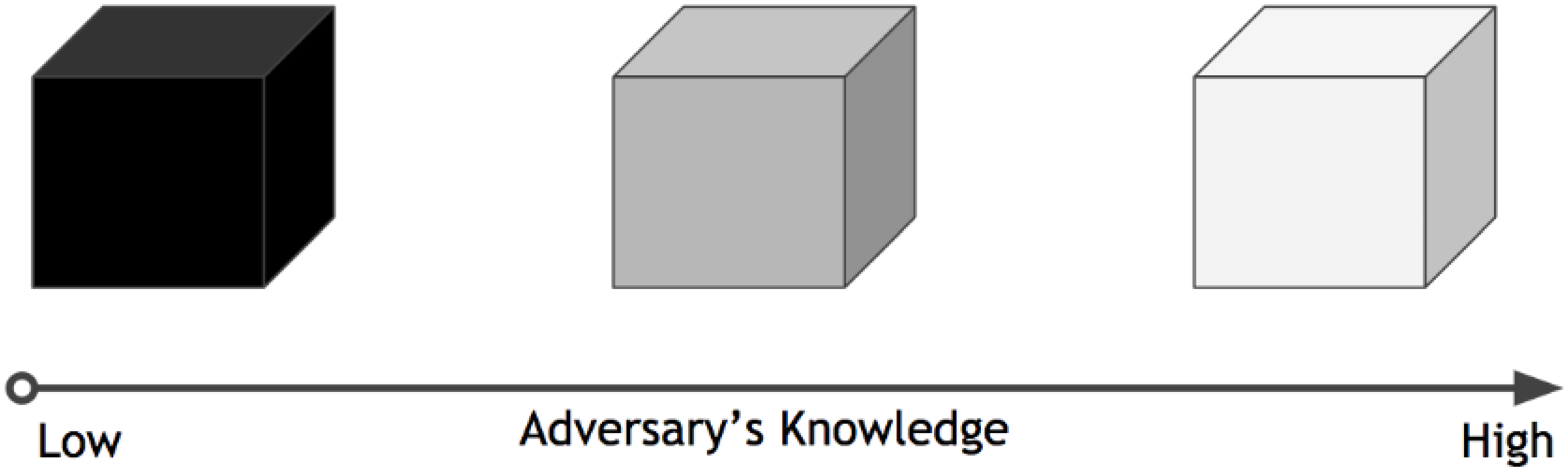


- Reinforcement learning agents can also be manipulated by adversarial examples…

Huang, Sandy, et al. "Adversarial attacks on neural network policies."

Adversarial Attack

- **Attack Methods**
  - Poisoning Attack
  - Evasion Attack
  - Targeted Attack
  - Non-Targeted Attack

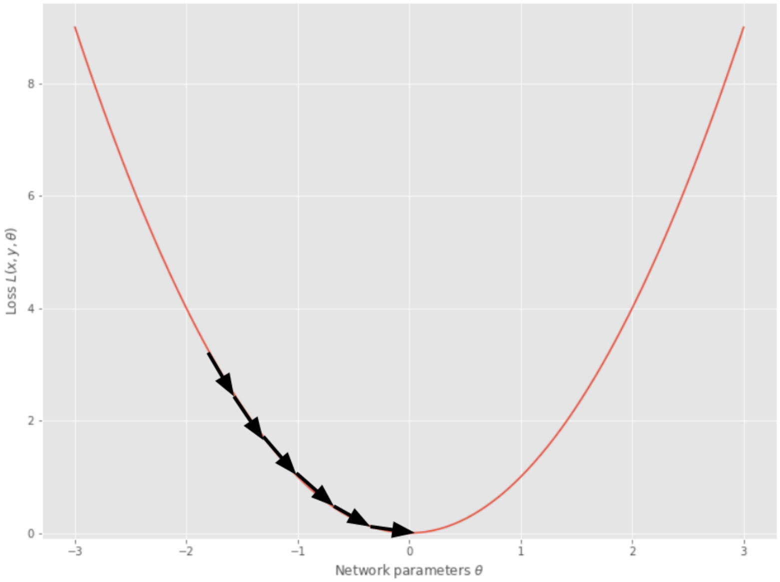


https://secml.github.io/class1/

Adversarial Attack

- **How to create an adversarial example**

Using gradient descent for train the neural networks



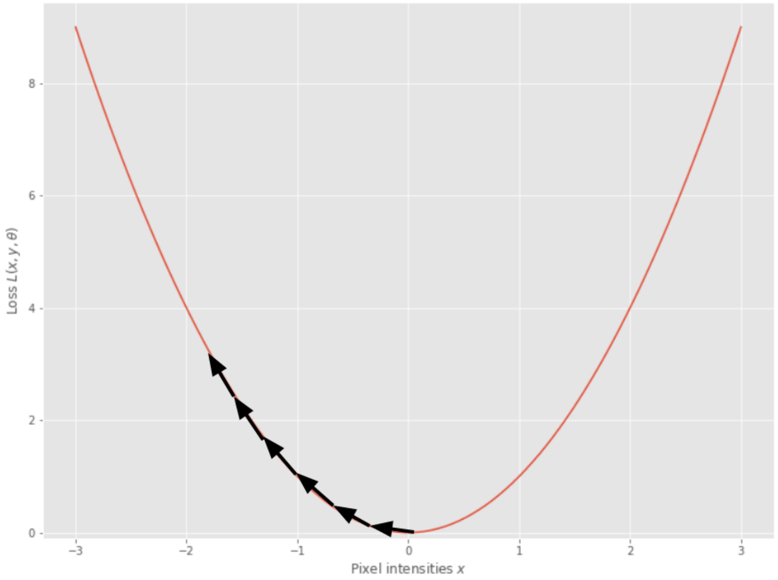
Define loss function

$$L(x, y, \theta) = (f_\theta(x) - y)^2$$

Update the parameters such that the loss will decrease

$$\theta' = \theta - \alpha \nabla_\theta L(x, y, \theta)$$

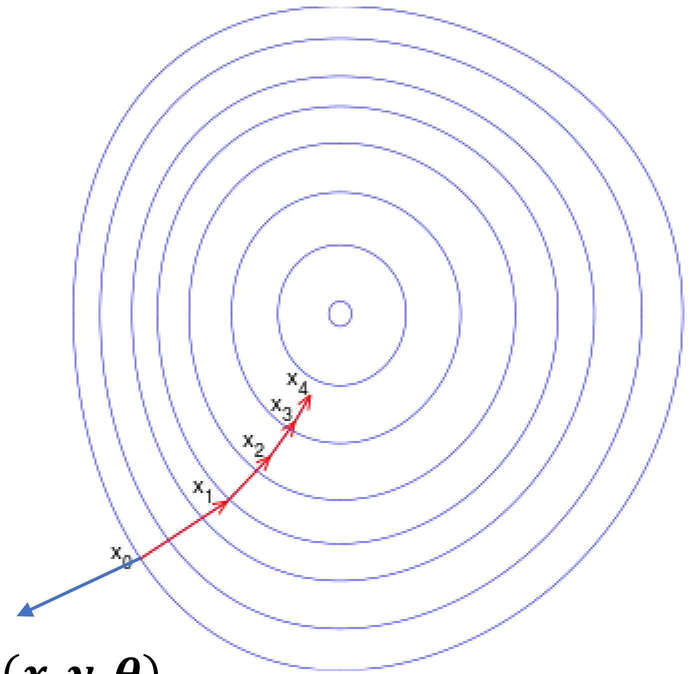
Using gradient ascent for creating an adversarial example



The model parameters will be held as constant, we could then update x in such a way that the expected loss of the model would increase.

$$\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\alpha \nabla_x L(x, y, \theta)}$$

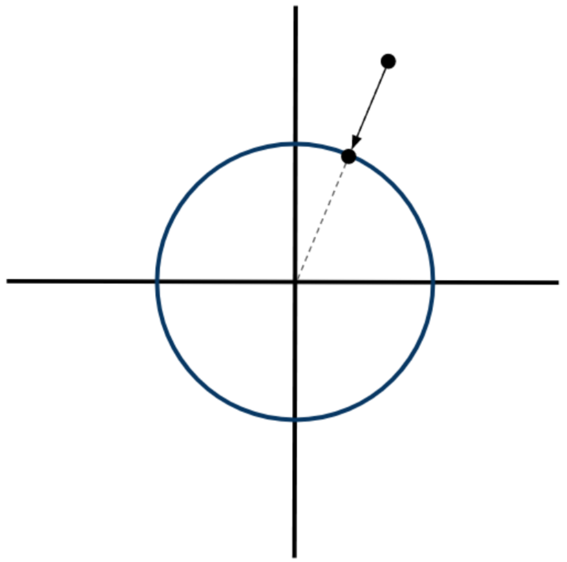
- **Fast Gradient-Sign Method** (Goodfellow et al. 2014)
  - Simplest method of creating an adversarial example.
  - Used as a benchmark.
  - Single step of gradient ascent.
  - Fix the perturbation on each pixel to be of fixed size, epsilon.



$$x' = x + \epsilon sign \nabla_x L(x, y, \theta)$$

Projected Gradient Descent

- **Projected Gradient Descent Attack**
  - White-box attack, attacker has a copy of target model's weights.
  - PGD attempts to find the perturbation that maximizes the loss of a model on a particular input while keeping the size of the perturbation smaller than a specified amount referred to as *epsilon.*
  - Constraint is usually expressed as the L^2 or L^∞ norm.
  - Detail steps of PGD
    1. Start from a random perturbation in the L^p space around a sample
    2. Take a gradient step in the direction of greatest loss
    3. Project perturbation back into L^p space if necessary
    4. Repeat



Projecting a point back into the L^2 ball in 2 dimensions

Feature Noise

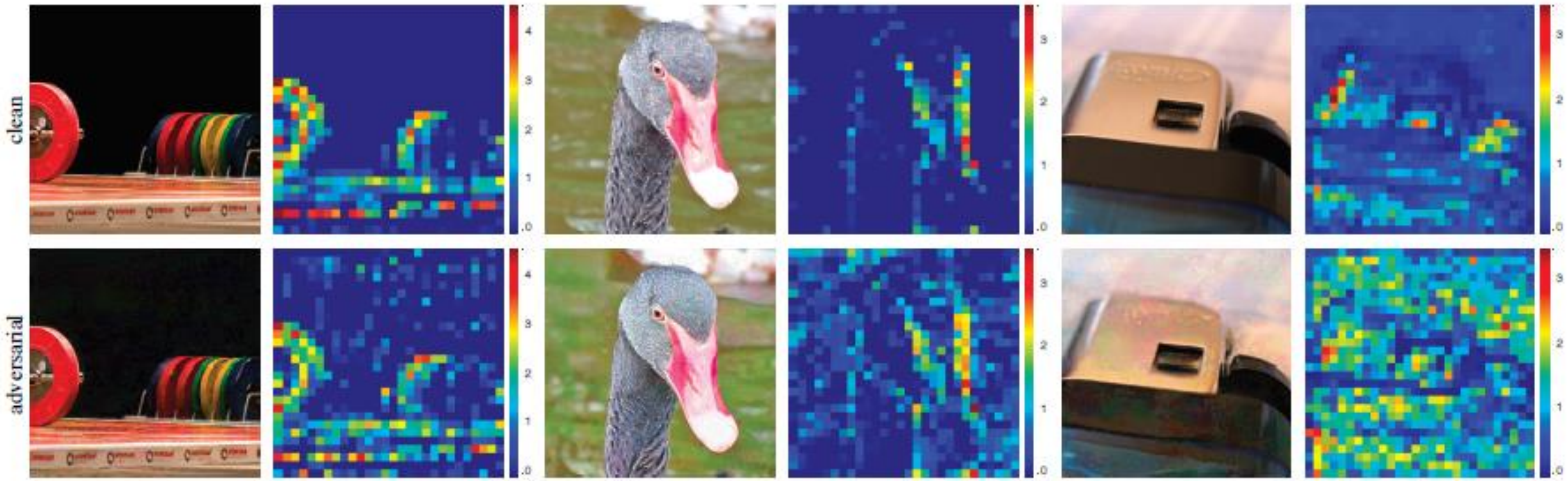


Figure 2. More examples similar to Figure 1. We show feature maps corresponding to clean images (top) and to their adversarial perturbed versions (bottom). The feature maps for each pair of examples are from the same channel of a res_3 block in the same ResNet-50 trained on clean images. The attacker has a maximum perturbation $\epsilon = 16$ in the pixel domain.

- The perturbations are constrained to be small at a the pixel level, no such constraints are imposed at the feature level in convolutional networks
- Assuming that strong activations that are hallucinated by adversarial images reveal why the model predictions are altered

Denoising Feature Maps

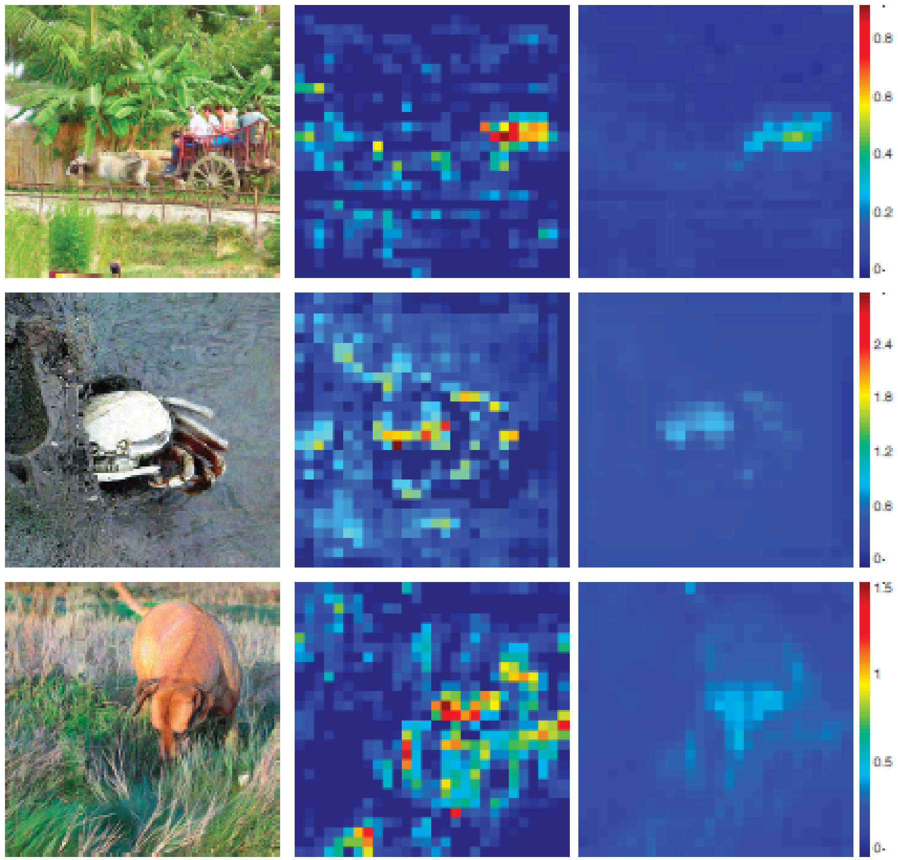
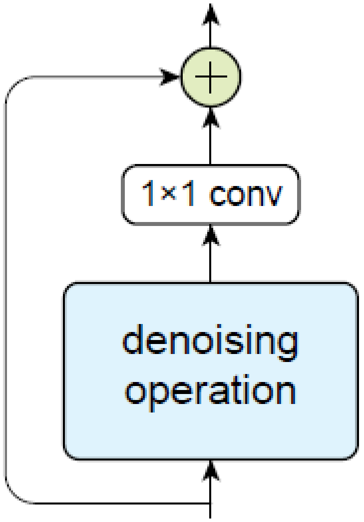
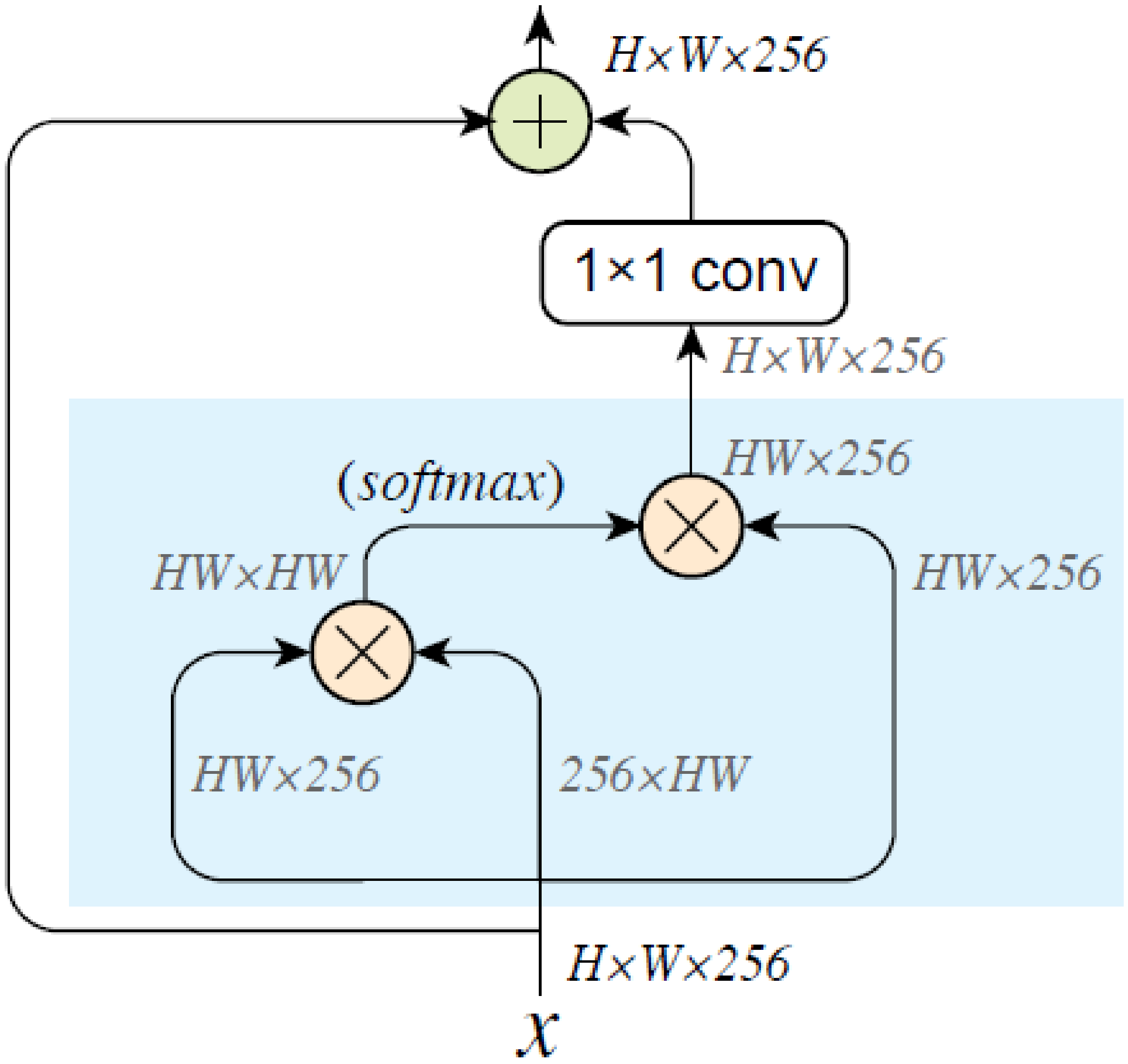


Figure 3. Adversarial images and their feature maps *before* (left) and *after* (right) the *denoising operation* (blue box in Figure 4). Here each pair of feature maps are from the same channel of a res_3 block in the same adversarially trained ResNet-50 equipped with (Gaussian) non-local means denoising blocks. The attacker has a maximum perturbation $\epsilon = 16$ for each pixel.

- Address this problem by **Feature Denoising**
- Feature denoising operations can successfully suppress much of the noise in the feature maps, and make the responses focus on visually meaningful content



Denoising Operations



- Residual connection can help the network to retain signals
- The tradeoff between removing noise and retaining signal is adjusted by the 1x1 convolution
- Denoising operations
  - **Non-local means**
  - Bilateral filters
  - Mean filters
  - Median filters

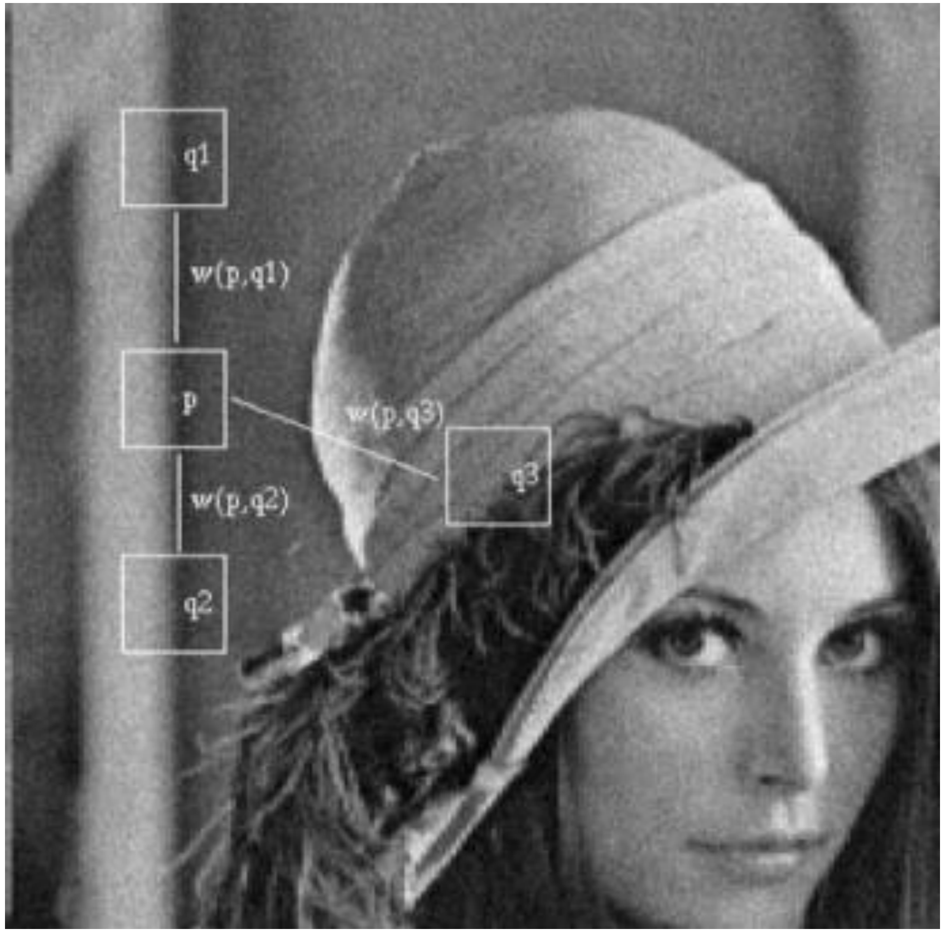
Denoising Operations

- **Image Denoising**

Buades, Antoni, Bartomeu Coll, and J-M. Morel. "A non-local algorithm for image denoising."

- **Non-local means Filter**
  - Non-local means filtering takes a mean of all pixels in the image, weighted by how similar these pixels are to the target pixel

$$NL[v](i) = \sum_{j \in I} w(i,j)v(j),$$



Denoising Operations

- **Non-local means**
  - Non-local means compute a denoised feature map y of an input feature map x by taking a weighted mean of features in all spatial locations $\mathcal{L}$

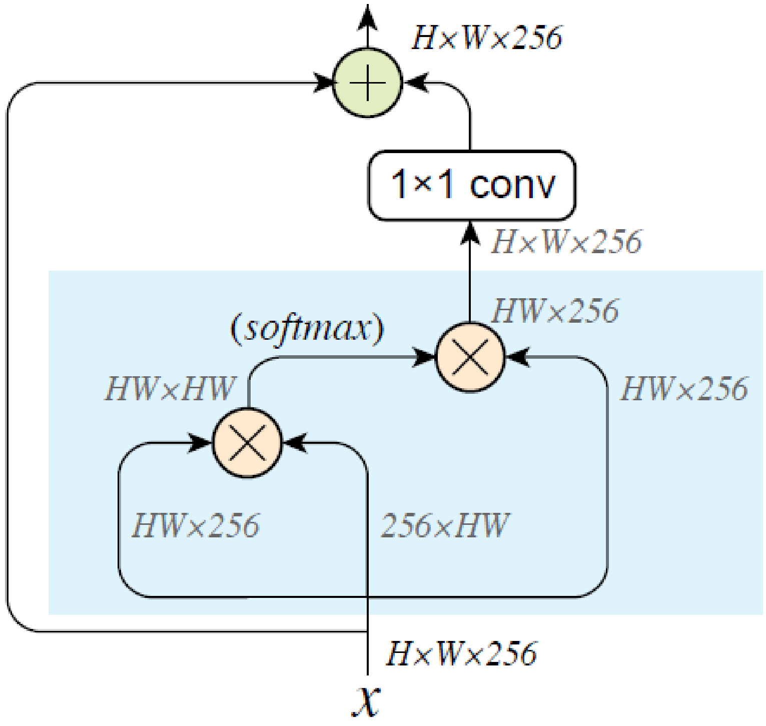
Feature-dependent weighting

$$y_i = \frac{1}{C(x)} \sum_{\forall j \in \mathcal{L}} f(x_i, x_j) \cdot x_j$$

Normalization

Gaussian (softmax) sets : $f(x_i, x_j) = e^{\frac{1}{\sqrt{d}}\theta(x_i)^T \phi(x_j)}$

Dot product sets : $f(x_i, x_j) = x_i^T x_j$



Adversarial Training

- **PGD attacker**
    - 20% of training batches use clean image, and 80% use adversarially perturbed images
    - Use the Projected Gradient Descent (PGD) as the white-box attacker

- **Distributed training with adversarial images**
    - A single SGD update is preceded by n-step PGD, the total amount of computation in adversarial training is n times bigger than standard (clean) training
    - Using synchronized SGD on 128 GPUs (Nvidia V100), Each mini-batch contains 32 images per GPU
    - 52 hours for the baseline ResNet-152 model

Experimental Result

Against White-box Attacks

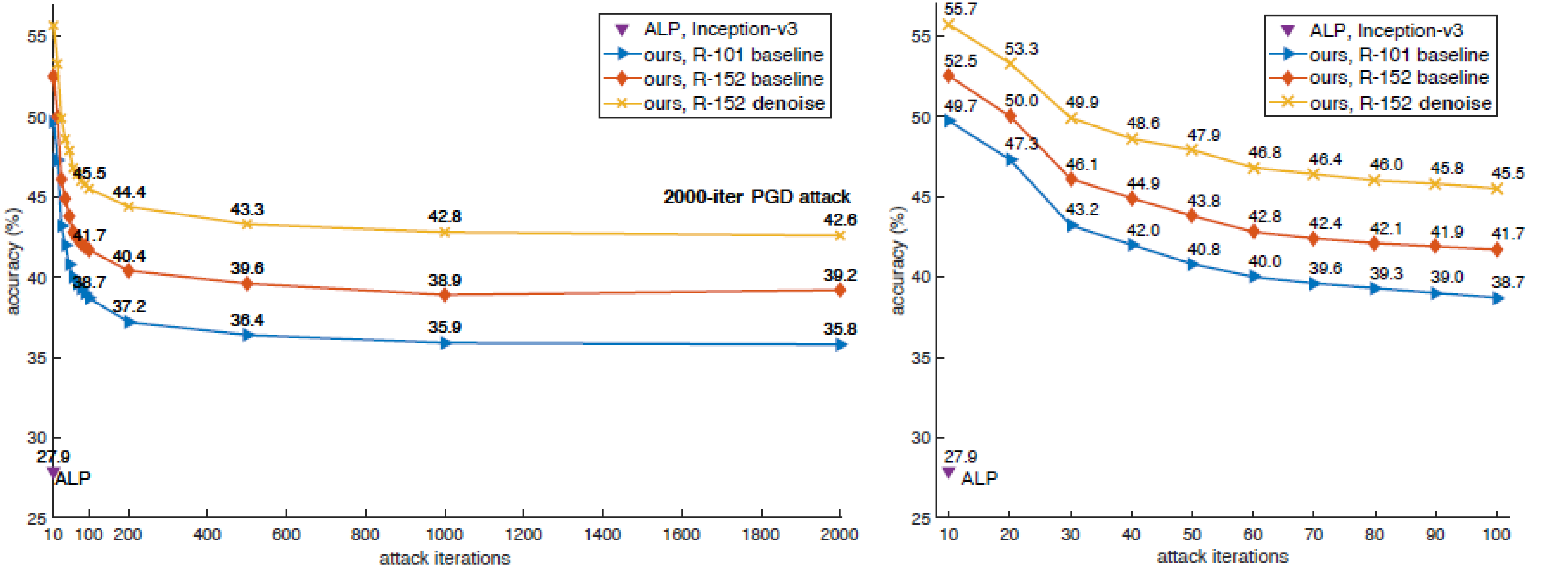
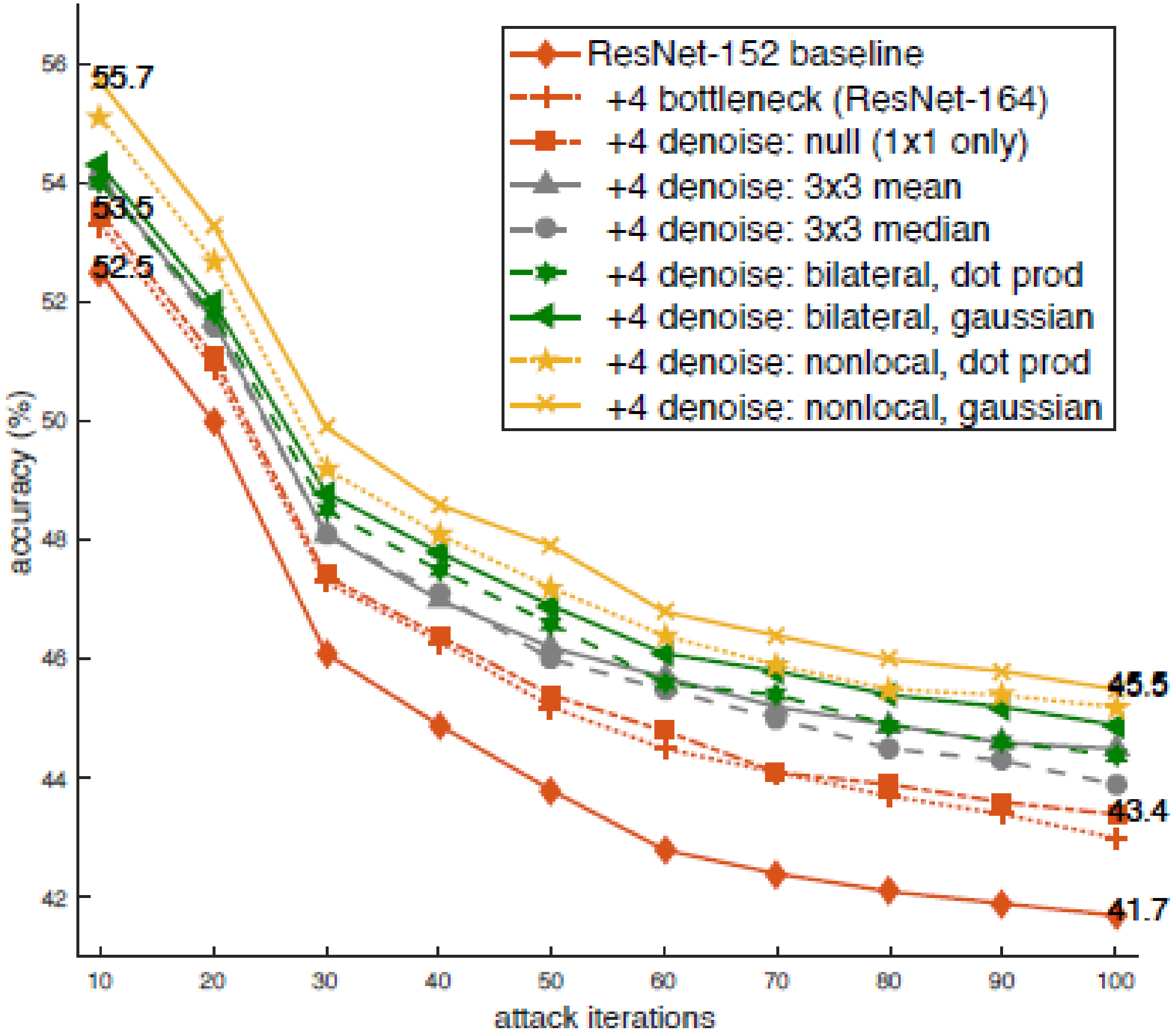


Figure 6. **Defense against white-box attacks on ImageNet**. The left plot shows results against a white-box PGD attacker with 10 to 2000 attack iterations. The right plot zooms in on the results with 10 to 100 attack iterations. The maximum perturbation is $\epsilon = 16$.

Experimental Result

Variants of denoising operations



Design decisions of the denoising block

attack iterations	10	100
non-local, Gaussian	55.7	45.5
removing 1×1	52.1	36.8
removing residual	NaN	NaN

Table 1. **Ablation: denoising block design** for defending against *white-box* attacks on ImageNet. Our networks have four (Gaussian) non-local means denoising blocks. We indicate the performance of models we were unable to train by "NaN".

- Denoising features in itself is not sufficient. As suppressing noise may also remove useful signals, it appears essential to properly combine the denoised features with the input features in denoising blocks.

Experimental Result

Against Black-Box Attacks

model	accuracy (%)
CAAD 2017 winner	0.04
CAAD 2017 winner, under 3 attackers	13.4
ours, R-152 baseline	43.1
+4 denoise: null (1×1 only)	44.1
+4 denoise: non-local, dot product	46.2
+4 denoise: non-local, Gaussian	**46.4**
+all denoise: non-local, Gaussian	**49.5**

Table 2. **Defense against black-box attacks on ImageNet.** We show top-1 classification accuracy on the ImageNet validation set. The attackers are the 5 best attackers in CAAD 2017. We adopt the CAAD 2018 "*all-or-nothing*" criterion for defenders. The 2017 winner has 0.04% accuracy under this strict criterion, and if we remove the 2 attackers that it is most vulnerable to, it has 13.4% accuracy under the 3 remaining attackers.

- "all-or-nothing" evaluation

  an image is considered correctly classified only if the model correctly classifies all adversarial versions of this image created by all attackers

Experimental Result

CAAD 2018 challenge results

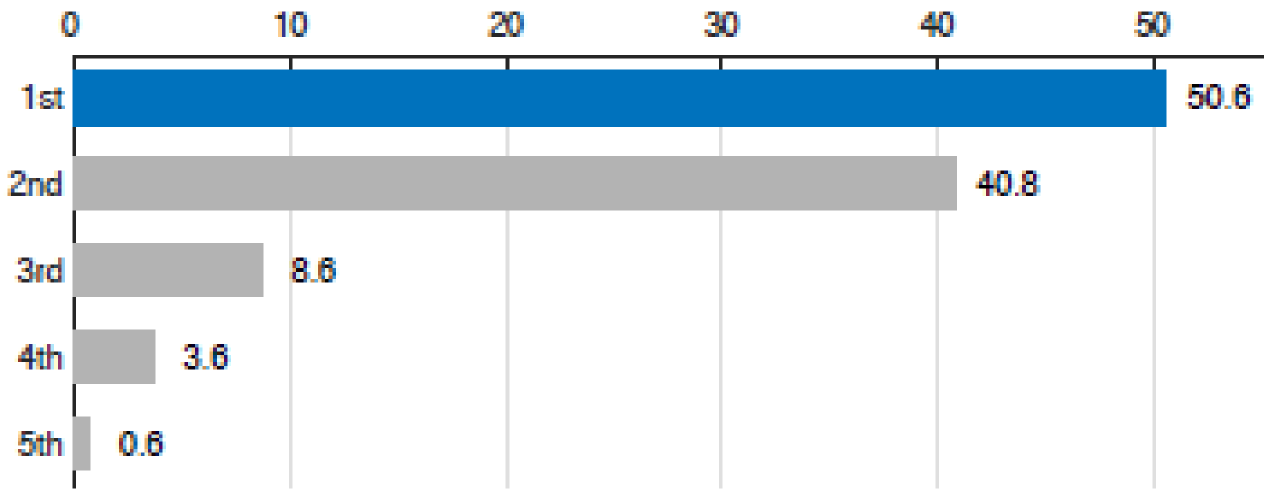


Figure 8. CAAD 2018 results of the adversarial defense track. The first-place entry is based on our method. We only show the 5 winning submissions here, out of more than 20 submissions.

- CAAD 2018 challenge – need to defend against 48 unknown attackers
- The winning model was based on using a ResNeXt-101 backbone with non-local denoising blocks added to all residual blocks

Experimental Result

Denoising Blocks in Non-Adversarial Settings

model	accuracy (%)
R-152 baseline	78.91
R-152 baseline, run 2	+0.05
R-152 baseline, run 3	-0.04
+4 bottleneck (R-164)	+0.13
+4 denoise: null (1×1 only)	+0.15
+4 denoise: 3×3 mean filter	+0.01
+4 denoise: 3×3 median filter	-0.12
+4 denoise: bilateral, Gaussian	+0.15
+4 denoise: non-local, Gaussian	+0.17

Table 3. **Accuracy on clean images** in the ImageNet validation set when trained on clean images. All numbers except the first row are reported as the accuracy difference comparing with the first R-152 baseline result. For R-152, we run training 3 times independently, to show the natural random variation of the same architecture. All denoising models show *no significant difference*, and are within ±0.2% of the baseline R-152's result.

- The denoising blocks could have special advantages in settings that require adversarial robustness
- ResNet-152 baseline with *adversarial training has 62.32% accuracy* when tested on *clean images*
- The tradeoff between adversarial and clean training to be the subject of future research

Q & A