Adversarial defense suggestion

Aurélien Fouque

17th July 2020

Defense presentation

Encryption inspired adversarial defense for visual classification (AprilPyone & Kiya 2020)

Reminder on attack and defense [1]:

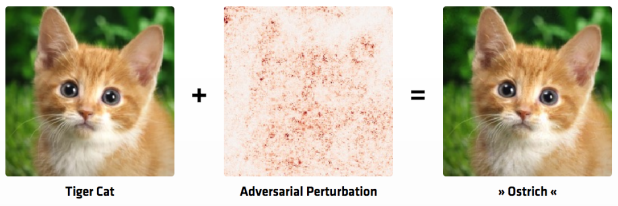
- original and modified images $(x; x')$
- given noise distance ε
- infinity norm $||u||_\infty = \max_i(|u_i|)$

$$\mathcal{D}(x; x') = ||x - x'||_\infty \leqslant \varepsilon \qquad (1)$$

- classifier $f(\cdot)$
- true and targeted classes $(y; z)$

$$f(x') \neq y \text{ or } f(x') = z \qquad (2)$$

Example [2]:

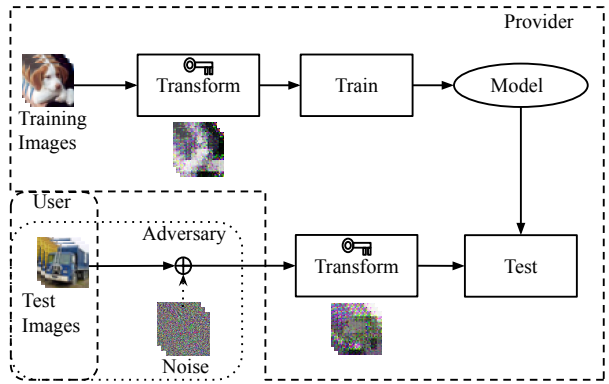


- defensive transform $t(\cdot)$

$$f\left(t(x')\right) = f\left(t(x)\right) = f(x) = y \qquad (3)$$

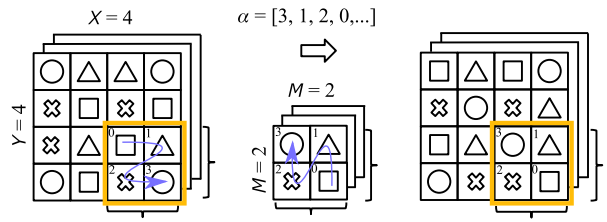
Overview of the defense [3]:

- transformation before both training and testing



Shuffling block-wise pixels:

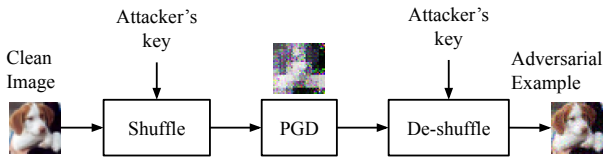
- size M, number of pixels $n = M^2$, number of possibilities $n!$
- random permutation vector α



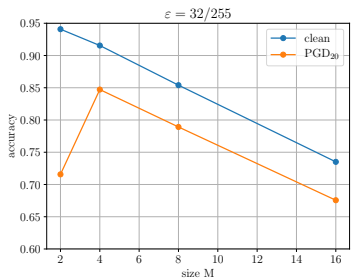
Interest for aidkit.ai

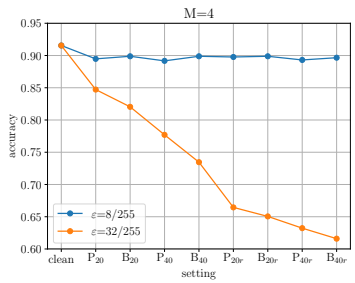
Attacks deployed, settings and results:

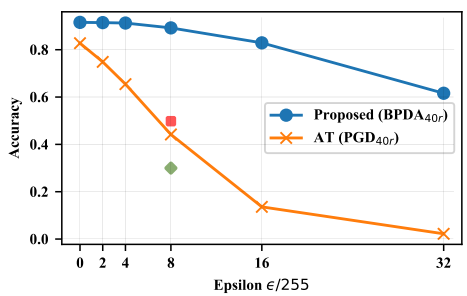
- "first order": projected gradient descent [4]
- adaptive: $\sim$ backward pass differentiable approximation [5]



- $\{20; 40\}$ iterations, with[out] random initialization







Good points:

- method seems to be robust
- $M = 4 \implies 2.09 \cdot 10^{13}$ possibilities
- good accuracy on test images
- recent publication
- easy principle and implementation

Axes for improvement:

- trying other M values $\in [5; 7]$
- using rectangles or other shapes
- shuffling independently RGB colours with different α values, trying >8 bit colour images
- using other norms and with grey/black boxes
- trying other databases than CIFAR-10 [6]

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus.
Intriguing properties of neural networks.
2013, arXiv:1312.6199 [cs.CV].

[2] DNN are not robust.
https://robust.vision/benchmark/about/.
Accessed: 13th July 2020.

[3] MaungMaung AprilPyone and Hitoshi Kiya.
Encryption inspired adversarial defense for visual classification.
2020, arXiv:2005.07998 [cs.LG].

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu.
Towards deep learning models resistant to adversarial attacks.
2017, arXiv:1706.06083 [stat.ML].

[5] Anish Athalye, Nicholas Carlini, and David Wagner.
Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples.
2018, arXiv:1802.00420 [cs.LG].

[6] Alex Krizhevsky.
Learning multiple layers of features from tiny images.
Master's thesis, University of Toronto, April 2009.
pdf.