DeepCloak: Masking Deep Neural Network Models for Robustness Against Adversarial Samples

Ji Gao, Beilun Wang, Zeming Lin, Weilin Xu, Yanjun Qi
Department of Computer Science, University of Virginia

Summary:

- Unnecessary features in the deep neural networks make the model vulnerable.
- Defend adversarial samples by removing unnecessary features.
- An efficient approach to remove unnecessary features without retraining the model.

Motivation:

Adversarial samples:

- Adversarial samples: deliberately generated samples to fool DNN classifiers.
- An adversarial sample x^′ can be defined as:

$x' = x + \Delta x, |\Delta x| < \epsilon$

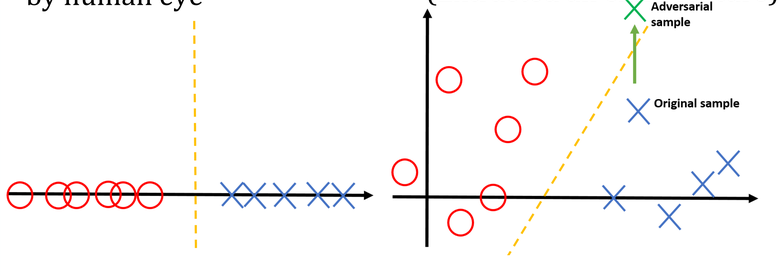
$F(x) \neq F(x')$

- An adversarial sample must be similar to its seed sample.
- Adversarial samples can greatly reduce the effectiveness of deep learning models.
- A Recent study [1] shows that extra unnecessary features extracted by the machine classifier are a vulnerability to adversarial samples.

Example of the vulnerability:

Truth, e.g., by human eye

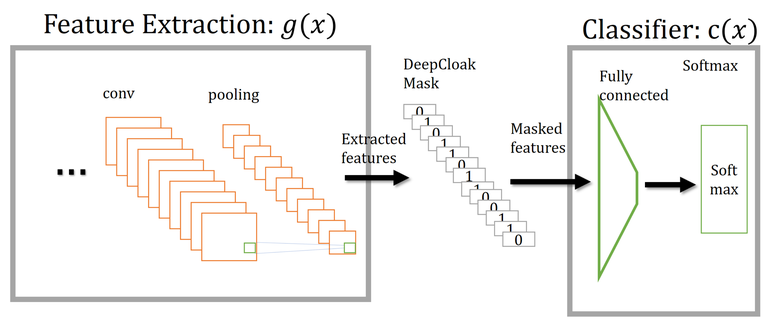
Machine Learning model (Extracted an extra feature)



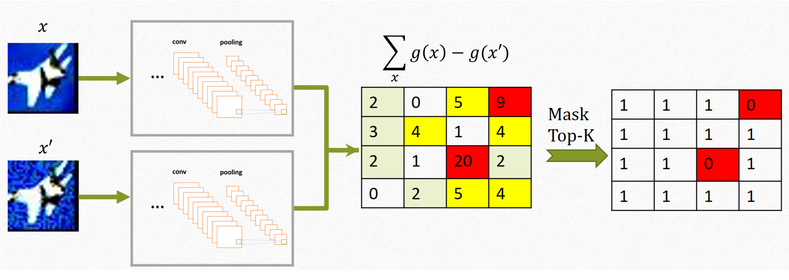
Method:

Key: Insert a mask layer in a DNN model right before the linear layer handling classification.

To use the mask: Model: $F(x)=g(c(x))$



The algorithm to learn the mask:

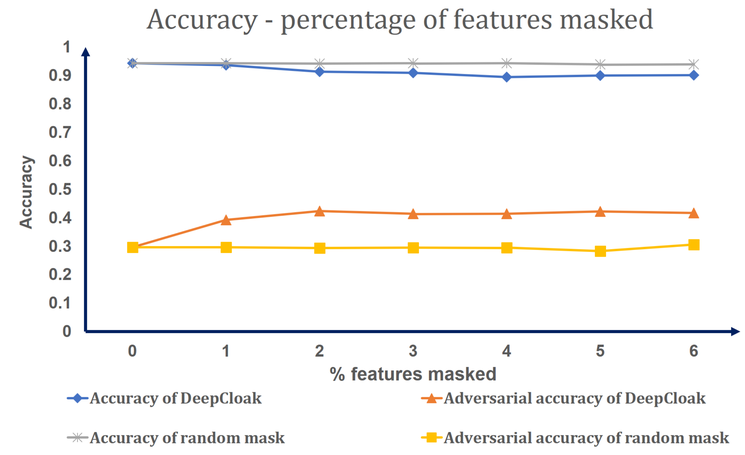


No retraining needed!

We've renamed our paper from **DeepMask** to **DeepCloak**.

Experiment result:

On Res-net152:



Get 10% increase with masking 1% nodes!

Reference:

[1] Wang Beilun, Ji Gao, and Yanjun Qi. "A Theoretical Framework for Robustness of (Deep) Classifiers Under Adversarial Noise." arXiv: 1612.00334 (2016).

[2] Gao, Ji, Beilun Wang, and Yanjun Qi. "DeepMask: Masking DNN Models for robustness against adversarial samples." arXiv preprint arXiv: 1702.06763 (2017).

Next: On other layers.