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: $x' = x + \Delta x$, $|\Delta x| < \epsilon$.
- 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)

Motivation:

Adversarial samples:

- Adversarial samples: deliberately generated samples to fool DNN classifiers.
- An adversarial sample $x'$ can be defined as: $x' = x + \Delta x$, $|\Delta x| < \epsilon$.
- 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.

Experiment result:

On Res-net152:

Get 10% increase with masking 1% nodes!

Reference:


Next: On other layers.