Existing Adversarial Attack Bound is Hard to Scale

- **Pixel Space**
  - Attack fails given a small bound; Attack is detectable on a larger bound.

- **Feature Space**
  - Attack is detectable on samples with high confidence.

Method - Quantile Bound on Gaussian Representation

- Throttle plane (TP) selection
- Internal distribution boundary constraint
- Adversarial sample generation with combined losses

Distribution of the Representation Matters

Which representation should we use for quantile bound? Look for Gaussian.

Summary

- Identify the scalability problem of existing adversarial attacks.
- Proposed the quantile bound on deep content features.
- Proposed an efficient way for optimizing the adversarial samples.
- Show the state-of-the-art trade-off between imperceptibility and attack success using the quantile bound.

Contact

- Email: xu1230@purdue.edu
- Home Page: Original

Bounded Adversarial Attack on Deep Content Features
Qiuling Xu, Guanhong Tao and Xiangyu Zhang @ Purdue University