Overcomplete Representations Against Adversarial Videos

ICIP 2021

Shao-Yuan Lo, Jeya Maria Jose Valanarasu, Vishal M. Patel
Johns Hopkins University
Recall: Adversarial Examples

\[ x_{adv} = x + \delta \]

\[ f(x_{adv}) \neq y \]
Recall: Adversarial Examples

• Deep networks are **vulnerable** to adversarial examples.

Goodfellow et al. Explaining and Harnessing Adversarial Examples. ICLR’15.
Adversarial Videos

• Video is a stack of consecutive images.

• A naïve way to generate adversarial videos:
  Use image-based method directly.

\[ x^{adv} = x + \epsilon \cdot \text{sign}(\nabla_x L(x, y; \theta)) \]

*Image*: \( x \in \mathbb{R}^{C \times H \times W} \)

*Video*: \( x \in \mathbb{R}^{F \times C \times H \times W} \)
Feature Denoising

• Remove adversarial perturbations in the feature domain instead of the image domain.

• Mean filter, median filter, bilateral filter, and non-local means.
Proposed Method: Overcomplete Representations

• A typical autoencoder downsamples features and learns undercomplete representations.

• OUDefend learns both undercomplete representations and overcomplete representations (upsample features)
Proposed Method: Overcomplete Representations

• **Undercomplete** representations have large receptive fields to collect global information, but they overlook local details.

• **Overcomplete** representations have opposite properties.

• OUDefend balances **global** and **local** features by learning those two representations.
Proposed Method: Overcomplete Representations

- Append OUDefend blocks to the target network (after each res block).
Adversarial Video Types

- PGD [Madry et al. ICLR’18]
- MultAV (Multiplicative Adversarial Video) [Lo et al. 2020]
- ROA (Rectangular Occlusion Attack) [Wu et al. ICLR’20]
- AF (Adversarial Framing) [Zajac et al. AAAI’19]
- SPA (Salt-and-Pepper Noise Attack) [Lo et al. 2020]
Experimental Results

- No Defense: Original network trained on clean data
- Madry [Madry et al. ICLR’18]: Original network trained by adversarial training (AT)
- Xie-A [Xie et al. CVPR’19]: Feature denoising (3D conv) network with AT
- Xie-B [Xie et al. CVPR’19]: Feature denoising (2D conv frame-by-frame) network with AT
- OUDefend: Proposed OUDefend network with AT

| Method     | #Params | Clean | PGD Linf | PGD L2 | MultAV | ROA   | AF   | SPA  | Avg_adv |
|------------|---------|-------|----------|--------|--------|-------|------|------|---------|
| No Defense | 33.0M   | 76.90 | 2.56     | 3.25   | 7.19   | 0.16  | 0.24 | 4.39 | 2.97    |
| Madry      | 33.0M   | 76.90 | 33.94    | 35.05  | 47.00  | 41.29 | 55.99| 55.99| 48.01   |
| Xie-A      | 33.7M   | 70.82 | 31.48    | 33.25  | 42.69  | 37.59 | 58.87| 49.14| 42.17   |
| Xie-B      | 34.8M   | 69.47 | 30.19    | 32.65  | 41.87  | 38.22 | 58.74| 49.14| 41.80   |
| OUDefend   | 33.6M   | **77.90** | **34.18** | **35.32** | **47.63** | **42.00** | **56.25** | **56.29** | **49.52** |

Dataset: UCF-101
Feature Visualization

PGD input  No defense  OUDefend  AF input  No defense  OUDefend
Conclusion

• Exploit both undercomplete and overcomplete representations
• Evaluate on 6 different attacks
• Show effectiveness with very small complexity increase