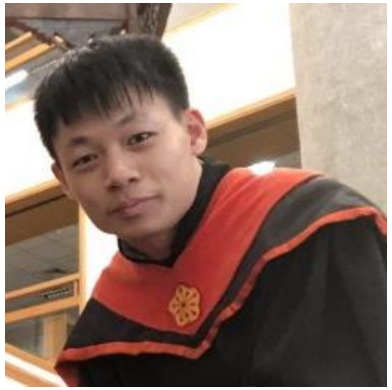
Overcomplete Representations Against Adversarial Videos

ICIP 2021



Shao-Yuan Lo, Jeya Maria Jose Valanarasu, Vishal M. Patel

Johns Hopkins University

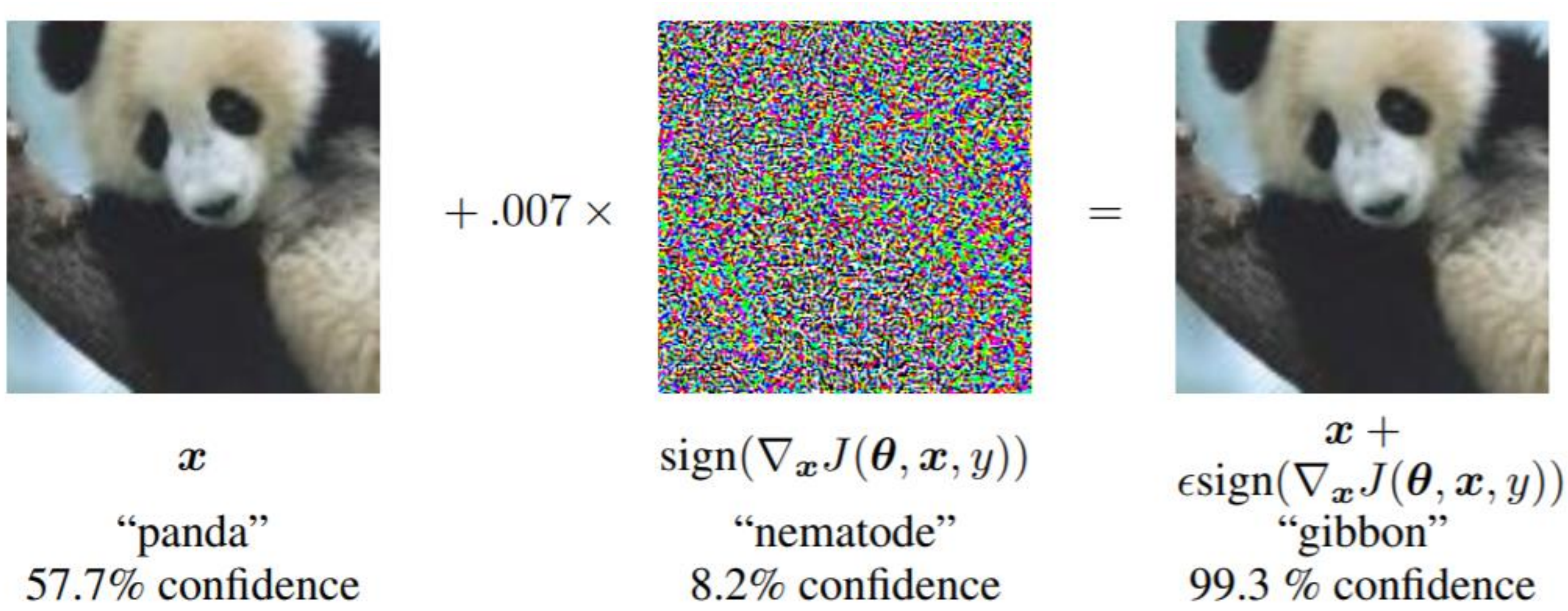
Recall: Adversarial Examples

$$x_{adv} = x + \delta$$

$$f(\boldsymbol{x}_{adv}) \neq y$$

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.

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

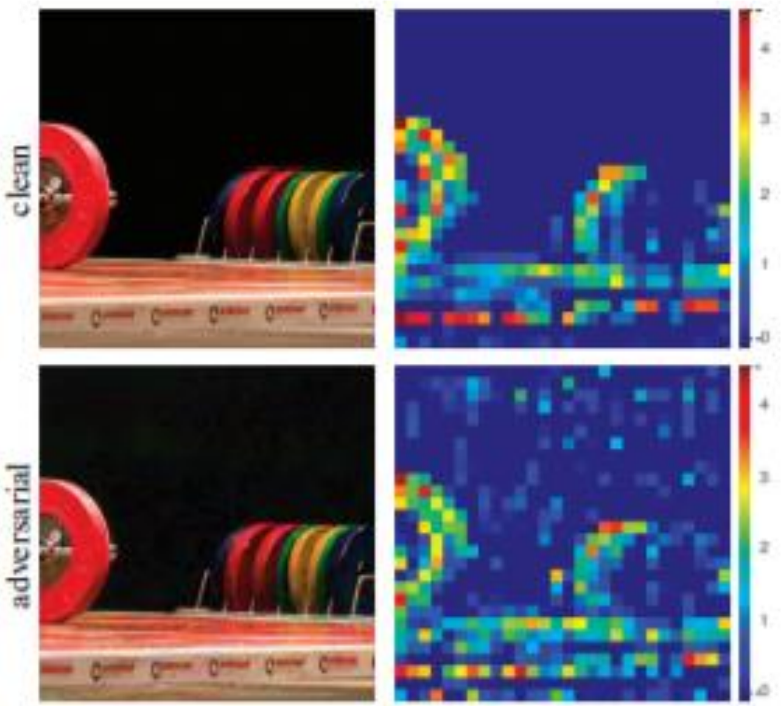
$$Image\colon\ x \in R^{C \times H \times W}$$

$$Video\colon\ x \in R^{\boldsymbol{F} \times C \times H \times W}$$

Feature Denoising

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

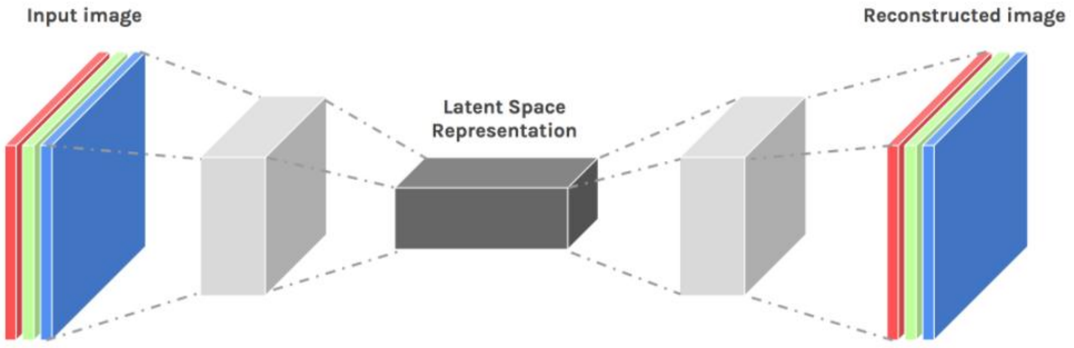
- Mean filter, median filter, bilateral filter, and non-local means.



Xie et al. Feature Denoising for Improving Adversarial Robustness. CVPR’19.

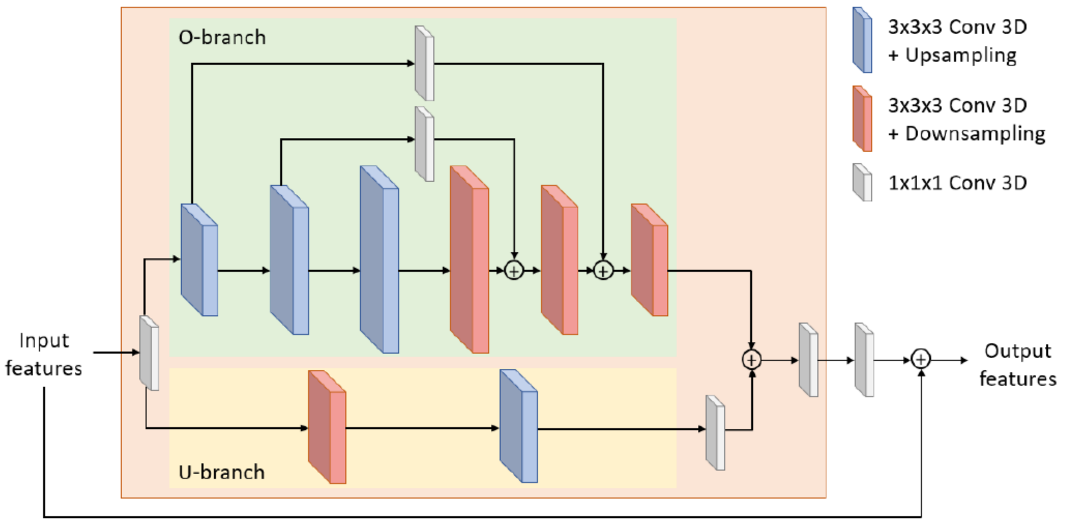
Proposed Method: Overcomplete Representations

- A typical autoencoder downsamples features and learns **undercomplete** representations.



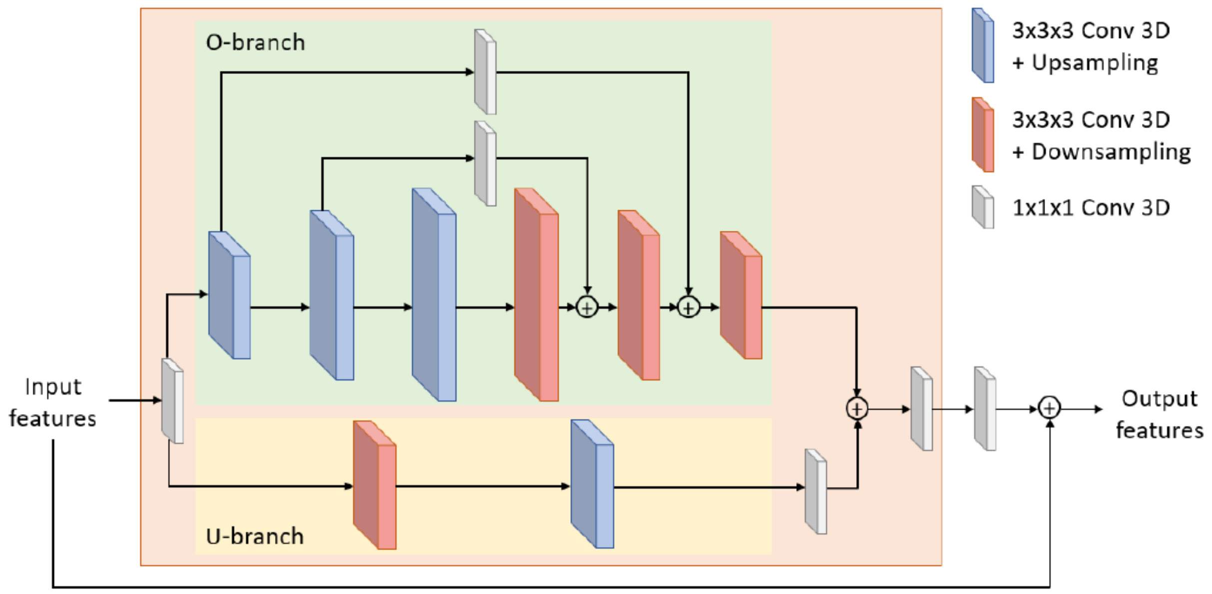
https://ai.plainenglish.io/convolutional-autoencoders-cae-with-tensorflow-97e8d8859cbe.

- OUDefend learns both **undercomplete** representations and **overcomplete** representations (upsample features)



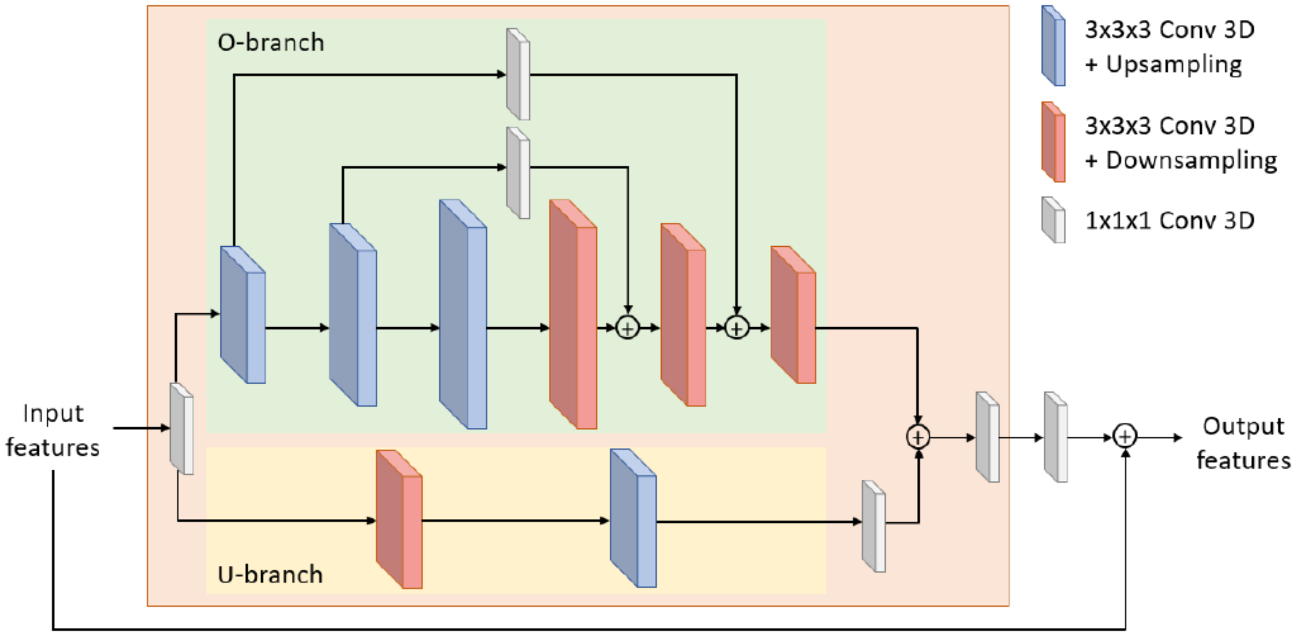
Proposed Method: Overcomplete Representations

- **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).



layer name	output size	18-layer
conv1	112×112	
conv2_x	56×56	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$

He et al. Deep Residual Learning for Image Recognition. CVPR'16.

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]

Clean PGD MultAV ROA AF SPA

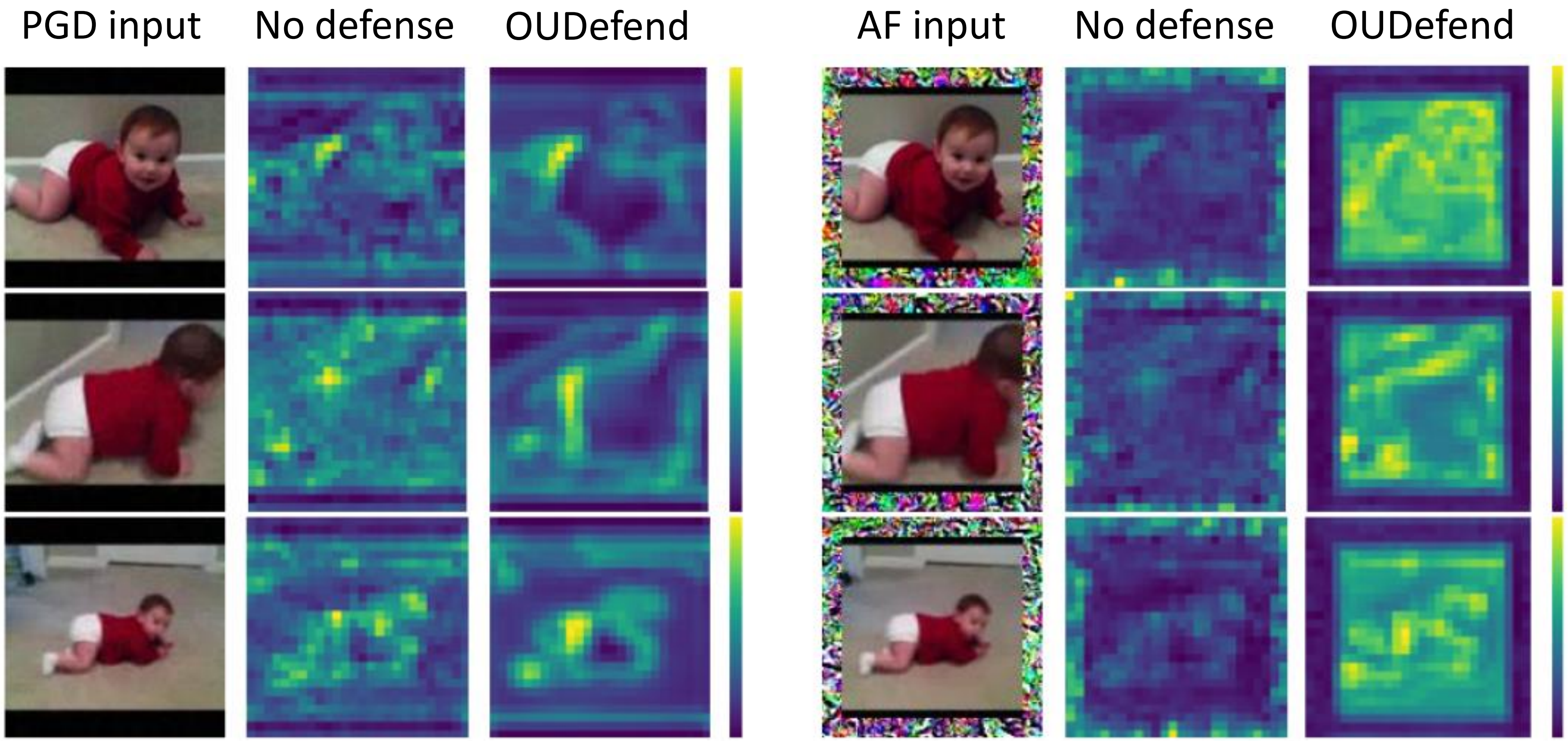
Experimental Results

Dataset: UCF-101

- 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**

Feature Visualization



Conclusion

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