LCANets: Lateral Competition Improves Robustness Against Corruption and Attack

Michael Teti
GRA, Los Alamos National Lab
Doctoral Candidate, Florida Atlantic University

mteti@lanl.gov

LA-UR-22-26490

Our Contributions

- Develop CNNs with sparse coding frontends called LCANets
- Competitive clean accuracy on action and image recognition
- SOTA robustness to corruptions and noise
- Perform first attacks with full knowledge of a sparse coding CNN layer
- Show how LCA frontends can augment robustness of adversarial training

Current CNNs Are Less Robust Than Biological Vision

- CNNs are often viewed as a rough model of biological object recognition [1]
- Previous work developed CNNs with biologically-motivated frontends [2]
  - Required collection and analysis of neurophysiological recordings
  - Left out sparsity and lateral competition observed in V1 [3, 4]

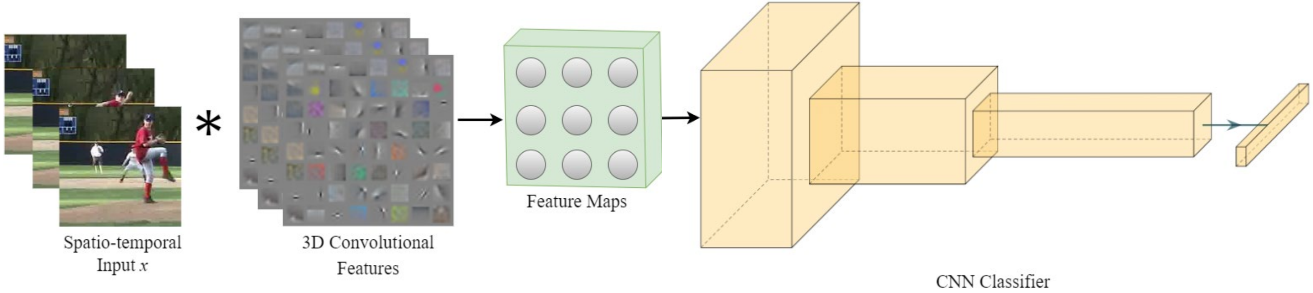
[1] Kubilius et al. 2019. Brain-Like Object Recognition With High Performing Shallow Recurrent ANNs.
[2] Dapello et al. 2020. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations.
[3] Yoshida and Ohki. 2020. Natural images are reliably represented by sparse and variable populations of neurons in visual cortex.
[4] Chettih and Harvey. 2019. Single-neuron perturbations reveal feature-specific competition in V1.

Sparse Coding CNN Layers

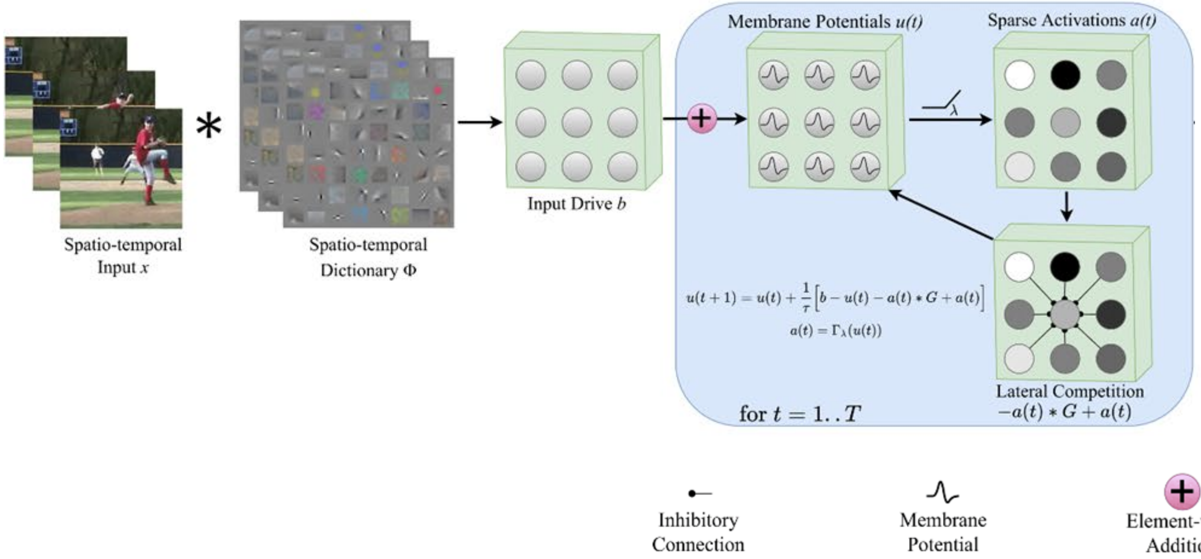
- Sparsity has been theorized to increase robustness of CNN layers [1, 2]
- Sparse coding frontends have been used to filter out noise and adversarial attacks computed on standard CNNs [2, 3, 4, 5]
  - Encode and reconstruct input image before classification by the CNN
  - Attacks had no or little knowledge of sparse coding layer
  - Not much comparison to other robust methods

[1] Subutai and Scheinkman. 2019. How can we be so dense? The benefits of using highly sparse representations.
[2] Paiton et al. 2020. Selectivity and robustness of sparse coding networks.
[3] Nguyen et al. 2020. Using models of cortical development based on sparse coding to discriminate between real and synthetically generated faces.
[4] Sun et al. 2019. Adversarial defense by stratified convolutional sparse coding.
[5] Kim et al. 2019. A neuromorphic sparse coding defense to adversarial images.

Standard CNN Architecture

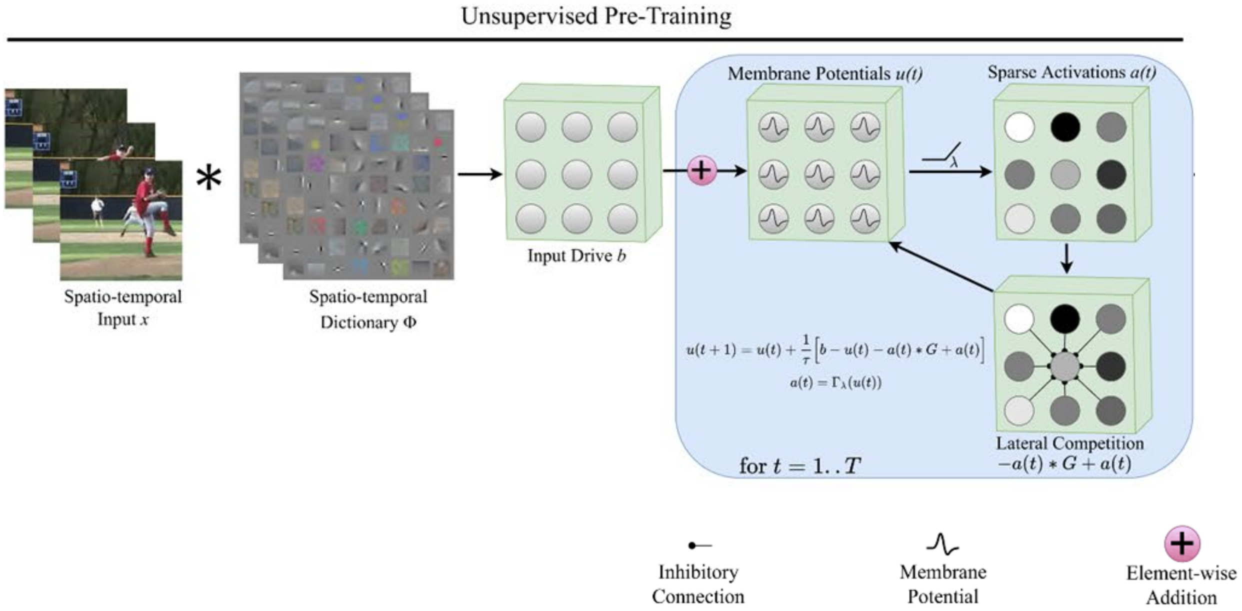


LCANet Architecture

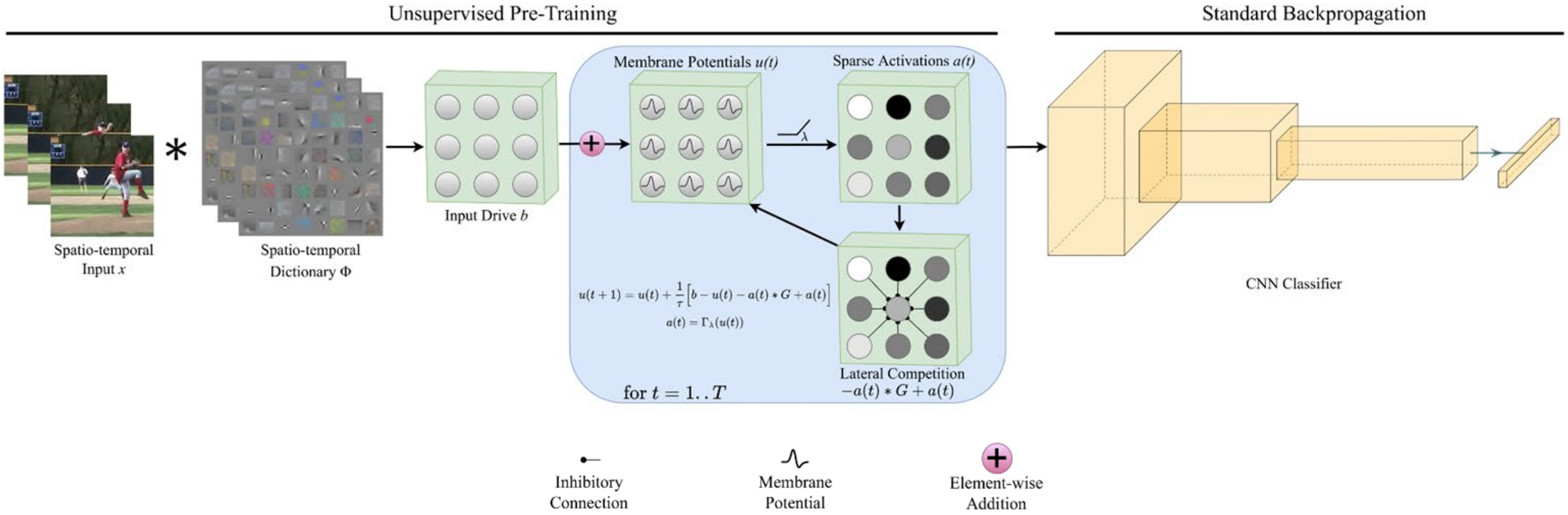

Spatio-temporal Input x
*
Spatio-temporal Dictionary Φ
Input Drive b
Membrane Potentials $u(t)$
Sparse Activations $a(t)$
λ
$u(t+1) = u(t) + \frac{1}{\tau}\left[b - u(t) - a(t) * G + a(t)\right]$
$a(t) = \Gamma_\lambda(u(t))$
for $t = 1..T$
Lateral Competition
$-a(t) * G + a(t)$
Inhibitory Connection
Membrane Potential
Element-wise Addition

- u(t) evolves over time
- Charged up/down by feature alignment with input
- Inhibited by neighboring active neurons
- Thresholded to compute sparse code

LCANet Architecture


Unsupervised Pre-Training
Spatio-temporal Input x
*
Spatio-temporal Dictionary Φ
Input Drive b
Membrane Potentials $u(t)$
Sparse Activations $a(t)$
λ
$u(t+1) = u(t) + \frac{1}{\tau}\left[b - u(t) - a(t) * G + a(t)\right]$
$a(t) = \Gamma_\lambda(u(t))$
for $t = 1..T$
Lateral Competition
$-a(t) * G + a(t)$
Inhibitory Connection
Membrane Potential
Element-wise Addition

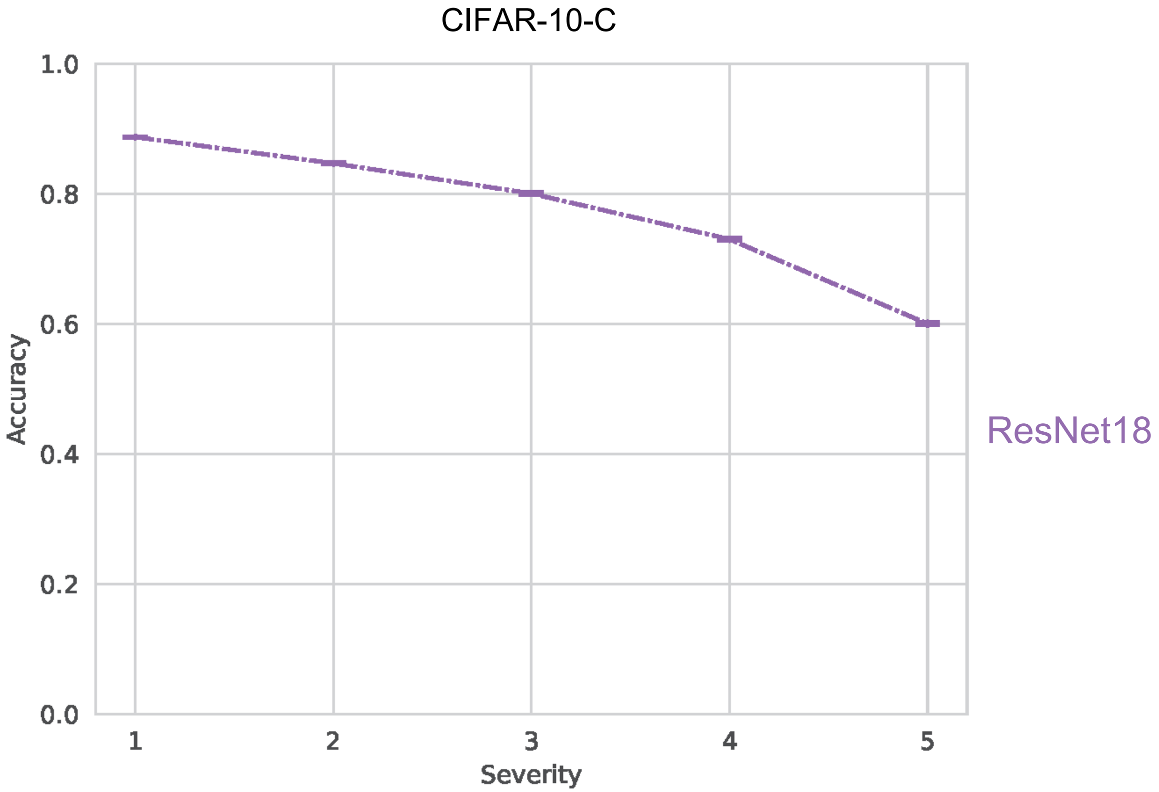
LCANet Architecture


Unsupervised Pre-Training
Standard Backpropagation
Spatio-temporal Input x
$*$
Spatio-temporal Dictionary Φ
Input Drive b
Membrane Potentials $u(t)$
Sparse Activations $a(t)$
λ
$u(t+1) = u(t) + \frac{1}{\tau}\left[b - u(t) - a(t) * G + a(t)\right]$
$a(t) = \Gamma_\lambda(u(t))$
for $t = 1..T$
Lateral Competition
$-a(t) * G + a(t)$
CNN Classifier
Inhibitory Connection
Membrane Potential
Element-wise Addition

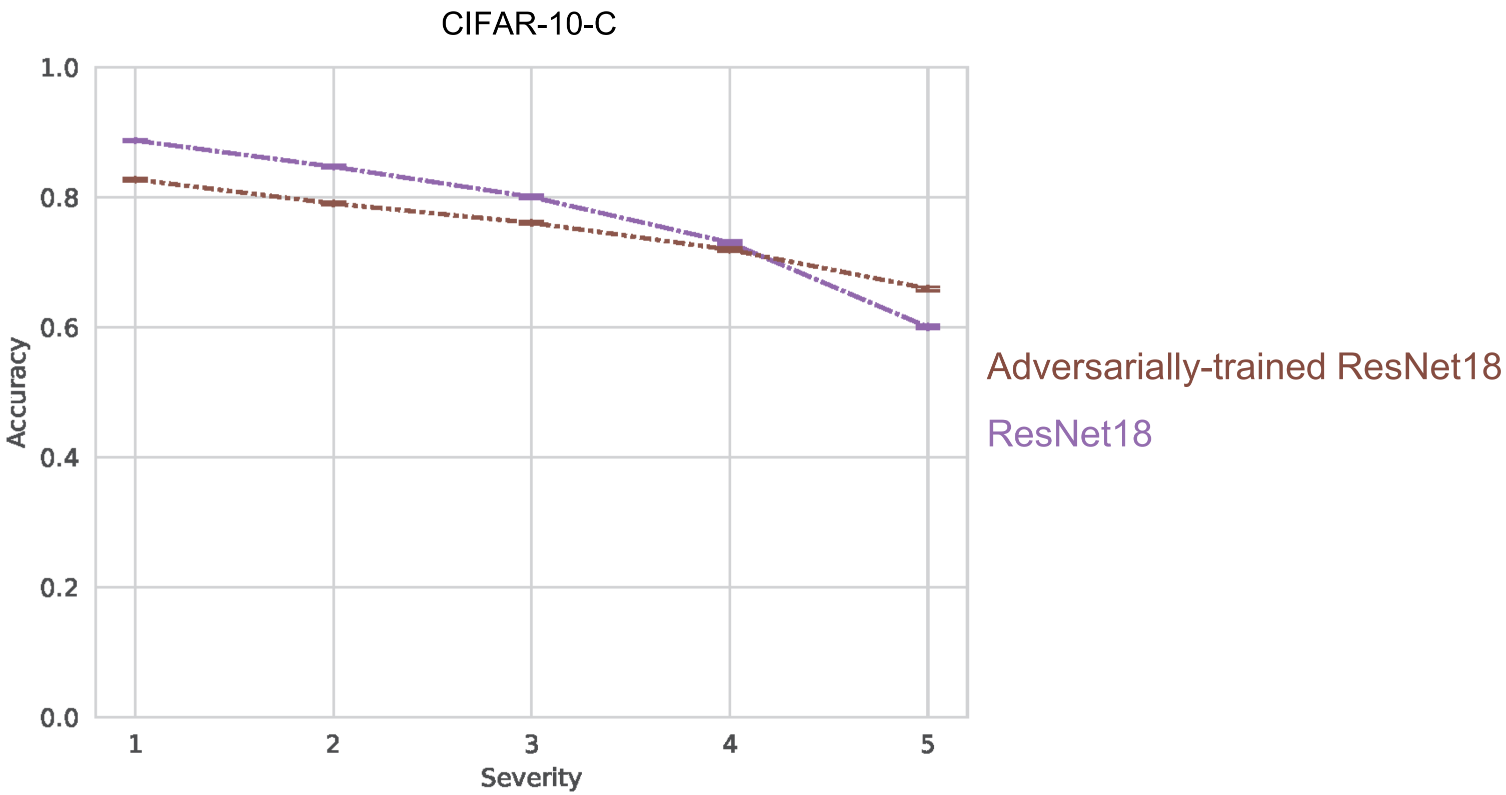
Tasks

- Action Recognition
  - UCF-101
  - HMDB-51
- Image Recognition
  - CIFAR-10
  - CIFAR-10-C

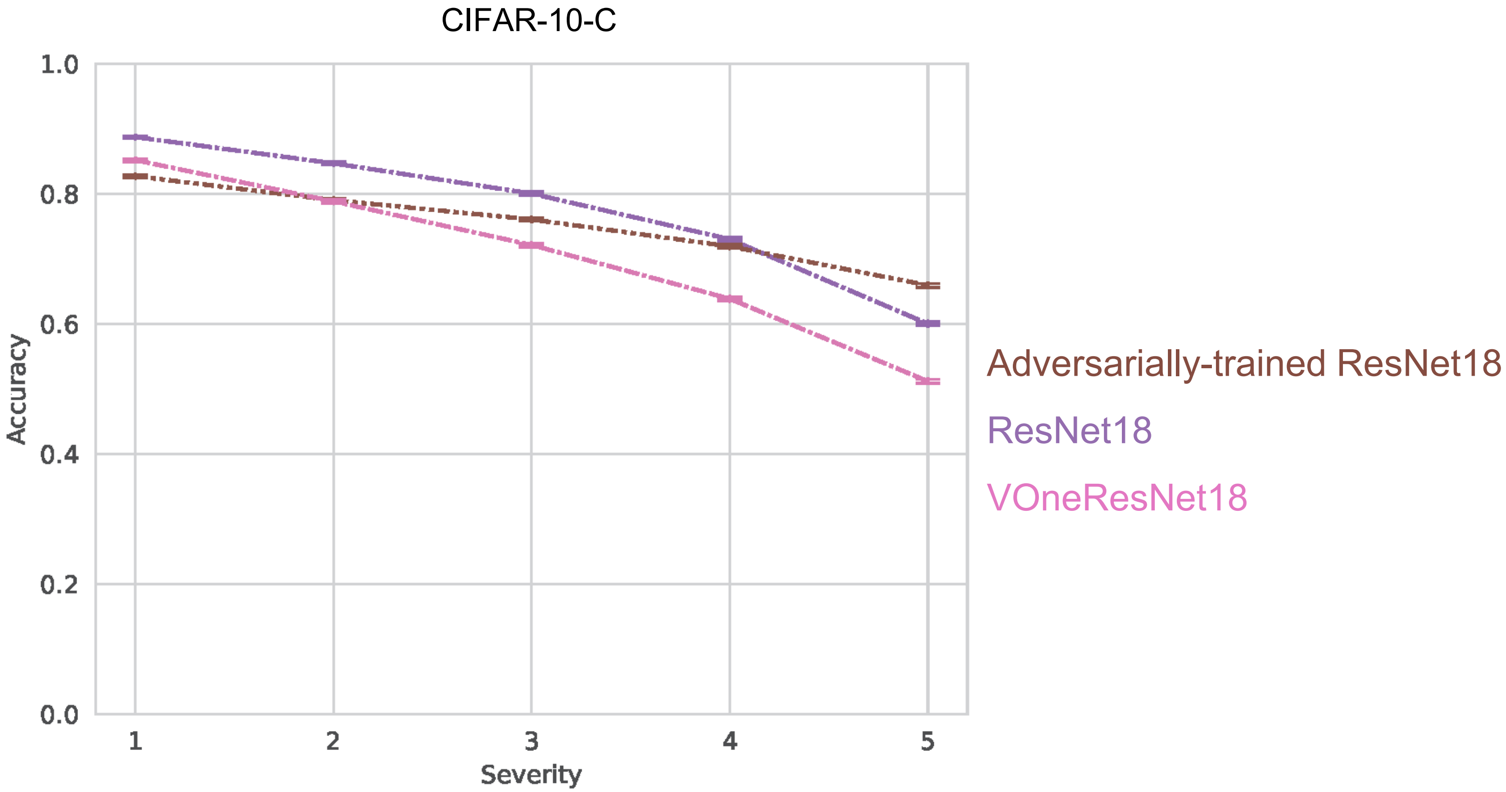
LCANets Are Robust to Corruptions



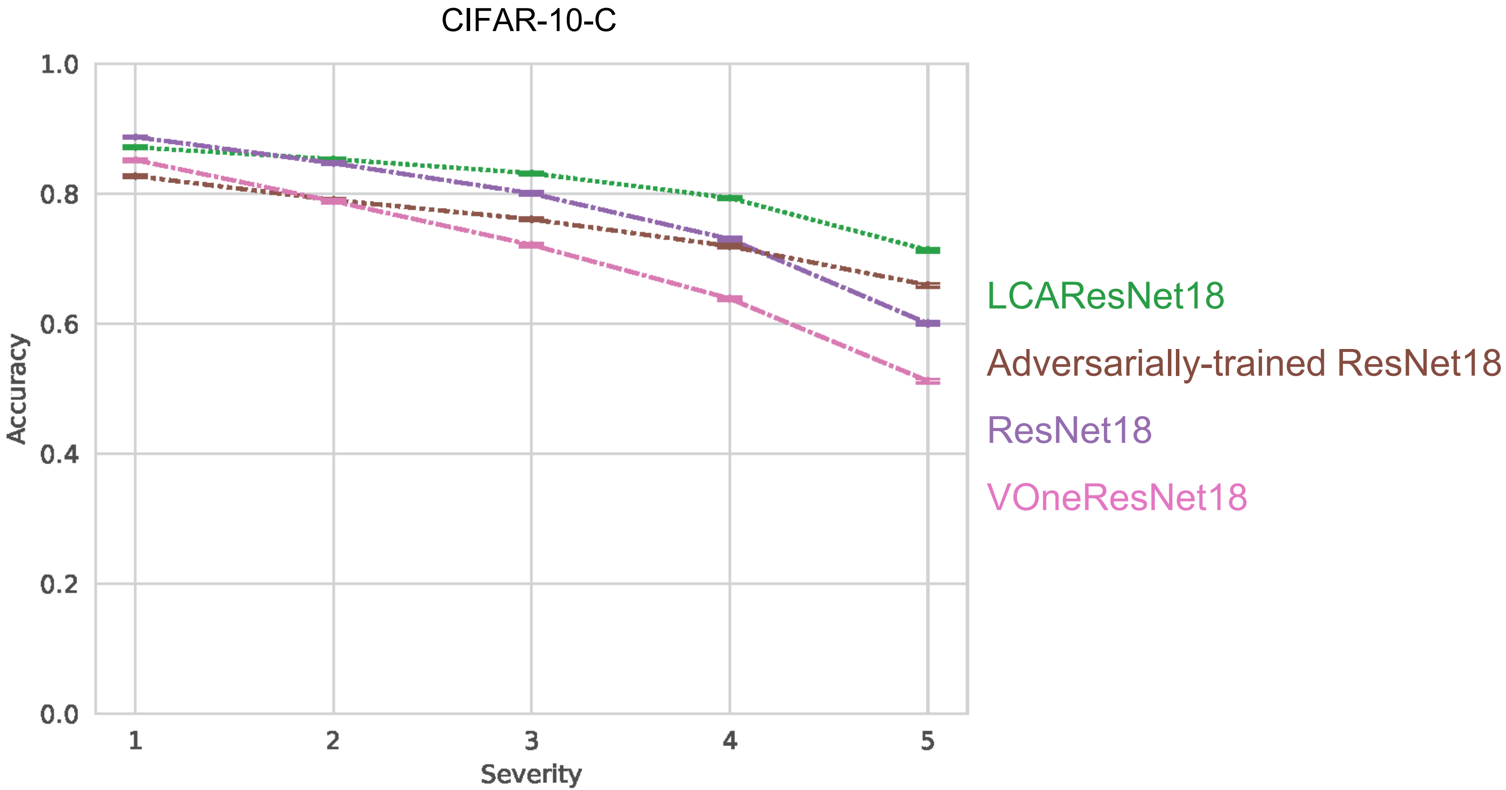
LCANets Are Robust to Corruptions



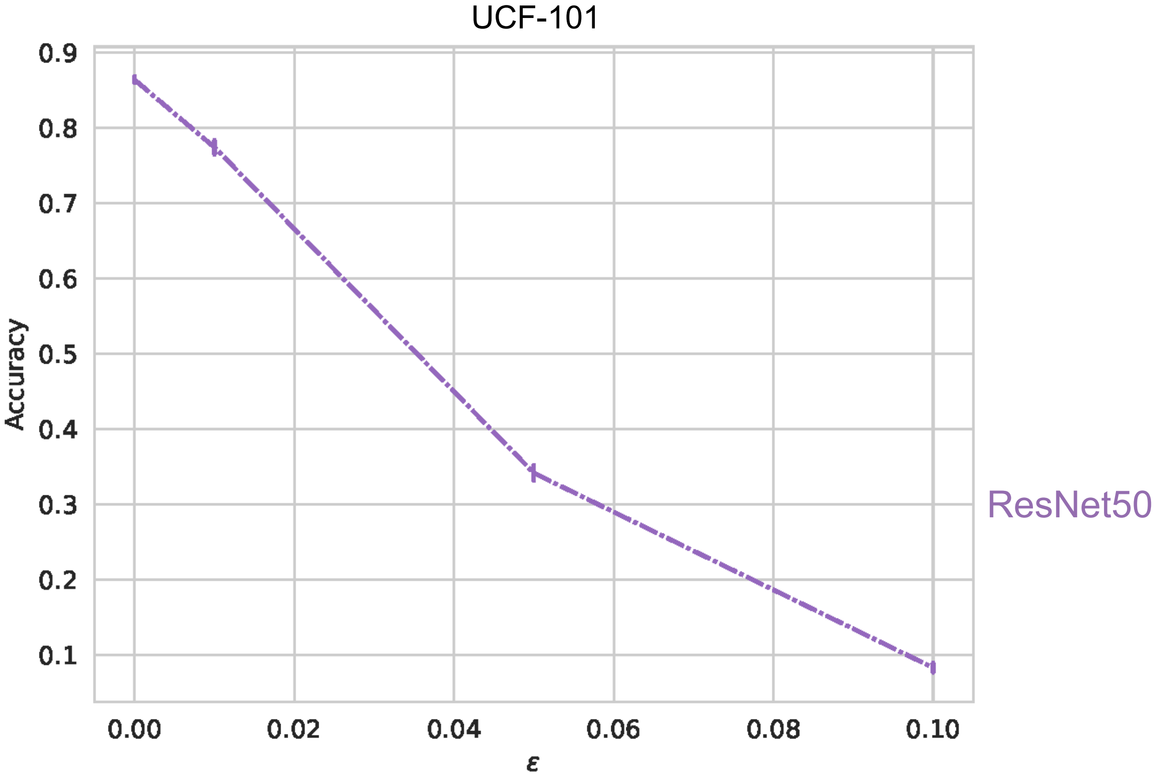
LCANets Are Robust to Corruptions



LCANets Are Robust to Corruptions

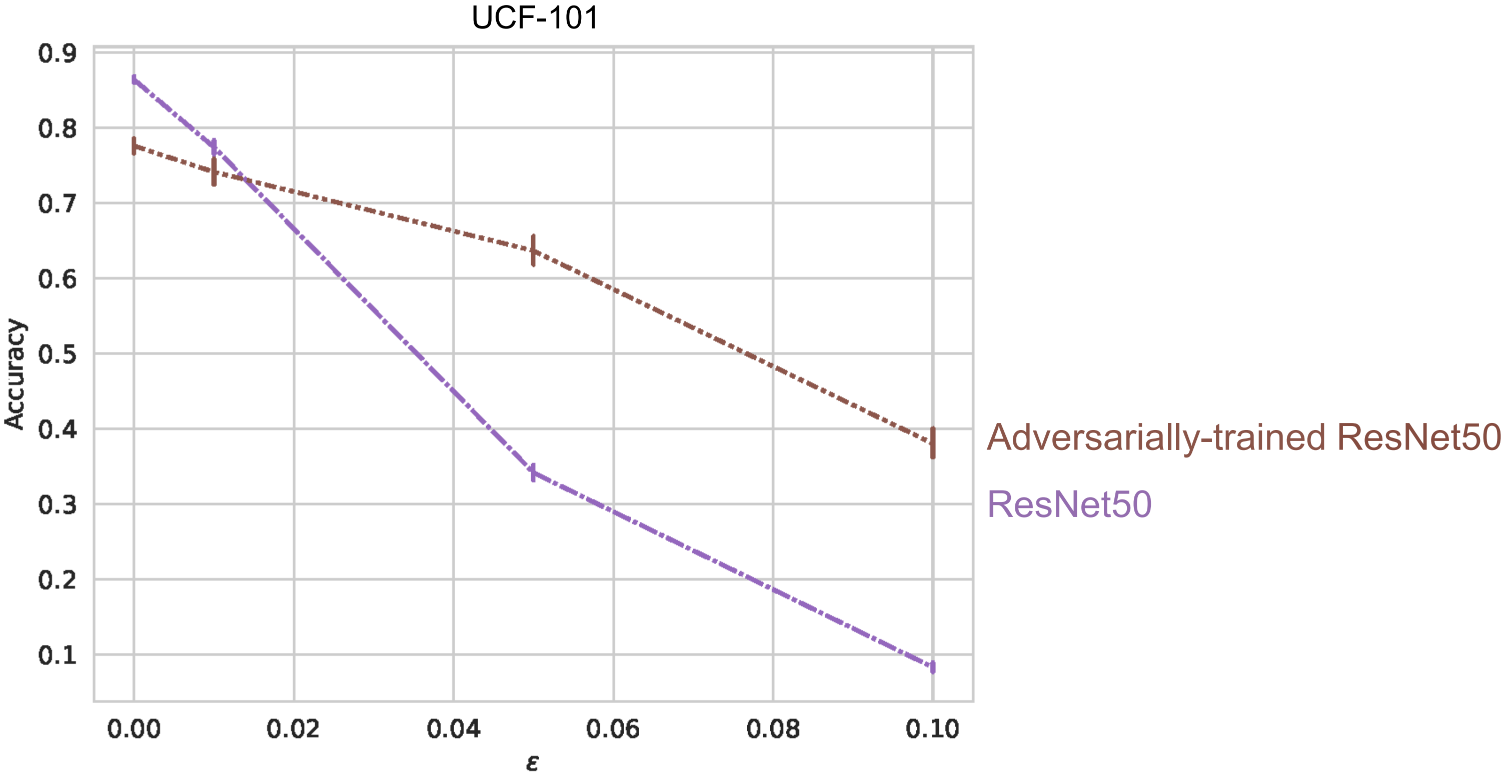


LCANets Are Competitive Under a Black-Box Attack



Ilyas et al. 2018. Prior convictions: Black-box adversarial attacks with bandits and priors. *ICLR 2018.*

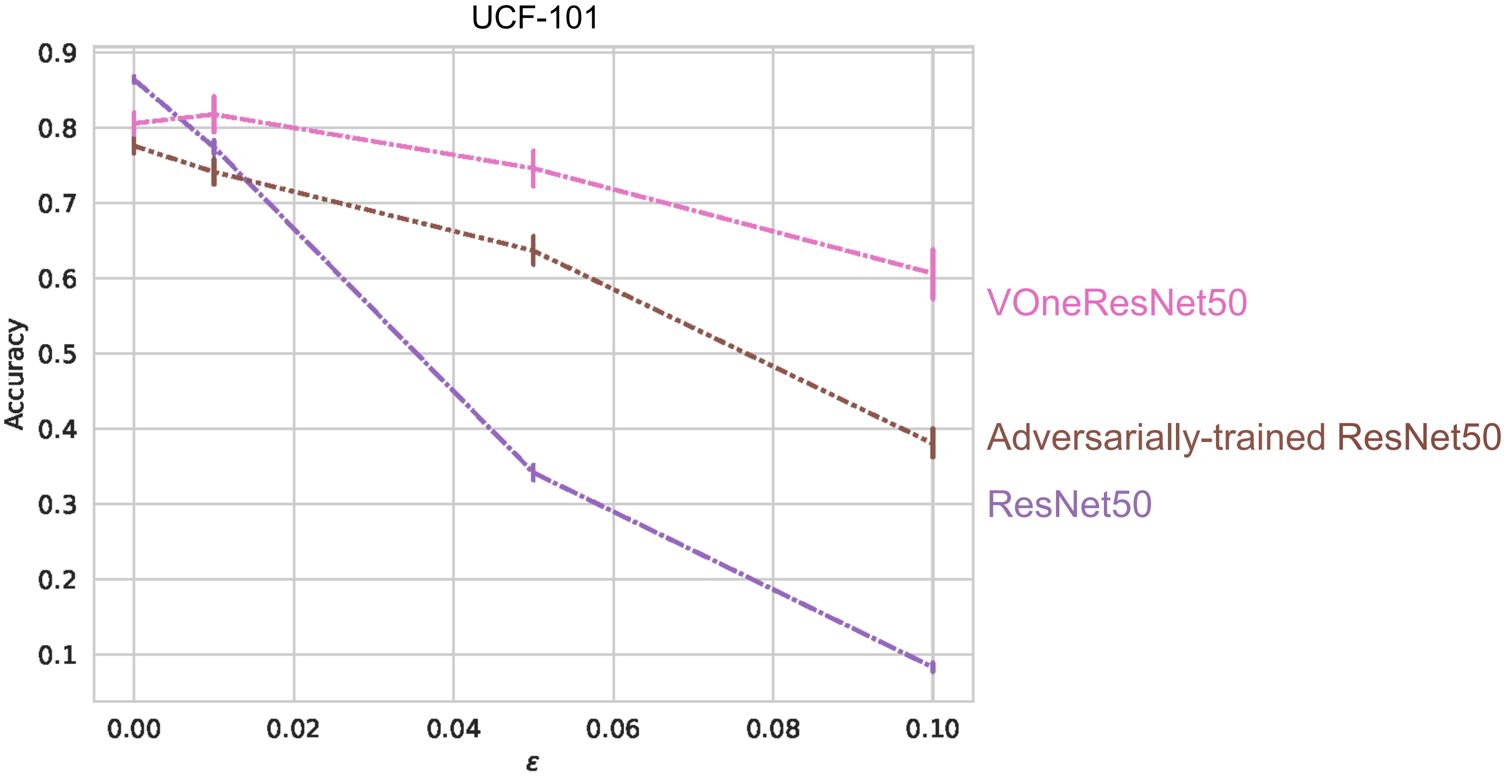
LCANets Are Competitive Under a Black-Box Attack



Los Alamos NATIONAL LABORATORY

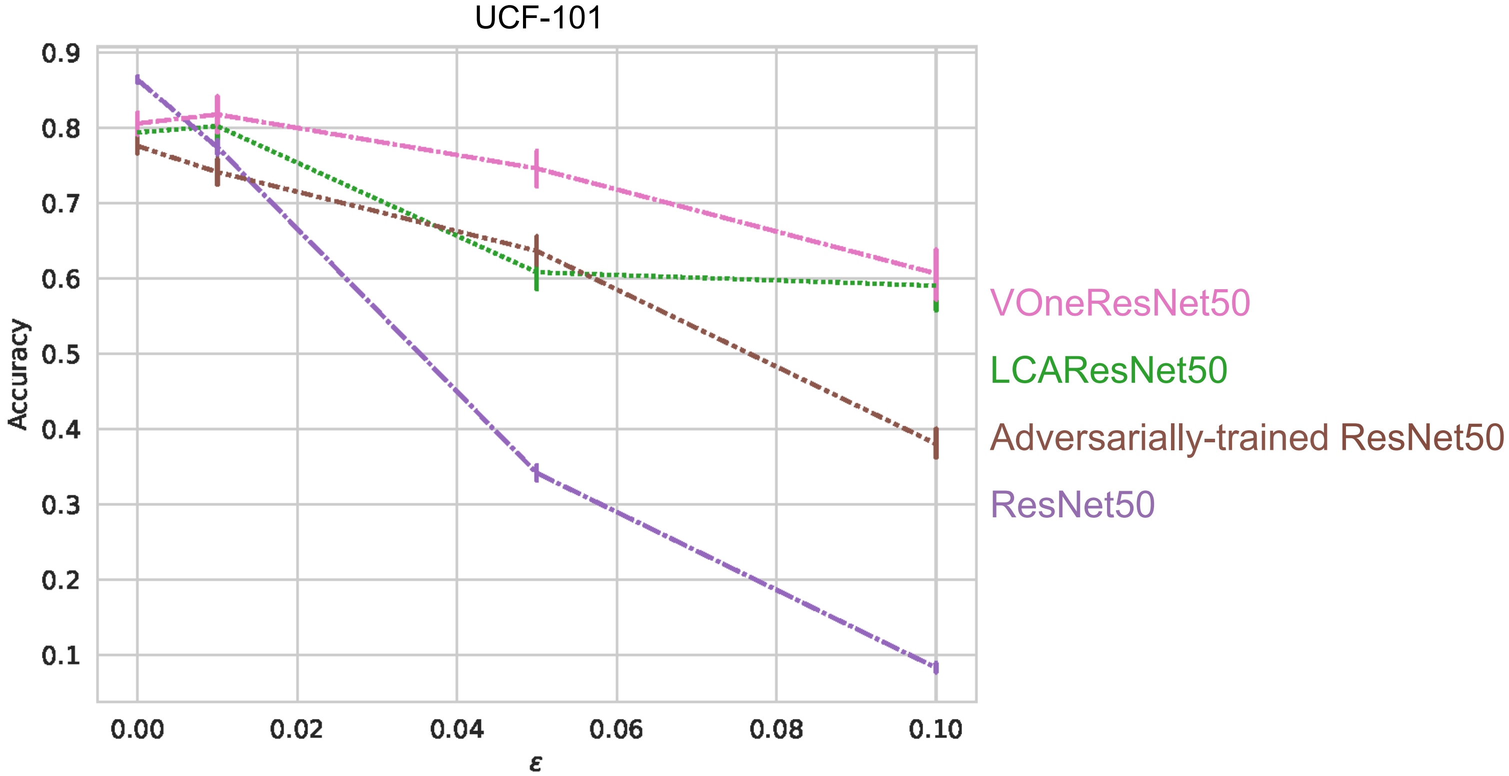
Ilyas et al. 2018. Prior convictions: Black-box adversarial attacks with bandits and priors. *ICLR 2018.*

LCANets Are Competitive Under a Black-Box Attack



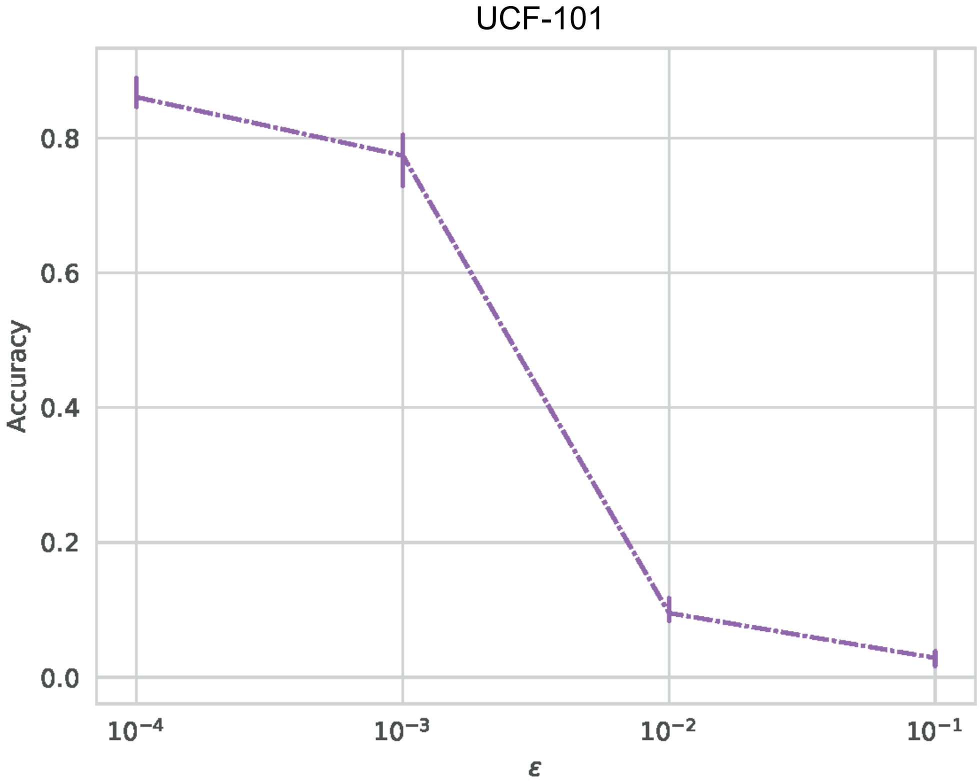
Ilyas et al. 2018. Prior convictions: Black-box adversarial attacks with bandits and priors. *ICLR 2018.*

LCANets Are Competitive Under a Black-Box Attack



Ilyas et al. 2018. Prior convictions: Black-box adversarial attacks with bandits and priors. *ICLR 2018.*

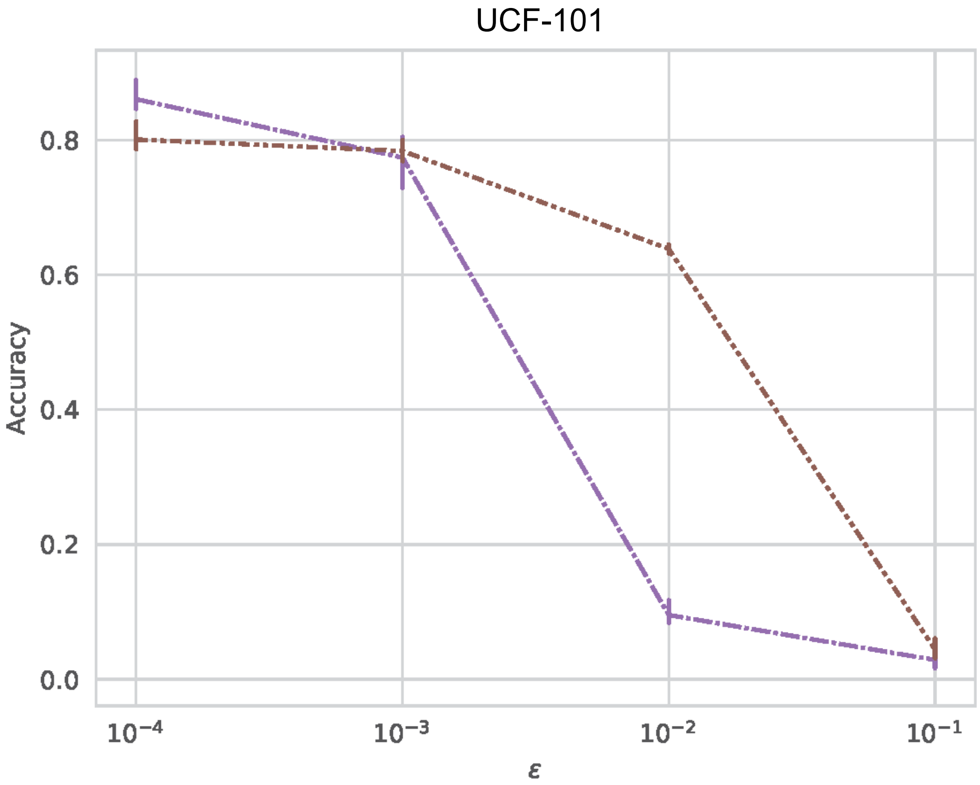
LCANets Are Not Robust to a Full White-Box Attack



ResNet50

Madry et al. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. *ICLR 2018.*

LCANets Are Not Robust to a Full White-Box Attack

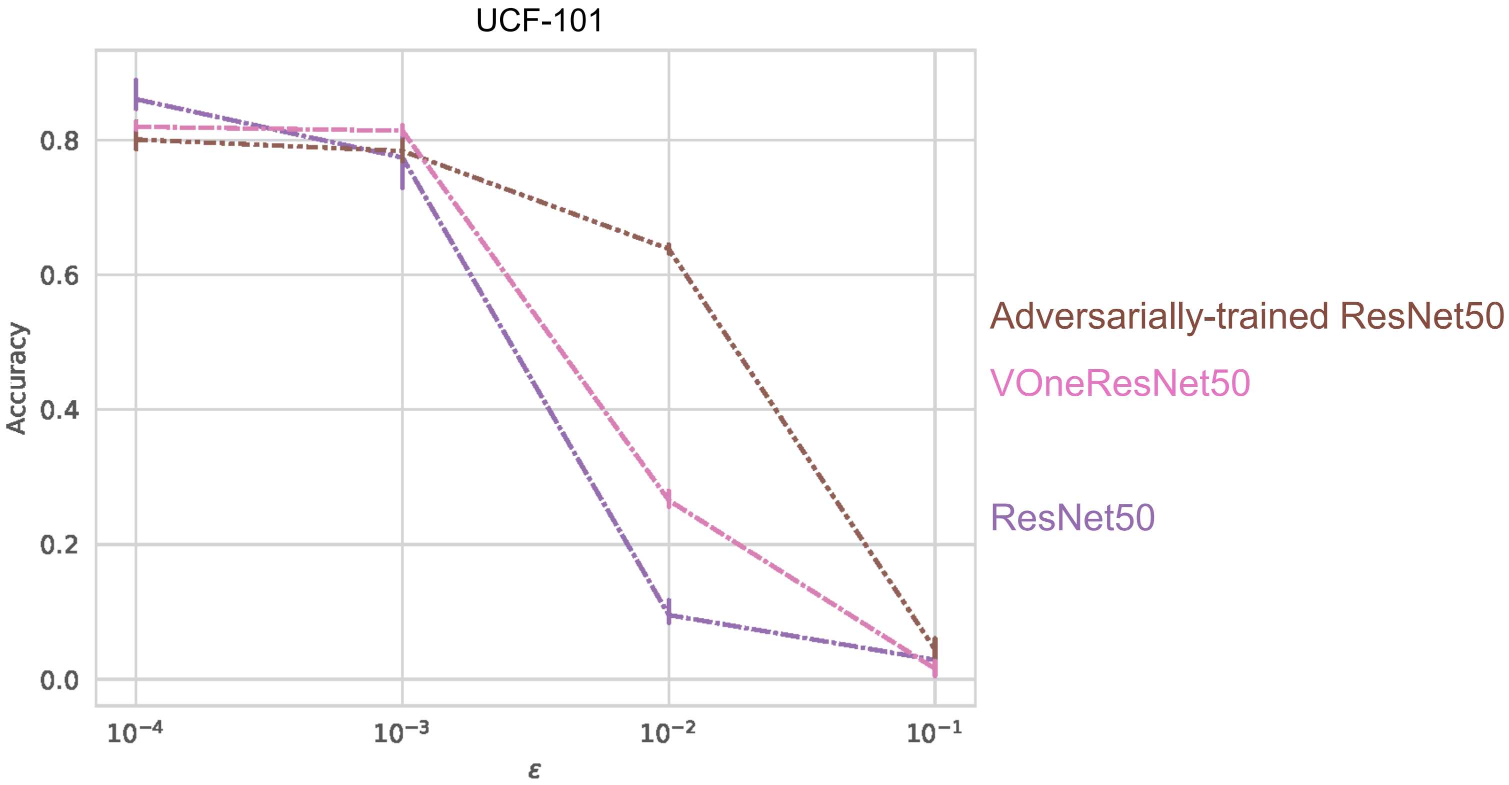


Adversarially-trained ResNet50

ResNet50

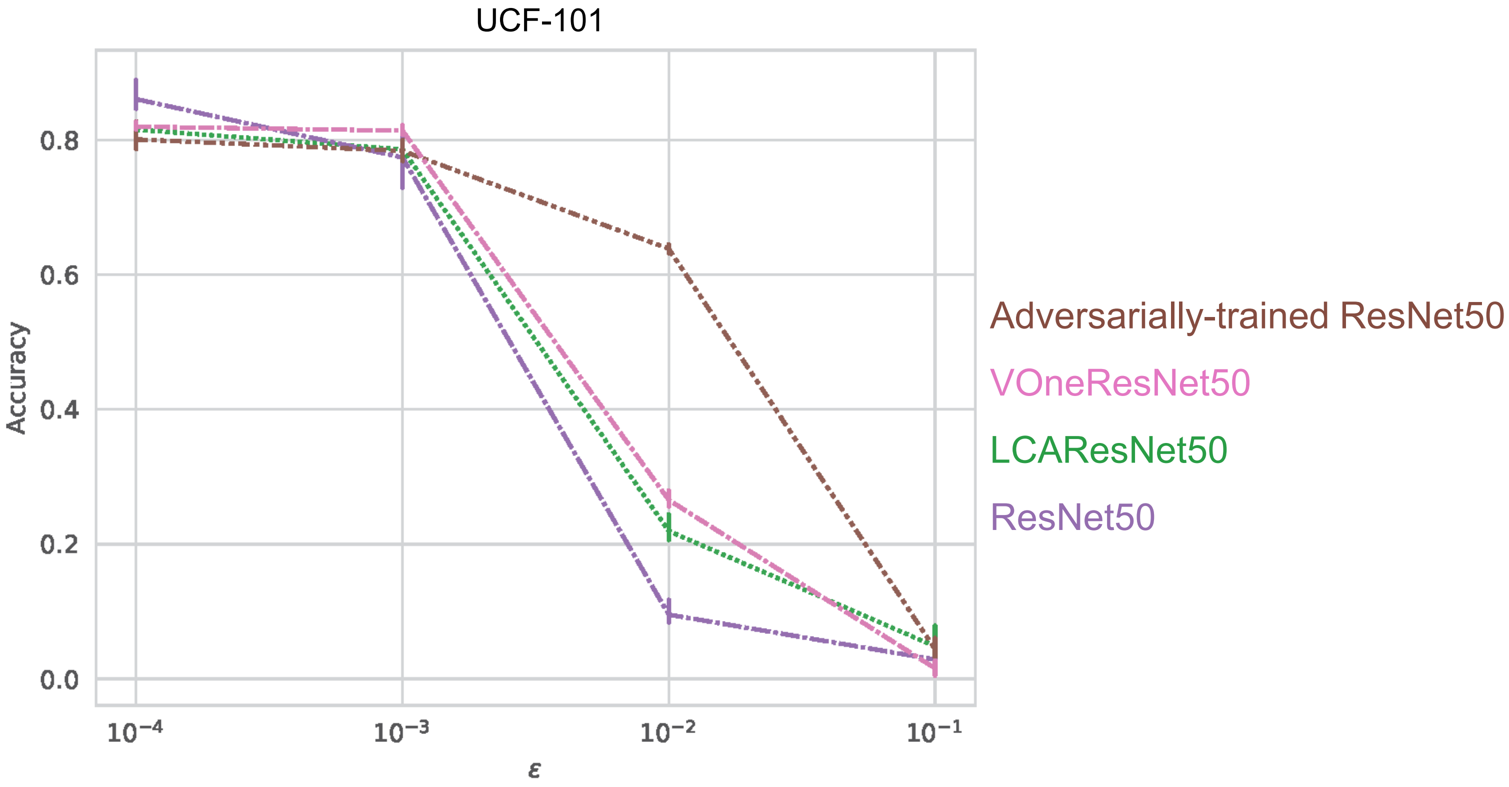
Madry et al. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. *ICLR 2018.*

LCANets Are Not Robust to a Full White-Box Attack



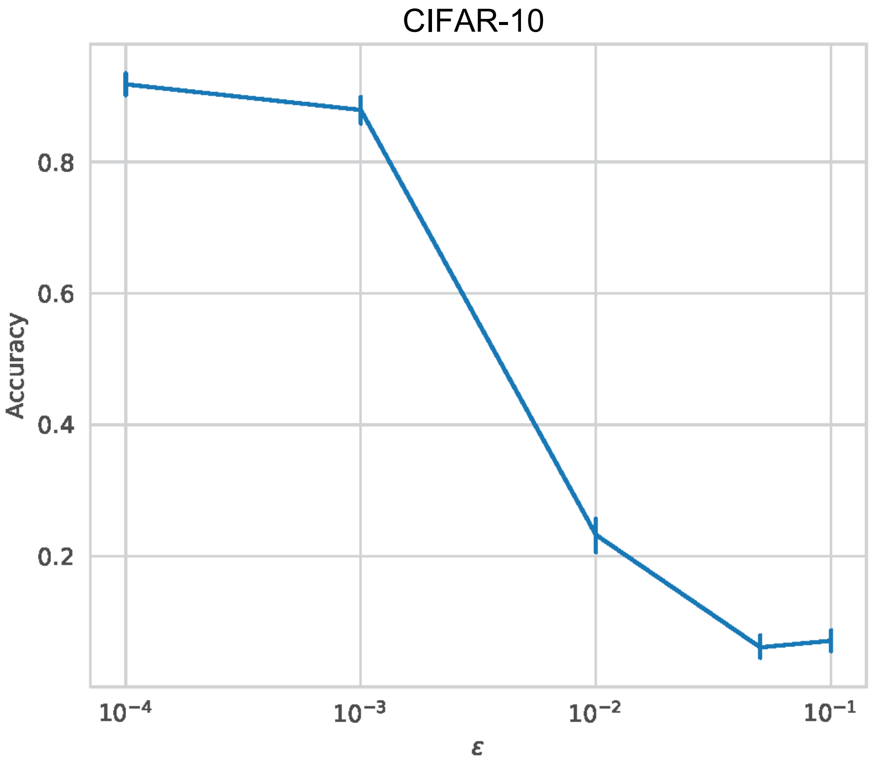
Madry et al. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. *ICLR 2018.*

LCANets Are Not Robust to a Full White-Box Attack



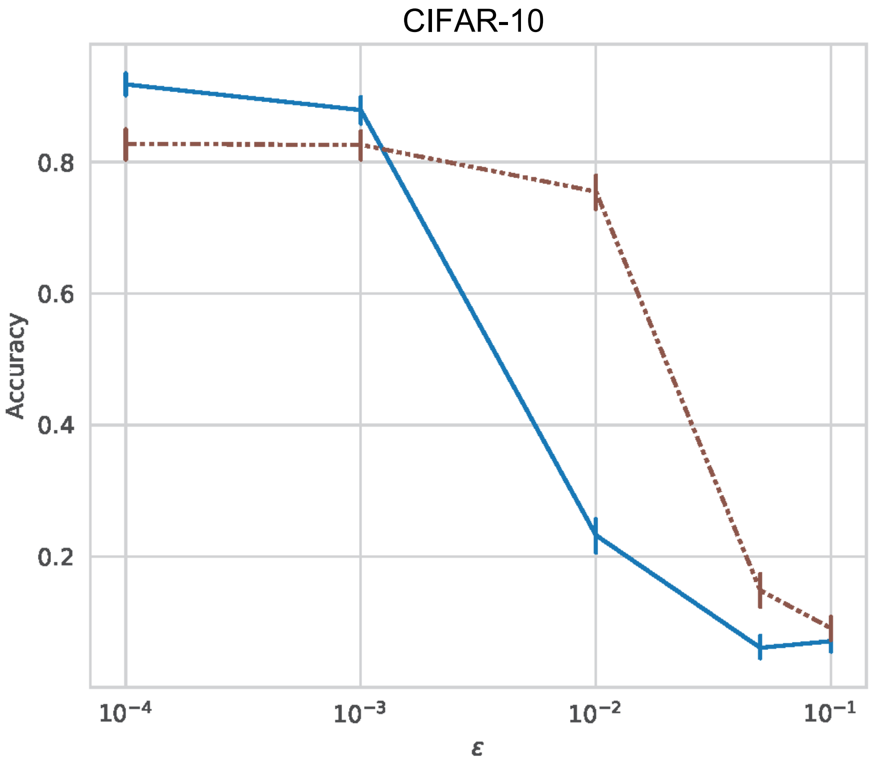
Madry et al. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. *ICLR 2018.*

LCA Frontends Can Augment Adversarial Training



ResNet18

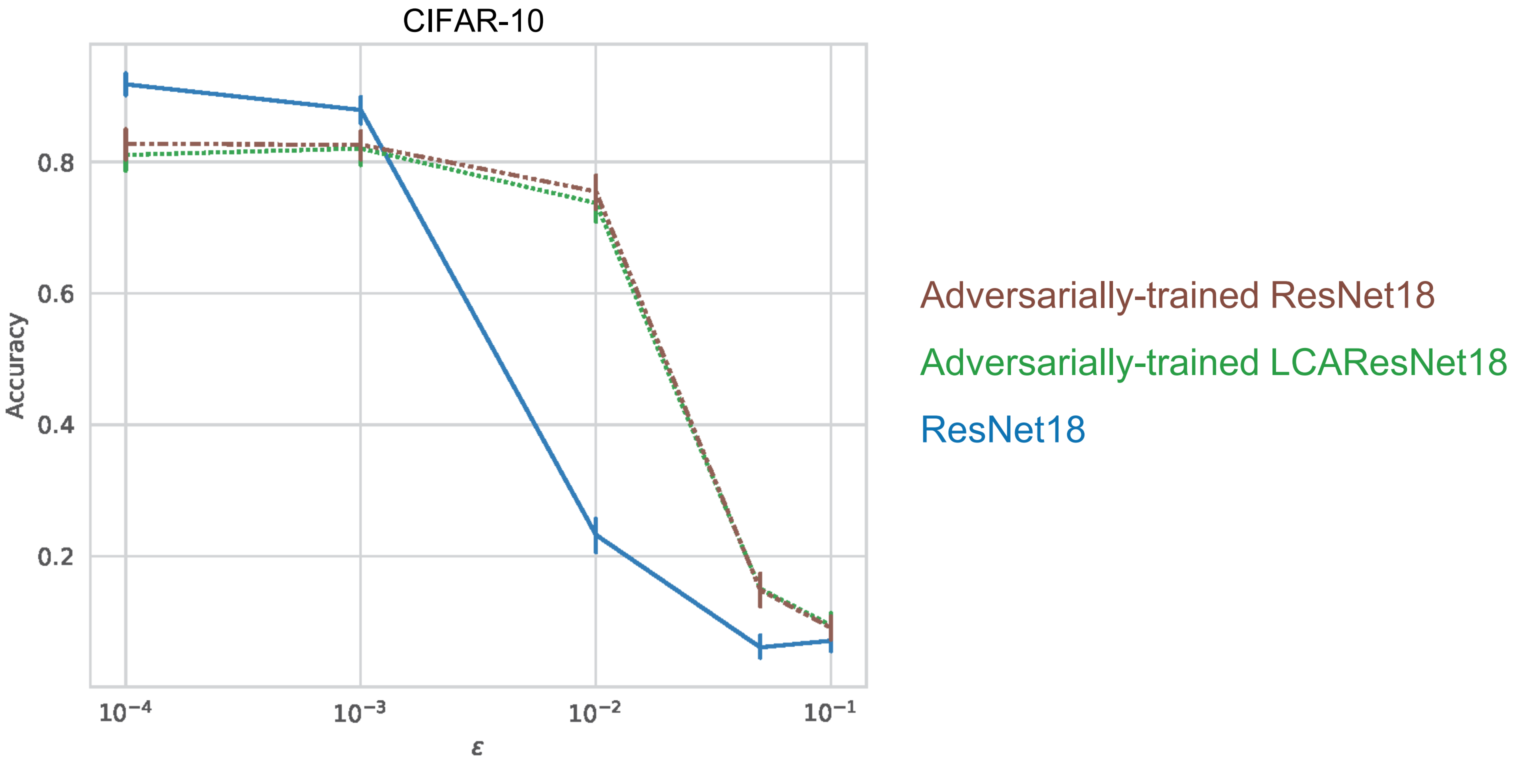
LCA Frontends Can Augment Adversarial Training



Adversarially-trained ResNet18

ResNet18

LCA Frontends Can Augment Adversarial Training



Thank You