CCN Poster

Pixel-Based Similarities as Alternative to Neural Data in CNN Regularization Against Adversarial Attacks

Elie Attias*, Cengiz Pehlevan, Dina Obeid**

Harvard John A. Paulson School of Engineering and Applied Sciences

* elieattias@g.harvard.edu , ** dinaobeid@seas.harvard.edu

Harvard John A. Paulson School of Engineering and Applied Sciences

Motivation

- Recent studies show that training CNNs with regularizers that promote brain-like representations, using neural recordings, improve model robustness.
- However, the requirement to use neural data severely restricts the utility of these methods.
- **Is it possible to develop regularizers that mimic the computational function of neural regularizers without the need for neural recordings?**

1. A neuroscience inspired objective to enhance robustness

We augment the CNNs objective function L_{task} with a term L_{sim} as in Li et al. [1]. Thus,

$$L = L_{task} + \alpha L_{sim}\,, \quad \text{where}$$

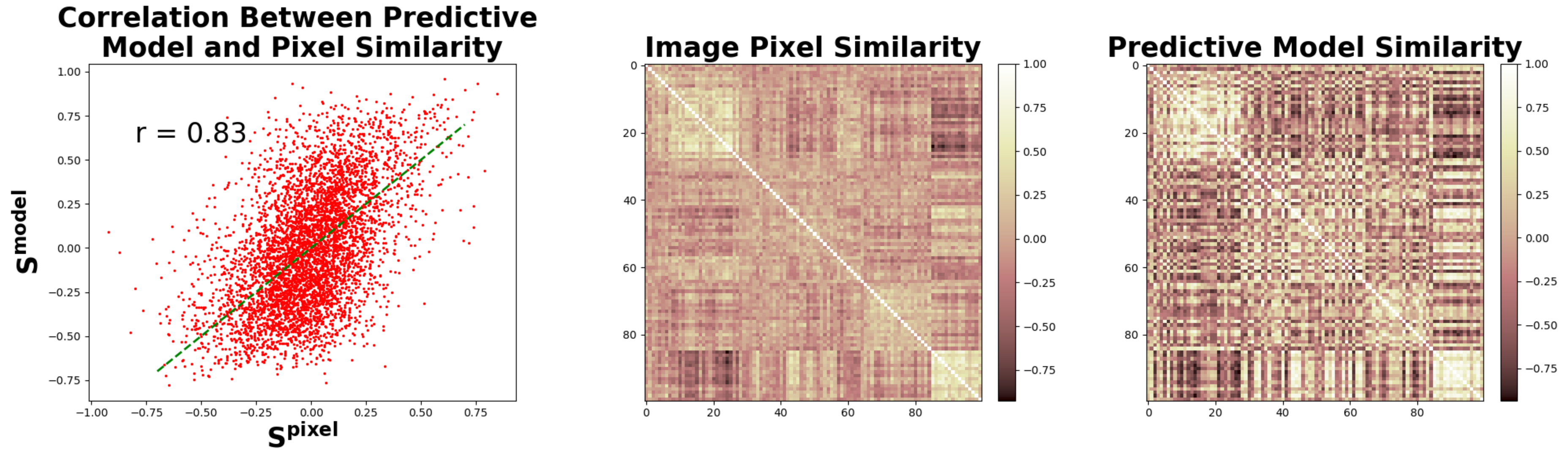
$$L_{sim} = \sum_{i \neq j} \left(\operatorname{arctanh}(S_{ij}^{CNN}) - \operatorname{arctanh}(S_{ij}^{target}) \right)^2 \quad \text{and} \quad S_{ij}^{CNN} = \sum_l \gamma_l S_{ij}^{CNN-l}.$$

S_{ij}^{CNN-l} is the mean-subtracted cosine feature similarity at layer l for image pairs (i, j); S_{ij}^{target} is the target similarity, and $\gamma_l \geq 0$ for all l, are trainable weights s.t. $\sum_l \gamma_l = 1$.

Li et al. [1] computed S_{ij}^{target} from a model trained to predict the neural recordings in mouse V1 for a given image. They showed that the regularized network is more robust.

2. Observation

We notice that similarities from the predictive model correlate with image pixel similarity.



3. Introducing a pixel-based regularizer

We introduce a new S_{ij}^{target} based on the images pixel similarity S_{ij}^{pixel}, defined as :

$$S_{ij}^{target} = \begin{cases} 1 & \text{if } S_{ij}^{pixel} > Th \\ -1 & \text{if } S_{ij}^{pixel} < -Th\,, \text{ where } Th \in (0,1)\,. \\ 0 & \text{otherwise} \end{cases}$$

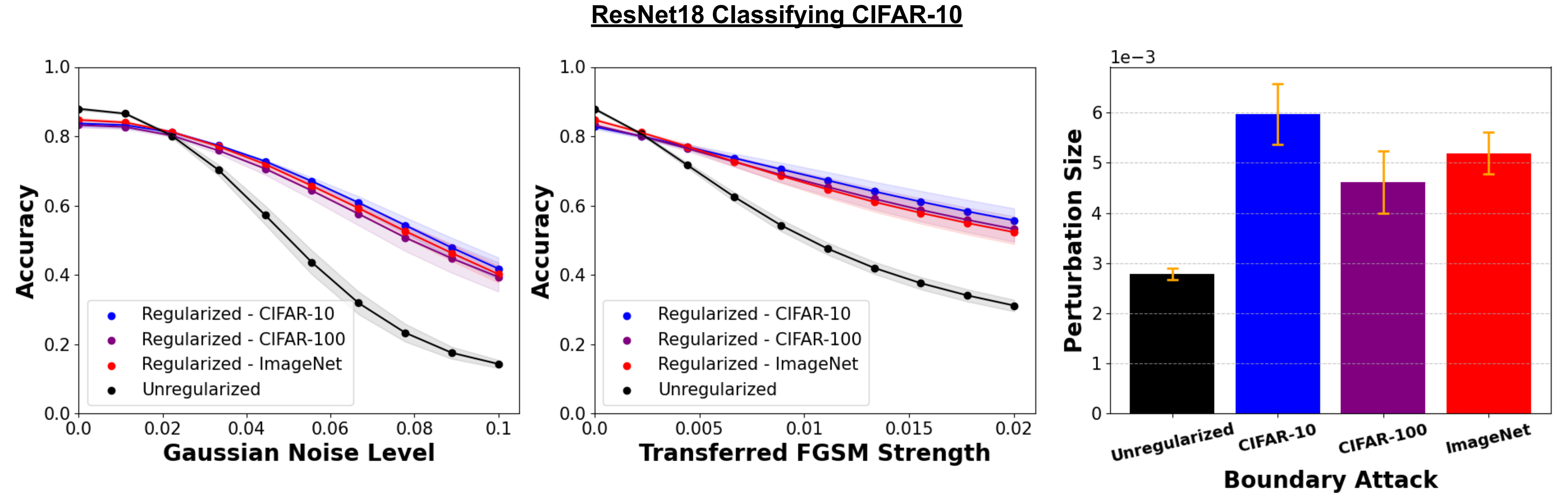
Th is a hyperparameter, chosen s.t. the task accuracy-robustness tradeoff is small.

4. Main results

For consistency, we show results from ResNet18 trained to classify CIFAR-10. Datasets are grayscale. We also tested other classification datasets (CIFAR-100, MNIST, Fashion MNIST), as well as colored images, and observed an increase in robustness.

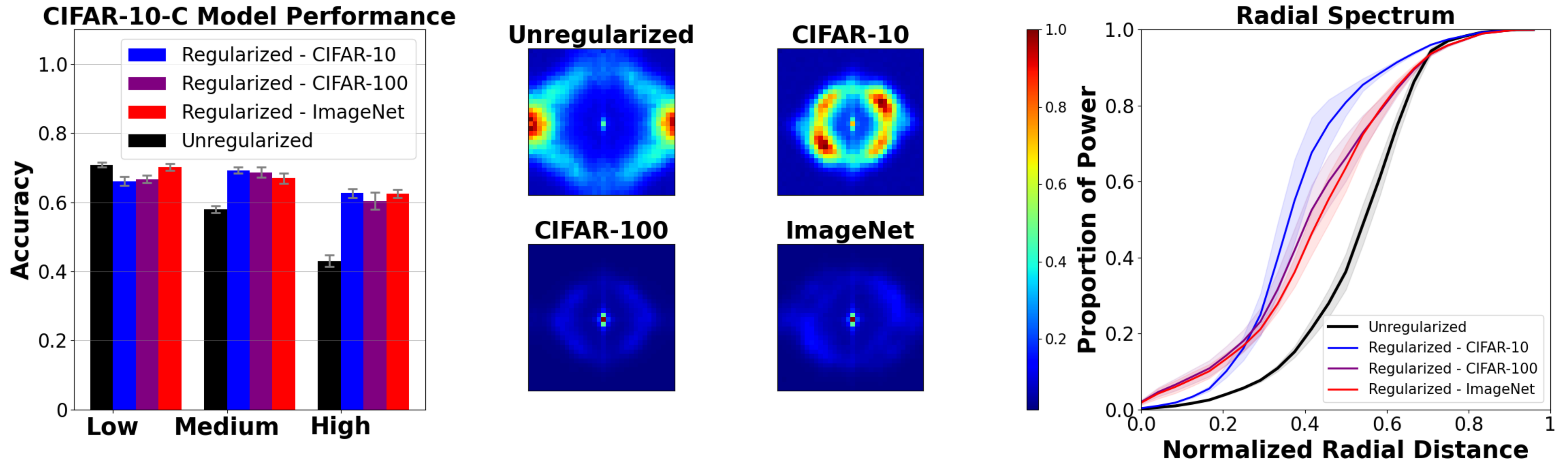
i. Increased robustness to a range of black-box adversarial attacks

Different datasets can be successfully used for regularization as we see below.



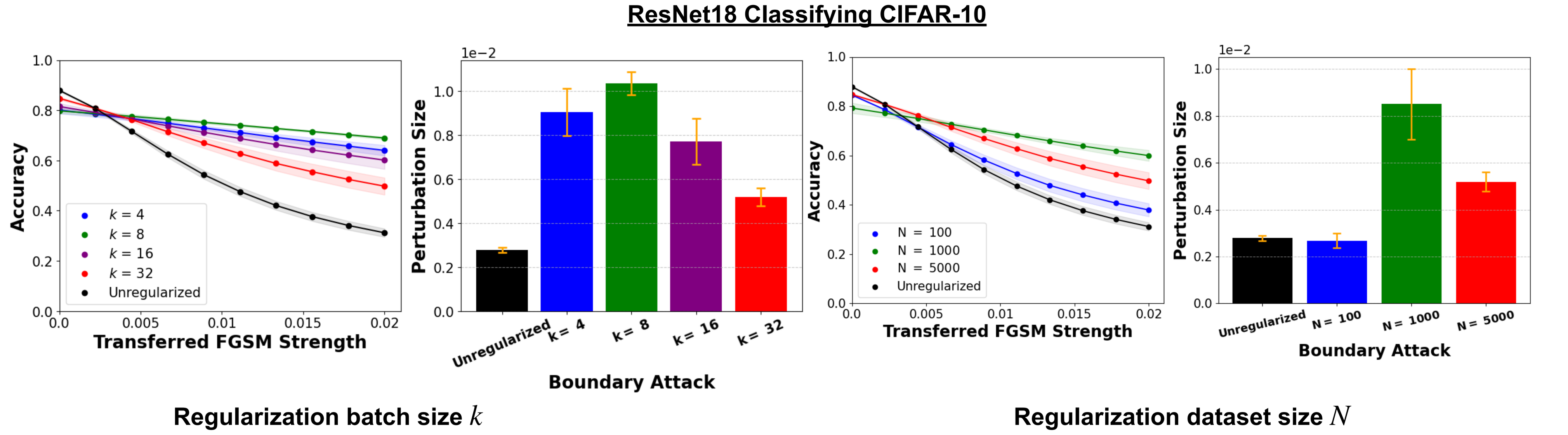
ii. Sensitivity to low frequencies

Using a boundary attack, we find that minimal adversarial perturbations contain mostly low frequencies, and that regularized models are insensitive to high frequency perturbations.

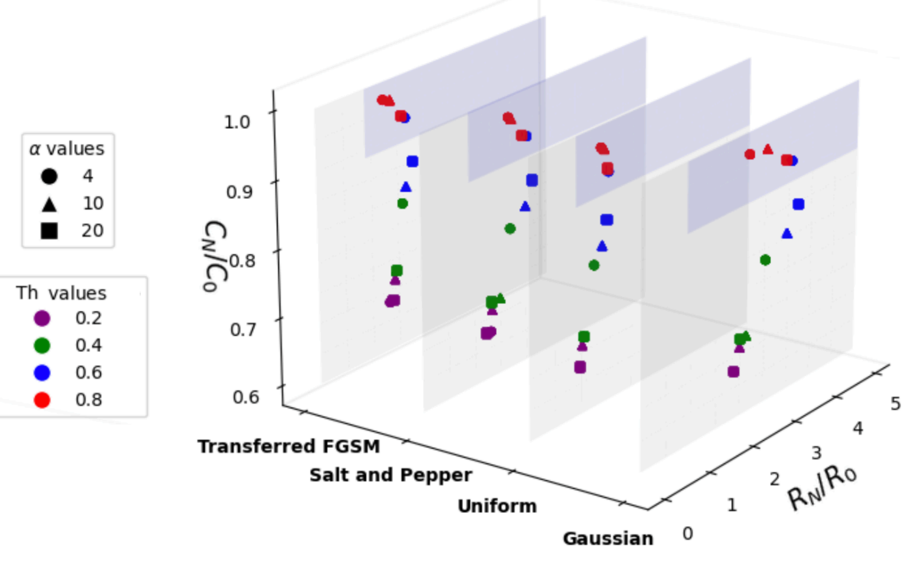


iii. Computational efficiency

We find that a small regularization batch size and dataset are sufficient to significantly increase robustness.



5. Accuracy-robustness tradeoff

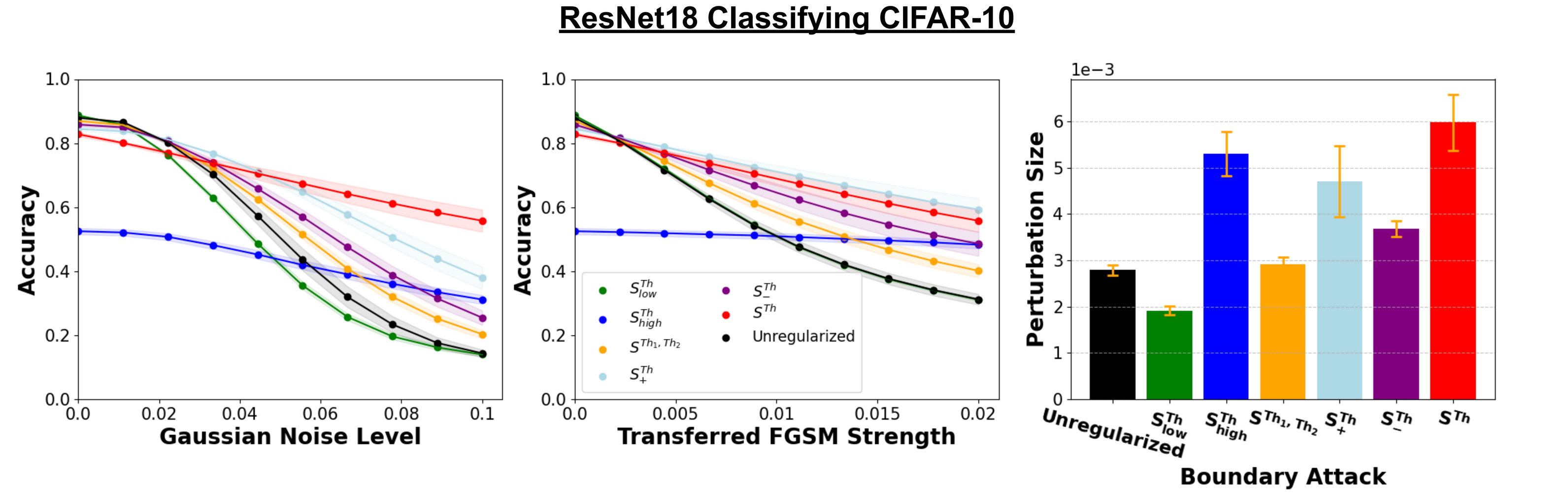


ResNet18 Classifying CIFAR-10 Regularized with ImageNet

C_0, R_0 : the accuracy at zero-distortion, and distortion of $\epsilon = 0.1$(random attacks) and 0.02 (transferred FGSM) for the **Unregularized model.**

C_N, R_N : same as above but for a **Regularized model.**

6. Investigating the different components of the regularizer



[1] Li et al., *Advances in neural information processing systems*, 32, 2019.