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

Elie Attias (elieattias@g.harvard.edu)
Harvard John A. Paulson School Of Engineering And Applied Sciences
29 Oxford St, Cambridge, MA 02138, USA

Cengiz Pehlevan (cpehlevan@seas.harvard.edu)
Harvard John A. Paulson School Of Engineering And Applied Sciences
29 Oxford St, Cambridge, MA 02138, USA

Dina Obeid (dinaobeid@seas.harvard.edu)
Harvard John A. Paulson School Of Engineering And Applied Sciences
52 Oxford St, Cambridge, MA 02138 USA

Abstract

Convolutional Neural Networks (CNNs) excel in many visual tasks but are highly sensitive to slight input perturbations that are imperceptible to the human eye, often resulting in task failures. Recent studies indicate that training CNNs with regularizers that promote brain-like representations, using neural recordings, can improve model robustness. However, the requirement to collect neural data restricts the utility of these methods. Is it possible to develop regularizers that mimic the computational function of neural regularizers without the need for direct neural recordings, thereby expanding the usability and effectiveness of these techniques? In this work, we inspect a neural regularizer introduced in Li et al. (2019) to extract its underlying strength. This regularizer uses neural representational similarities, which we find also correlate with pixel similarities. Motivated by this finding, we introduce a new regularizer that retains the essence of the original but is computed using only image pixel similarities, eliminating the need for neural recordings. We show that our regularizer significantly advances model robustness for a wide range of black box attacks. Our work opens the door to explore how biologically motivated loss functions can be used to drive the performance of artificial neural networks using a method accessible to the broader machine learning community.

Keywords: Neuroscience; Machine Learning; CNN; Adversarial Attacks; Image Classification

Introduction

Convolutional Neural Networks (CNNs) have achieved high performance on a variety of visual task such as image classifications, image segmentation, object recognition etc. Despite their huge success, Szegedy et al. (2013) found that adding small perturbations to an image that are nearly imperceptible to the human eye, can mislead the network to misclassify that image. These perturbed images were coined adversarial images and represent a large threat to computer vision models. Recent studies have shown that deep neural networks trained to emulate brain-like representations, are more resistant to adversarial attacks (Li et al., 2019; Safarani et al., 2021; Li et al., 2023). In particular, Li et al. (2019) showed that by adding a loss term or regularizer that drives the CNN to align its representational similarities (Kriegeskorte et al., 2008), with those of mouse primary visual cortex (V1), significantly increases the network's robustness to Gaussian noise and adversarial attacks. Using auxiliary loss functions to steer models towards brain-like representations is referred to as neural regularization.

In this work, we take a deeper look at this similarity loss term or regularizer introduced in Li et al. (2019). We explore how this biologically inspired loss term can be used to increase the robustness of deep neural networks in a simple and accessible way, that does not require the use of large scale neural recordings, which can be quite costly to obtain and may require additional computational cost to process.

We evaluate model robustness on black box attacks, where the attacker does not have access to model parameters as opposed to white box attacks. We show that the similarity loss term drives the network to be more robust towards a wide range of black box attacks when the similarity targets are extracted from the regularization image dataset directly rather than from neural responses. We also demonstrate the success of this method by applying it to different datasets. Our work shows how that the similarity loss term can be utilized to increase the robustness of a model using a simple method that is accessible to any user. Our work also opens the door to explore how biologically inspired loss functions can be broadly used to enhance the performance of artificial neural networks.

A Similarity Loss Neural Regularizer

Li et al. (2019) introduced a similarity loss L_{sim} as a neural regularizer to enhance the robustness of CNNs against adversarial attacks. The total loss function L has the form,

$$L = L_{\text{task}} + \alpha L_{\text{sim}} \tag{1}$$

with L_{sim} defined as,

$$L_{\text{sim}} = \Big(\text{arctanh}(S_{ij}^{\text{CNN}}) - \text{arctanh}(S_{ij}^{\text{target}})\Big)^2 \tag{2}$$

α is a parameter that sets the regularization strength, and S_{ij}^{CNN} and S_{ij}^{target} are the CNN's and target's representational similarity between images i and j, respectively.

Li et al. (2019) used a ResNet (He et al., 2016) trained for image classification task. The neural data was collected from the primary visual cortex (V1) while the mouse was looking at images from ImageNet dataset. However, Li et al. (2019) did not use the neural responses directly to compute S_{ij}^{target}, rather they utilized a predictive model (Sinz et al., 2018) to estimate neural responses, a step which they argued is key for denoising the measured neural data.

Method

We inspect the similarity loss L_{sim} in eq.(2) to distill the method introduced in Li et al. (2019) and make it more accessible to general users who usually do not have access to neural recordings. We first observe that the response similarities computed from the predictive model, which is used as a proxy for the neural data, correlate well with the similarities computed directly from image pixels (Fig.1). This observation may not be surprising since V1 is the first visual processing area in the cortex.

Indeed using image pixel similarities as target similarities in L_{sim}, that is setting S_{ij}^{target} to S_{ij}^{pixel} enhances the model's robustness to some attacks (see the blue plots in Fig. 2 and Fig. 3). However, we find that if we modify the target similarities by defining

$$S_{ij}^{Th} = \begin{cases} 1, & \text{if } S_{ij}^{\text{pixel}} > Th, \\ -1, & \text{if } S_{ij}^{\text{pixel}} < -Th, \\ 0, & \text{if } |S_{ij}^{\text{pixel}}| \leq Th \end{cases} \tag{3}$$

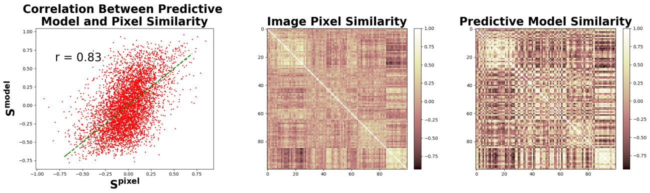


Figure 1: Correlation between the predictive model used in Li et al. (2019) and image pixel similarity. We trained more than 3 models on 6 distinct scans to predict neural responses and averaged their resulting representational similarity. We observe that the representational similarity correlate with the image pixel similarity.

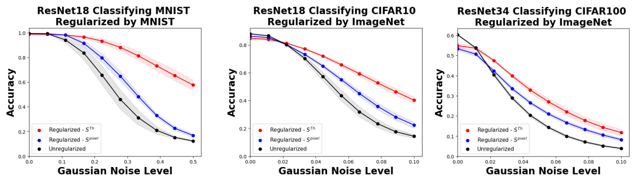
we can further enhance the model's robustness to a wide range of black box attacks, as shown in section Main Results below. $Th \in (0, 1)$ is a tunable thresholding parameter. Therefore, in our method S_{ij}^{target} is simply S_{ij}^{Th}. We note that in practice we modified the value 1 and -1 in eq. (3) by a very small number ε since we apply the arctanh function to S_{ij}^{target} (eq.(2)).

Main Results

Note that we tune the (Th, α) pair for specific classification-regularization datasets combinations.

1. Robustness to Random Noise

We show that regularizing a CNN with S^{Th} as target similarity in the regularizer leads to a large increase in the model robustness to random noise when compared to an unregularized model. In Fig.2 we show examples of ResNets trained on image classification tasks and regularized by different datasets. Fig 2a shows that a ResNet18 trained to classify MNIST dataset is significantly more robust at high noise level without incurring any loss on distortion-free images. A similar gain in robustness is observed when training on CIFAR10 and CIFAR100 while regularized by ImageNet (Fig 2b, 2c) at the expense of a slight loss in accuracy to distortion-free images.

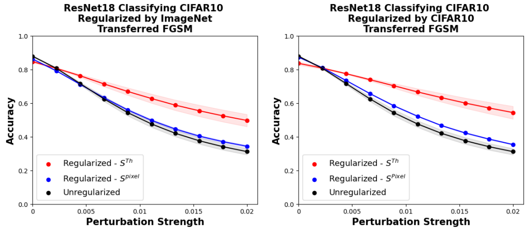


(a) (b) (c)

Figure 2: Robustness to Gaussian noise is shown for **(a)** models classifying MNIST when regularized by MNIST images ($\alpha = 4$, $Th = 0.2$) and **(b)** models classifying CIFAR10 and **(c)** CIFAR100 that are regularized on ImageNet images ($\alpha = 10$ and $Th = 0.8$). Shaded areas represent the standard error of the mean (SEM) across seven seeds per model.

2. Robustness to Stronger Black Box Attacks

The increase in robustness achieved by using S^{Th} as target similarity in the regularizer, also holds for stronger attacks, such as transferred Fast Gradient Sign Method attack (FGSM) (Goodfellow et al., 2014) and decision-based Boundary Attack Brendel et al. (2017) (results not shown here). In Fig.3 we show that a ResNet18 trained to classify CIFAR10 and regularized by ImageNet (Fig.3a) and CIFAR10 (Fig.3b) shows an increase in robustness against transferred FGSM attack.



(a) (b)

Figure 3: A ResNet18 classifying CIFAR10, regularized on **(a)** ImageNet with ($\alpha = 10$, $Th = 0.8$) and **(b)** CIFAR10 with ($\alpha = 10$, $Th = 0.8$) is evaluated against transferred FGSM (Goodfellow et al., 2014) perturbations from an unregularized model. Shaded areas represent the SEM across seven seeds per model.

3. Robustness Across Dataset Combinations

Although we only show examples of specific classification-regularization dataset combinations, we observe that regularizing across various combinations of datasets leads to an increase in the model's robustness against a wide range of black box attacks (Gaussian, Salt and Pepper and Uniform noise, transferred FGSM and decision-based Boundary Attacks).

Experimental Set up

All models were trained by stochastic gradient descent on a NVIDIA A100-SXM4-40GB GPU. Models classifying CIFAR100, CIFAR10, MNIST were trained during 60, 40 and 20 epochs respectively. Training and regularizing a ResNet18 and a ResNet34 on CIFAR10 and CIFAR100 took in average 34 min and 48min to run, respectively. We used a batch size of 64 for the classification pathway and a batch of 16 image pairs for the regularization pathway. We use the same learning schedule as in Li et al. (2019). Models were trained using Pytorch (Paszke et al., 2017). All code will soon be publicly available on Github.

Conclusion

Extracting working principles of the brain to advance AI is a long term goal of neuroscience. Working towards this goal, we showed that a neuroscience inspired regularizer capitalising on pixel representations can successfully increase CNN robustness without the requirement of measuring and processing large scale neural recordings. Our method has proved to be successful for various combinations of classification datasets (MNIST, FashionMNIST, CIFAR10, CIFAR100) and regularization datasets (MNIST, FashionMNIST, CIFAR10, CIFAR100 and ImageNet), as well as for a wide range of black box attacks while preserving a high accuracy on distortion-free images. This work is an encouraging step towards dissecting the working of neural regularizers to explore how they can be used to enhance machine learning model performance in a way that is accessible to the broader machine learning community.

References

Brendel, W., Rauber, J., & Bethge, M. (2017). Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 249.

Li, Z., Brendel, W., Walker, E., Cobos, E., Muhammad, T., Reimer, J., . . . Tolias, A. (2019). Learning from brains how to regularize machines. *Advances in neural information processing systems*, *32*.

Li, Z., Ortega Caro, J., Rusak, E., Brendel, W., Bethge, M., Anselmi, F., . . . Pitkow, X. (2023). Robust deep learning object recognition models rely on low frequency information in natural images. *PLOS Computational Biology*, *19*(3), e1010932.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., . . . Lerer, A. (2017). Automatic differentiation in pytorch.

Safarani, S., Nix, A., Willeke, K., Cadena, S., Restivo, K., Denfield, G., . . . Sinz, F. (2021). Towards robust vision by multi-task learning on monkey visual cortex. *Advances in Neural Information Processing Systems*, *34*, 739–751.

Sinz, F., Ecker, A. S., Fahey, P., Walker, E., Cobos, E., Froudarakis, E., . . . Tolias, A. (2018). Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. *Advances in neural information processing systems*, *31*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.