UMassAmherst

College of Information & Computer Sciences

Making Neural Networks Robust To Adversarial Attacks

Aditya Kuppa, Piusha Gullapalli

Department of Computer Science, University of Massachusetts Amherst

Abstract

Convolutional Neural Networks (CNNs) have been successful in computer vision applications such as image classification, but they are vulnerable to adversarial attacks. Adversarial attacks involve adding imperceptible perturbations to input images, causing the network to misclassify them. Traditional defense mechanisms against adversarial attacks are easily bypassed, making it challenging to defend against such attacks. Researchers have developed several approaches to improve CNNs' robustness against adversarial attacks, such as adding noise to the input data or adversarial training. This paper proposes a new approach to enhance CNNs' robustness by adding new channels to the input image, representing the difference between the original image and its shifted version. The method improves the network's performance on adversarial images by providing additional context information to the network. The proposed approach outperforms traditional CNNs on both clean and perturbed datasets, demonstrating its effectiveness in defending against adversarial attacks.

Methodology

We propose a novel approach to improve the performance of Convolutional Neural Networks (CNNs) in identifying objects in images by incorporating additional “context” channels into the input image. Specifically, the new channels that we add to the input image, come from the differences between the original image, and its left-shifted, right-shifted, up-shifted, down-shifted, and noisy versions. The intuition is to provide CNNs with additional contextual information and make them more spatially aware, potentially improving their ability to detect and classify objects in images.

We have run our experiments using SqueezeNet,and VGG-19 deep neural network architectures. As expected, the VGG-19 model performs better on the CIFAR-10 dataset that we used for experiments. The models were evaluated on both the standard, and perturbed versions of dataset. We have used Fast Gradient Sign Method(FGSM) to add perturbations to the images in the test dataset.

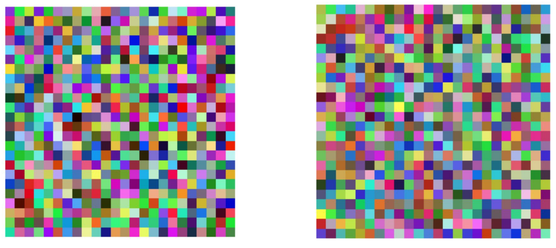

Figure 1: Weights of the network before and after training

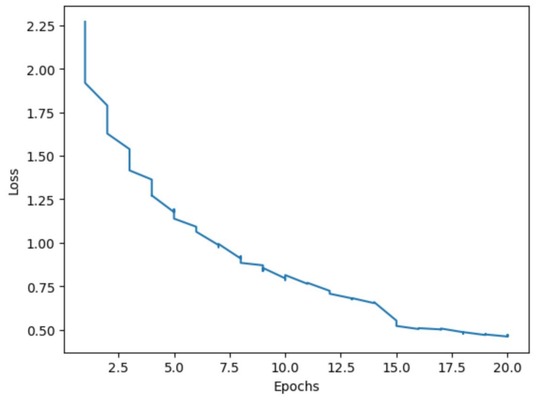

Figure 2: Loss vs epoch graph

Results

Model	Accuracy on Standard Dataset		Accuracy on Adversarial Dataset	
	Without context	With context	Without context	With context
SqueezeNet	47%	50%	9%	14%
VGG-19	70%	80%	11%	19%

Table 1: Accuracy on different models with and without context on standard and adversarial dataset

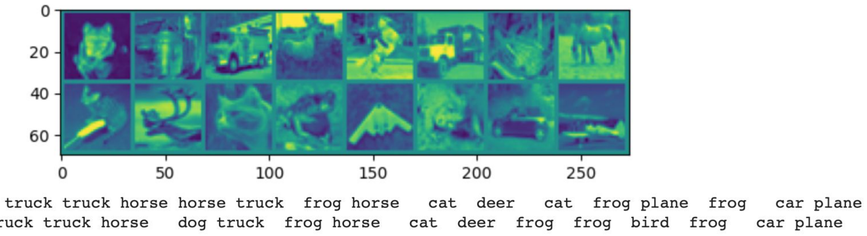

Figure 3: Images and their true and predicted labels

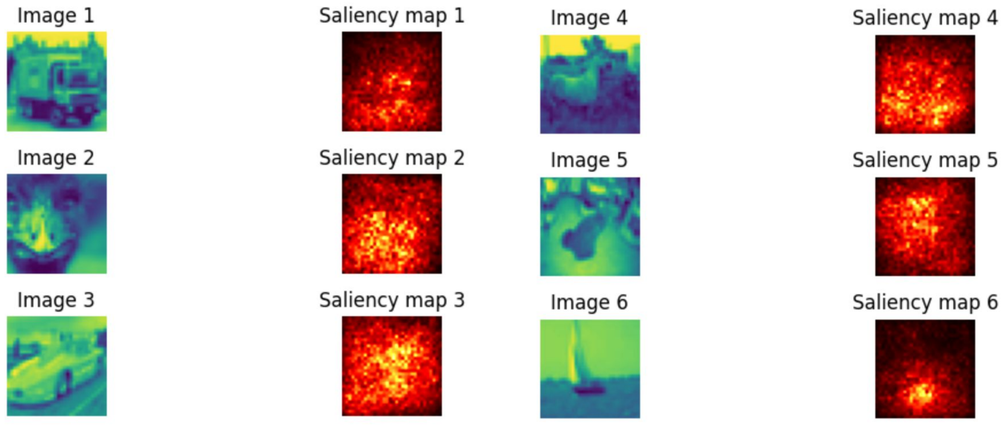

Figure 4: Some of the images with their corresponding saliency maps

References

1. Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2017). Synthesizing robust adversarial examples. arXiv preprint arXiv:1707.07397.
2. Zhang, H., Yang, C., Zhang, X., & Jia, Y. (2020). DVS-Attacks: Adversarial Attacks on Dynamic Vision Sensors for Spiking Neural Networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 0, 1-9.
3. Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533.
4. Captum. (n.d.). Captum - Model Interpretability for PyTorch. Retrieved May 5, 2023, from https://captum.ai/

Conclusion and Future Work

Dealing with adversarial attacks on CNNs has been an active area of research for years now. In this paper, we propose a way to deal with adversarial attacks based on FGSM. The proposed idea takes motivation from the DVS camera and mitigates the issue by adding extra channels called "context" channels to the input images of a neural network. Experimental results have shown that our proposed approach outperforms a standard neural network on an adversarial dataset, thus making it robust to adversarial attacks up to some extent. The potential future work could experiment with transformations other than shifting, and adding noise, and analyzing the performance on other kinds of FGSM attacks, like iterative FGSM to say.