Making Neural Networks Robust To Adversarial Attacks

Aditya Kuppa
Department of Computer Science
University of Massachusetts Amherst
vkuppa@umass.edu

Piusha Gullapalli
Department of Computer Science
University of Massachusetts Amherst
pgullapalli@umass.edu

Abstract

Convolutional Neural Networks(CNNs) are widely used in the field of Computer Vision for successfully performing various tasks like Image Classification, Object Detection, Image Segmentation, etc. However, these models do not inherently encode invariances to certain types of image transformations, such as translations or rotations, making them vulnerable to adversarial attacks. This paper presents a method to improve the performance of CNN models on adversarial images by training them on context-added images. "Context" is simply the difference between the original image and the transformed image. The context information is added to the original image in the form of additional channels. Our proposed approach outperforms a standard CNN like VGG-19, Squeezenet, etc. on the standard CIFAR-10 dataset by 3%, and the perturbed CIFAR-10 dataset by 5%, indicating increased robustness towards adversarial attacks.

1 Introduction

Convolutional Neural Networks (CNNs) have shown remarkable success in computer vision applications, particularly in image classification tasks. However, CNNs are vulnerable to adversarial attacks, which can cause the network to misclassify the input data, leading to potentially harmful consequences. Adversarial attacks are a type of attack that involves adding imperceptible perturbations to the input image, resulting in a misclassification by the network.

Adversarial attacks are challenging to defend against since they can be crafted with high precision and easily bypass traditional defense mechanisms. These attacks are particularly concerning in real-world applications, such as autonomous driving and security systems, where the reliability and robustness of the network are crucial.

In recent years, researchers have focused on developing defense mechanisms to improve the robustness of CNNs against adversarial attacks. One such approach is to add noise to the input data, which makes it difficult for the attacker to craft an adversarial example. Another approach is adversarial training, which involves training the network on normal and adversarial examples to improve its robustness.

In this paper, we propose a novel approach to improve the robustness of CNNs against adversarial attacks. Our approach involves adding new channels to the input, where each channel represents the difference between the original image and its shifted version. The shifted images are generated by shifting the original image by one pixel in different directions. The different channels provide the network with additional information about the input, making it more difficult for attackers to craft adversarial examples.

We demonstrate the effectiveness of our proposed approach by evaluating it on benchmark datasets. Our results show that our approach improves the robustness of CNNs against adversarial attacks

while maintaining high accuracy on clean data. Adversarial attacks pose a significant threat to the effectiveness of CNNs, and our approach offers a novel way to defend against these attacks.

2 Related Work

Spatially Transformed Adversarial Examples (1):
In 2017, Athalye et al. proposed a novel approach to generate adversarial examples by applying spatial transformations to the input images. They showed that applying geometric transformations, such as rotation, scaling, and translation, to the input image could lead to significant misclassification by the network. They also demonstrated that the adversarial examples generated using this method were robust to different defense mechanisms. Their work highlighted the vulnerability of CNNs to spatial transformations and showed the importance of considering these transformations in defense mechanisms against adversarial attacks.

DVS-Attacks: Adversarial Attacks on Dynamic Vision Sensors for Spiking Neural Networks (2):
In 2020, Zhang et al. proposed a novel method for generating adversarial attacks on Dynamic Vision Sensors (DVS) used in Spiking Neural Networks (SNNs). They exploited the temporal dynamics of DVS sensors to generate perturbations that could lead to misclassification by the network. They showed that their method could be used to attack both SNNs and traditional CNNs that processed the output of DVS sensors. Their work highlighted the vulnerability of DVS sensors and the need for developing robust defense mechanisms against temporal adversarial attacks.

Adversarial Examples in the Physical World (3):
In 2016, Kurakin et al. proposed the fast gradient sign method (FGSM) for generating adversarial examples. The FGSM algorithm generates perturbations by taking the sign of the gradient of the loss function concerning the input image. The resulting perturbation is then added to the input image to generate the adversarial example. The FGSM method is simple and computationally efficient, making it a popular choice for generating adversarial examples. Kurakin et al. showed that the generated adversarial examples could transfer across different models and architectures, making them a significant threat to the security of deep learning systems. Their work highlighted the need for developing robust defense mechanisms against adversarial attacks and sparked significant research efforts.

3 Proposed Approach

We propose a novel approach to improve the performance of Convolutional Neural Networks (CNNs) in identifying objects in images by incorporating additional channels into the input image. Specifically, we add four channels to the input image, representing the differences between the original image and its left, right, up, and down-shifted versions.

CNNs are designed to be pixel-level discriminators and may not know the spatial relationships between different image parts. By including the difference channels in the input image, we aim to provide CNNs with additional contextual information and make them more spatially aware, potentially improving their ability to detect and classify objects in images.

To evaluate the effectiveness of our approach, we conduct experiments on the CIFAR-10 dataset, a widely used benchmark for image classification. We compare the performance of the baseline CNN model with that of our proposed approach. The experimental results demonstrate the efficacy of our approach in improving the performance of CNNs for image classification tasks.

Our proposed approach has the potential to be further extended and applied to other CNN-based image classification models. The additional contextual information provided by the difference channels can potentially enhance the performance of other computer vision tasks beyond object recognition.

4 Experimental Results

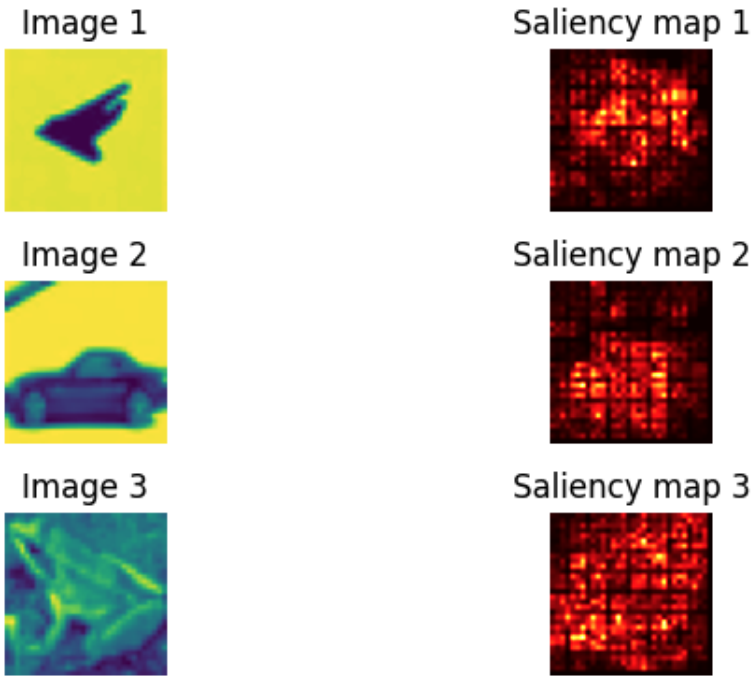


Figure 1: The image and its corresponding saliency map (4)

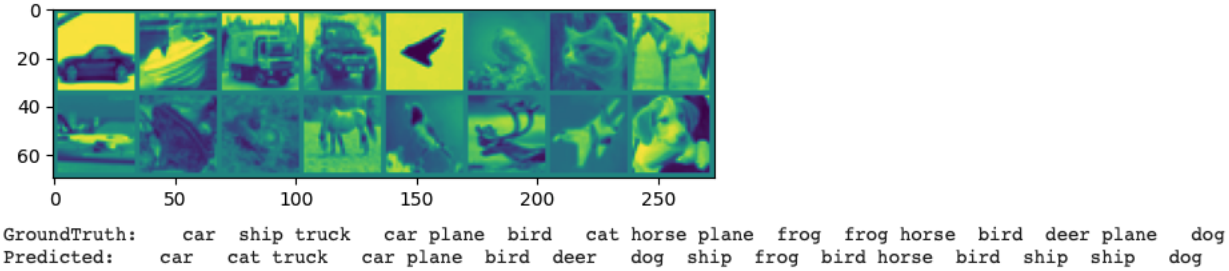


Figure 2: The images, true and predicted labels

4.1 Datasets

We have used CIFAR-10 dataset to obtain the initial results. CIFAR-10 is a benchmark dataset of 60,000 32x32 RGB images divided into 10 classes, with 6,000 images per class. The classes are airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. It is a widely used dataset for image classification and object recognition tasks. The dataset is split into 50,000 training images and 10,000 test images, with a balanced distribution of classes in both sets. The training set is further divided into five subsets, each containing 10,000 images, which are used for cross-validation purposes. We plan to experiment with datasets like CIFAR-100, MNIST, etc. in the future.

The adversarial dataset for testing purposes has been created by perturbing the images in the CIFAR-10 test set using the Fast Gradient Sign Method(FGSM). The FGSM attack works by taking the gradient of the loss function with respect to the input data and then adding a small perturbation to the input in the direction of the gradient. The size of the perturbation is controlled by a hyperparameter called the epsilon value, which determines the maximum allowable distortion of the input.

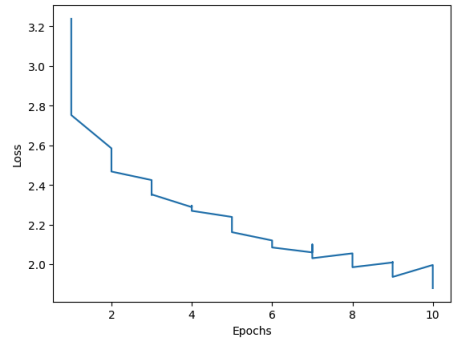
4.2 Models

We have run our experiments using SqueezeNet, a deep neural network architecture designed to achieve high accuracy with fewer parameters than traditional deep neural networks. SqueezeNet has a very small memory footprint compared to other deep neural networks, making it suitable for training on CPUs. Though the model has not shown great accuracy in our case, it has empirically verified the improvement from our proposed approach. We also plan to experiment with heavier neural networks like VGG-19, ResNet, etc. in the future.

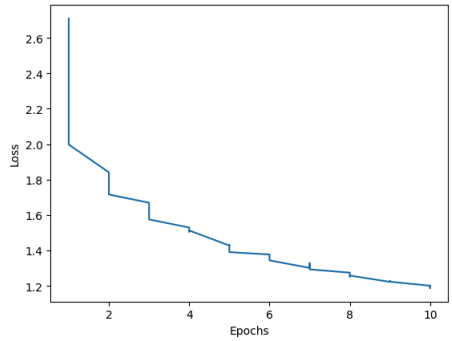
4.3 Results

The initial results obtained from training SqueezeNet on the CIFAR-10 dataset using our proposed approach are promising. A SqueezeNet with Adam optimizer trained on CIFAR-10 without the addition of context channels to the input images gives an average accuracy of 56% on the standard test set and an accuracy of 12% on the adversarial test set. However, the same neural network that is trained on CIFAR-10 by adding context channels to the input images gives an average accuracy of 59% on the standard test set and an accuracy of 17% on the adversarial test set. Therefore, it can be inferred that the proposed approach makes the neural networks robust to adversarial attacks up to some extent.

5 Performance Analysis



Accuracy of the SqueezeNet on the 10000 test images: 56 %



Accuracy of the SqueezeNet on the 10000 test images: 59 %

Figure 3: The loss with the epochs and the accuracy before and after adding the additional channels

```
Adversarial accuracy of the SqueezeNet on the 10000 test images: 12 %
Adversarial accuracy of the SqueezeNet on the 10000 test images: 17 %
```

Figure 4: The adversarial accuracy before and after adding the additional channels

Our experiments demonstrate that our proposed approach improves the model's accuracy on regular and adversarial test sets. Specifically, on the regular test set, the model accuracy improves by 3%, while on the adversarial test set, the model accuracy improves by 5%. These improvements indicate

that our approach successfully enhances the robustness and resistance of the model to unexpected inputs and adversarial attacks.

However, there are potential reasons why the model is performing poorly. One reason could be overfitting, where the model is too complex, or insufficient data to train the model properly. Another reason could be limited contextual information, where the additional information provided by our approach may not fully capture the spatial relationships between different image parts. Optimization issues, such as inappropriate optimization techniques or suboptimal optimizer, could also be a factor. Finally, the model architecture may not be well-suited for the task, potentially lacking enough capacity or the ability to capture the necessary features for accurate classification. We hope to address these issues in our future work to improve the performance of our approach further.

6 Future Work

There are many directions to experiment along - in terms of the model choice, type of adversarial attack, dataset choice, etc. The following are the major experiments (in the order of importance) that we plan to perform before the final presentation:

- Evaluate other models like VGG-19, ResNet, etc.
- Add more context channels involving transformations other than shifting.
- Experiment on other datasets like CIFAR-100, MNIST, etc.
- Analyze performance on similar types of adversarial attacks like iterative FGSM.

7 Conclusion

Dealing with adversarial attacks on CNNS has been an active area of research for years now. Multiple approaches have been proposed and published to mitigate different types of adversarial attacks. In this paper, we have proposed one such 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. The context channels are derived by taking the difference between the original image and the transformed image. 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.

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/