Enhanced Robustness by Symmetry Enforcement

Longwei Wang*, Aashish Ghimire*, KC Santosh*, Zheng Zhang†, Xueqian Li†
*Department of Computer Science, University of South Dakota
†Department of Computer Science and Software Engineering, Auburn University

***Abstract*—Convolutional Neural Networks excel in various applications but remain susceptible to adversarial attacks. Even minute alterations, like a single-pixel shift, could drastically mislead cutting-edge state-of-the-art models. In this study, we explore this vulnerability – the adversarial example problem – attributing it primarily to limited training samples. Such a case leads to overfitting and deviation from optimal models. To overcome this challenge, we propose integrating a variety of symmetry-invariant operations into network model designs. This strategy maximizes the use of available training data, amplifies the neural network's expressive capacity, and empowers its robustness. Our experiments demonstrate the effectiveness of this approach against random perturbations in test data while concurrently enhancing their generalization capabilities. By augmenting deep learning architectures with symmetry-invariant layers, we strive to mitigate vulnerabilities, enhancing both robustness and generalization adaptability.**

***Index Terms*—Robustness, Convolutional Neural Networks, Adversial Attacks, Symmetry, Generalization**

I. Introduction

Convolutional Neural Networks (CNNs) have emerged as the go-to architecture for computer vision tasks and have been used extensively as a backbone for various advanced artificial intelligence (AI) systems. However, these foundational models remain susceptible to adversarial examples—subtle input perturbations that mislead models [1]. As these attacks exploit the inherent sensitivity of CNNs to subtle changes in input, they pose a significant challenge to the deployment of these networks in safety-critical applications. While several defense mechanisms have been proposed, *e.g.* data augmentation [2] [3], the persistent nature of the adversarial example problem calls for innovative solutions that overcome the limitations of existing approaches, address the root causes of vulnerability, and are immune against optimization-crafted perturbations. Issues with adversarial robustness stem from the significant distinction between optimal and overfitted decision boundaries in models due to limited training data. Models that are overfitted, are more susceptible to misclassification from slight perturbations, thereby heightening their vulnerability to adversarial attacks. To address this, we propose an architectural advancement in deep neural networks, shifting the focus from data augmentation to symmetry enforcement as a guiding principle in network design through the integration of diverse symmetries—translation, rotation, and scaling. The motivation for symmetry enforcement stems from the observation that symmetric structures exhibit inherent stability and resilience in the face of variations. By integrating CNN architectures with symmetrical characteristics, we aim to reduce the impact of adversarial perturbations on model predictions, thereby fortifying the reliability of these models in real-world settings.

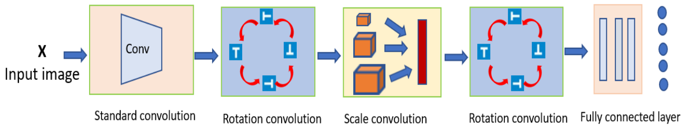

Fig. 1. The proposed symmetry enforcement method for CNNs in which rotation and scale invariant layers are integrated.

II. Proposed Symmetry Enforcement Method for Convolutional Neural Networks

In this work, we aim to enhance generalization in CNNs by integrating symmetries. We utilize several methods to implement symmetry, aiming to approximate perturbation invariance through various invariances attained via symmetry operations within the model. This section introduces the concept of symmetry enforcement, outlining the mechanisms through which symmetry can be effectively integrated. Symmetry in objects means unchanged properties despite transformations like rotation, scaling, or translation. In neural networks, these transformations should yield consistent outputs, treating them as symmetry operations. This principle is crucial for adversarial examples, where even minor perturbations should not change the network output. Achieving this perturbation invariance is essential for improving adversarial robustness.

In our work, we propose enhancing standard CNNs by adding a rotation-equivariant layer and a scale-equivariant layer. This integration aims to impart the network with rotation, scale, and translation invariance. Fig. 1 depicts the proposed symmetry enforcement method in which rotation and scale invariant layers are integrated into the CNN architecture.

A. Rotation Invariance

The rotation invariance is achieved through integrating *G*-convolution [4] (rotationally equivariant convolution) into the deep neural network architecture. Similar to the translational convolution in CNNs, the *G*-convolution applies rotational operations to filters, thereby achieving rotation equivariance. Consequently, this allows different rotated versions of the same input to produce the same output label in the neural network.

B. Scale Invariance

The scale invariance is attained by integrating the scale-invariant convolution layer [5] as a layer in the network. It enables the CNN model to handle input images of varying sizes and generate fixed-size outputs. It works by dividing the feature map (output from the previous convolutional layers) into bins at different scales and applying pooling (usually max pooling) in each bin.

C. Translation Invariance

The translation invariance is inherently integrated into the model since CNNs are specially designed to be translation-invariant. The huge success of CNNs has been substantially attributed to their unique property of translation-invariance. By using standard convolution and pooling operations, the desired translation-invariance property is achieved.

III. Experimental Setup, Results, and Discussion

To assess the effectiveness of our proposed symmetry enforcement technique, we implement a symmetry-enforced CNN architecture and compare it with a baseline standard CNN without symmetry enforcement. The symmetry-enforced model incorporates architectural symmetry as detailed in Section II. We utilize CIFAR-100 as a benchmark dataset which comprises 60,000 32x32 color images in 100 classes for object recognition. The models training was conducted on a single node HPC.

In Fig. 2, the symmetry-enforced CNN model demonstrates better learning and higher accuracy during training, with improved initial generalization on the test set. However, its fluctuating test accuracy suggests possible overfitting or learning instability over time. This model outperforms the standard CNN in training, but the latter shows a significant early drop in test accuracy, indicating less robustness and generalization. On perturbed data, the symmetry-enforced model maintains relatively stable accuracy, hinting at its invariance to noise and distortions. This robustness may stem from its ability to capture essential, invariant data features, indicating a more generalized representation. Despite limited training data, our symmetry-enhanced model demonstrates improved expressive capability, thus increasing its resilience to adversarial alterations. The performance on the random rotated CIFAR 100 data in Fig. 3. The higher accuracy of symmetry-enforced model suggests that it is better at handling the transformations applied to the dataset, likely due to its architecture being more robust to variations in the input data.

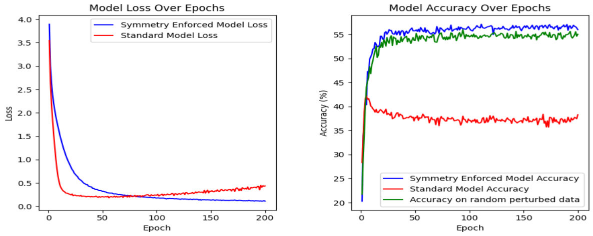

Fig. 2. Comparison using CIFAR 100 data:test accuracy and training convergence

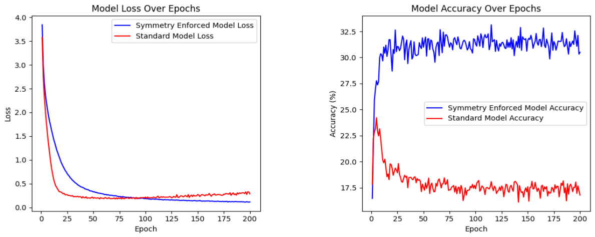

Fig. 3. Comparison using transformed CIFAR 100 data

IV. Conclusion

In this work, we proposed a symmetry enforcement method to enhance the robustness of CNNs against adversarial examples. By incorporating various symmetries, such as rotation and scaling, into existing CNN models to improve their robustness, this approach leads to the development of perturbation invariance within the models. As a result, the enhanced models demonstrate greater generalizability to inputs that are shifted, rotated, or scaled. The introduction of symmetry operations not only optimizes the use of training data but also significantly expands the expressive capabilities of the network, contributing to increased adversarial robustness, and showcasing the potential of symmetry to create more versatile, reliable, and trustworthy AI systems.

References

[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. ”Explaining and harnessing adversarial examples.” arXiv preprint arXiv:1412.6572 (2014).

[2] Taylor, Luke, and Geoff Nitschke. ”Improving deep learning with generic data augmentation.” In 2018 IEEE symposium series on computational intelligence (SSCI), pp. 1542-1547. IEEE, 2018.

[3] Wang, L., Wang, C., Li, Y. and Wang, R., 2021. Improving robustness of deep neural networks via large-difference transformation. Neurocomputing, 450, pp.411-419.

[4] Cohen, Taco, and Max Welling. ”Group equivariant convolutional networks.” In International conference on machine learning, pp. 2990-2999. PMLR, 2016.

[5] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. ”Spatial pyramid pooling in deep convolutional networks for visual recognition.” IEEE transactions on pattern analysis and machine intelligence 37, no. 9 (2015): 1904-1916.