Integrating Non-Classical Receptive Fields of the Primary Visual Cortex into CNNs Enhances Adversarial Robustness

Ehsan Ur Rahman Mohammed (mrahm326@uwo.ca)
Western University, London, Ontario, Canada

Elham Bagheri (ebagher5@uwo.ca)
Western University, London, Ontario, Canada
Vector Institute for Artificial Intelligence, Canada

Soniya (soniya@uwo.ca)
Western University, London, Ontario, Canada

Apurva Narayan (apurva.narayan@uwo.ca)
Western University, London, Ontario, Canada
University of Waterloo, Waterloo, Ontario, Canada

Yalda Mohsenzadeh (ymohsenz@uwo.ca)
Western University, London, Ontario, Canada
Vector Institute for Artificial Intelligence, Canada

Abstract

The adversarial robustness of artificial intelligence models remains a critical challenge for their deployment in real-world applications. Brain-inspired approaches have gained traction recently due to the superior adversarial robustness of human vision. In this study, we propose a novel architecture named nCRF-SurroundNet. nCRF-SurroundNet performs surround modulation on the responses obtained after passing images through a Gabor filter bank. Our enhancements are guided by neurophysiological evidence that receptive fields in the primary visual cortex (V1) extend beyond the classical receptive field—the region directly sampled by a convolutional kernel—to encompass the nonclassical receptive field (nCRF). The nCRF plays a crucial role by inhibiting or amplifying the output of the classical receptive field, depending on the context around the classical receptive field. The proposed model is evaluated against standard convolutional architectures, including AlexNet, ResNet50, and the original VOneNet, using two benchmark datasets, CIFAR10 and ImageNet100. Performance is assessed under the threat of adversarial attacks, specifically using projected gradient descent (PGD) and Carlini and Wagner (C&W) attack methods. The results on CIFAR10 and ImageNet100 demonstrate that our proposed models achieve significantly higher robustness against PGD attacks than existing models, while also maintaining relatively strong performance against CW attacks. In particular, on ImageNet100, our models outperform all baselines under PGD and stronger variants of the C&W attack.

Keywords: adversarial attacks; brain-inspired artificial intelligence; primary visual cortex; receptive fields; non-classical receptive fields; surround modulation; adversarial robustness

Introduction

Deep neural networks remain vulnerable to adversarial attacks—imperceptible perturbations that lead to misclassifications. This vulnerability undermines their deployment in high-stakes applications such as autonomous vehicles and medical diagnostics. In contrast, the human visual system exhibits remarkable resilience to such perturbations. Neuroscientific research attributes this to context-aware processing mechanisms, including the role of non-classical receptive fields (nCRFs) in the primary visual cortex (V1). These nCRFs modulate the response of classical receptive fields (CRFs) based on surrounding visual context, enabling enhanced contour integration and robustness to noise.

Brain-inspired architectures have recently gained traction for improving robustness. VOneNet, for example, augments CNNs with a Gabor filter bank simulating V1 responses. However, it lacks dynamic context modulation. Neuroscientific studies (Spillmann, Dresp-Langley, & Tseng, 2015; David, Vinje, & Gallant, 2004) highlight that nCRFs facilitate sparse coding, scene segmentation, and predictive perception through contextual inhibition or facilitation. In computer vision, early works (Grigorescu, Petkov, & Westenberg, 2003; Wei, Lang, & Zuo, 2013) used nCRF-like mechanisms for contour detection. More recently, push-pull inhibition and predictive coding dynamics have been proposed (Strisciuglio, Lopez-Antequera, & Petkov, 2020; Choksi et al., 2021). However, these models often lack end-to-end integration with deep architectures or do not systematically explore adversarial robustness. Our work bridges this gap by embedding biologically grounded surround modulation into standard CNN pipelines.

We introduce nCRF-SurroundNet, a biologically inspired architecture that explicitly integrates nCRF mechanisms into convolutional neural networks (CNNs). Unlike prior approaches that rely on fixed filter banks (e.g., Gabor filters in VOneNet), our model dynamically modulates CRF responses using surround-derived features, enabling adaptive context-aware feature encoding. We show that this neurophysiological principle not only enhances adversarial robustness but also improves interpretability and generalization to common

corruptions.

Proposed Model

nCRF-SurroundNet integrates a novel surround modulation block at the input stage of CNNs. Input images are convolved using a Gabor filter bank to extract oriented edge features, which are then passed through a non-linear activation and noise injection stage. These CRF responses are modulated using context signals from isotropic Gaussian-based nCRF kernels. Unlike subtractive inhibition, we employ divisive suppression, improving numerical stability and mimicking biological non-linearities.

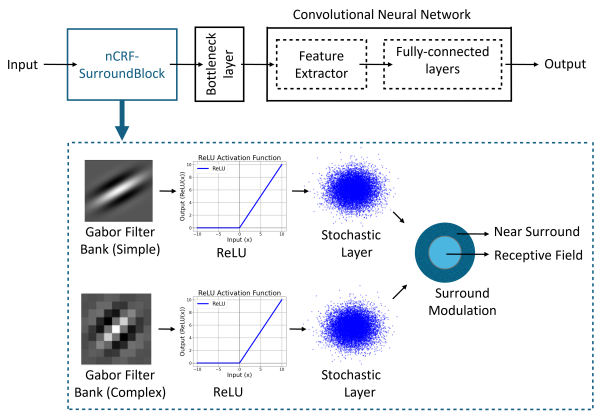

Figure 1: Architecture of the proposed nCRF-SurroundNet model.

Experimental Setup

The architecture was tested on two benchmark datasets: CIFAR-10 and ImageNet100, under strong adversarial threat models, namely Projected Gradient Descent (PGD) (Madry, 2017) and Carlini & Wagner (C&W) (Carlini & Wagner, 2017)attacks. Evaluation metrics included adversarial accuracy across multiple threat levels, feature separability via Linear Discriminant Analysis (LDA) (Zhao, Zhang, Yang, Zhou, & Xu, 2024), and interpretability using ScoreCAM (Wang et al., 2020) heatmaps. Additionally, robustness against 19 types of common corruptions (e.g., blur, noise, JPEG compression) was examined to assess generalization beyond adversarial contexts.

Results

Our experiments reveal the following high-level findings:

- Adversarial Robustness: nCRF-SurroundNet consistently outperforms both standard CNNs and fixed Gabor-based models (e.g., VOneNet) under strong PGD attacks.
- Feature Representation: LDA projections indicate superior inter-class separability and intra-class cohesion in nCRF-SurroundNet compared to baselines.
- Interpretability: ScoreCAM heatmaps show the model attends more precisely to semantically relevant object regions, reducing background sensitivity.
- Generalization: The model demonstrates improved robustness against a wide spectrum of image corruptions, highlighting its potential for real-world deployment.

Conclusion

This study introduces nCRF-SurroundNet, a novel architecture incorporating biologically inspired surround modulation to enhance adversarial robustness. By bridging the gap between classical and non-classical receptive fields in CNNs, the model achieves improved contextual feature processing, interpretability, and resilience. Future work will explore multi-layer surround modulation, integration of feedback pathways, and real-world attack scenarios including physical and patch-based perturbations.

References

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)* (pp. 39–57).

Choksi, B., Mozafari, M., Biggs O'May, C., Ador, B., Alamia, A., & VanRullen, R. (2021). Predify: Augmenting deep neural networks with brain-inspired predictive coding dynamics. *Advances in Neural Information Processing Systems*, *34*, 14069–14083.

David, S. V., Vinje, W. E., & Gallant, J. L. (2004). Natural stimulus statistics alter the receptive field structure of v1 neurons. *Journal of Neuroscience*, *24*(31), 6991–7006.

Grigorescu, C., Petkov, N., & Westenberg, M. A. (2003). Contour detection based on nonclassical receptive field inhibition. *IEEE Transactions on image processing*, *12*(7), 729–739.

Madry, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Spillmann, L., Dresp-Langley, B., & Tseng, C.-H. (2015). Beyond the classical receptive field: The effect of contextual stimuli. *Journal of Vision*, *15*(9), 7–7.

Strisciuglio, N., Lopez-Antequera, M., & Petkov, N. (2020). Enhanced robustness of convolutional networks with a push–pull inhibition layer. *Neural Computing and Applications*, *32*(24), 17957–17971.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., ... Hu, X. (2020). Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops* (pp. 24–25).

Wei, H., Lang, B., & Zuo, Q. (2013). Contour detection model with multi-scale integration based on non-classical receptive field. *Neurocomputing*, *103*, 247–262.

Zhao, S., Zhang, B., Yang, J., Zhou, J., & Xu, Y. (2024). Linear discriminant analysis. *Nature Reviews Methods Primers*, *4*(1), 70.