NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

Visualizing Interpretations of Deep Neural Networks

Applying Interpretation Techniques on ConvNeXt model

Student: Ta Quynh Nga Supervisor: A/P Li Boyang

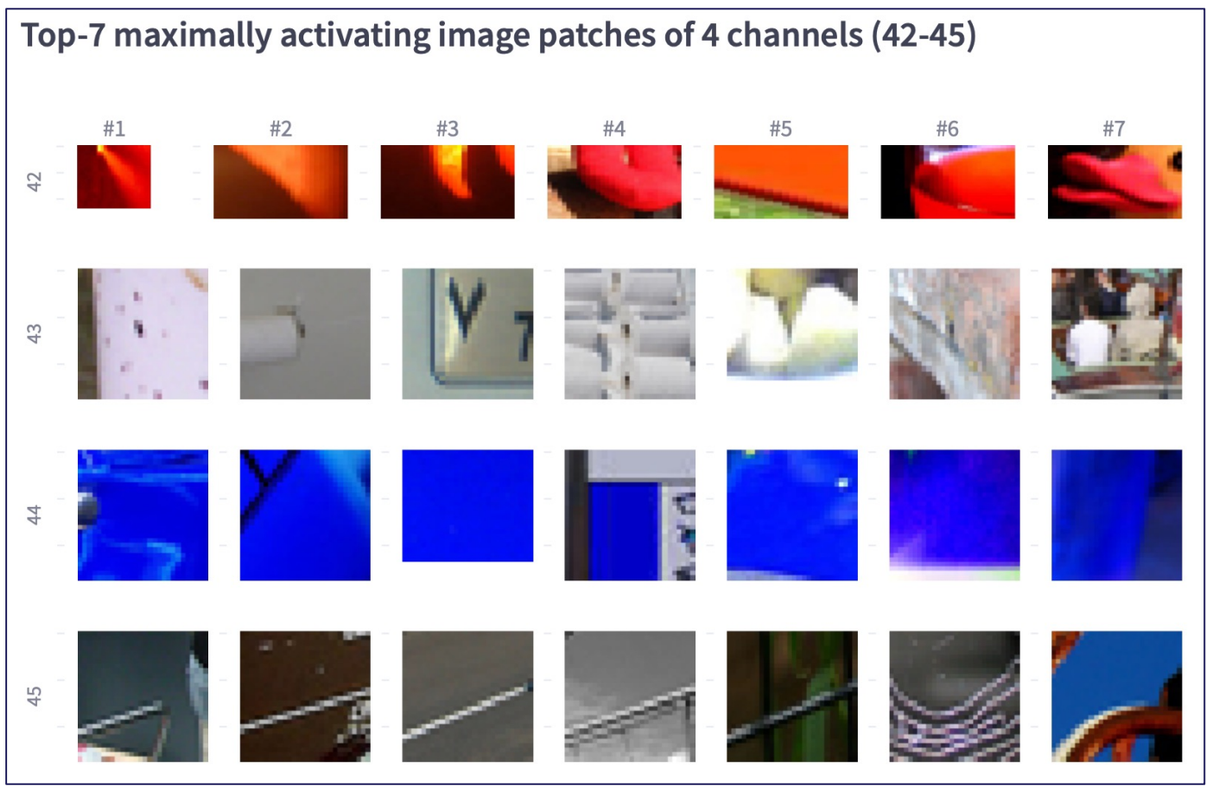
BACKGROUND & PROJECT OBJECTIVES

Convolutional Neural Networks and Vision Transformers have improved computer vision, but their lack of interpretability can have consequences in critical applications. Visualizing deep neural network interpretations can identify biases and errors, improving transparency and trustworthiness. This research is crucial for enhancing the application of deep neural networks in critical domains.

This project aims to create a web application[1] to facilitate the interpretation of the ConvNeXt model, a state-of-the-art convolutional neural network. The application applied three techniques as described below.

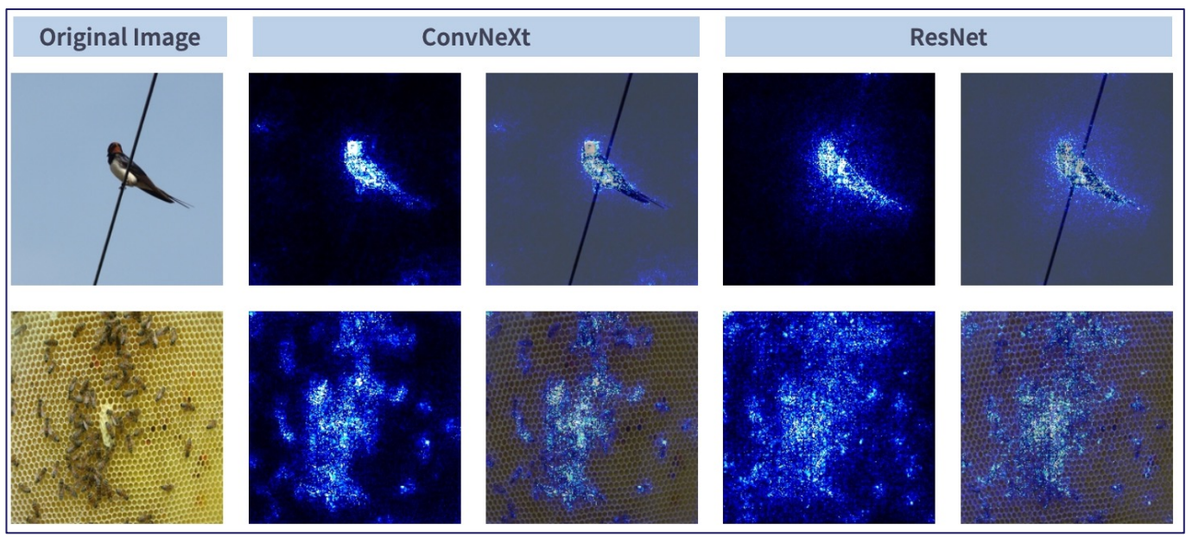
1. Maximally Activating Patches

This technique visualizes the most activating image patches to investigate the patterns or features that maximally activate a filter in a deep neural network layer. This approach provides insight into the model's feature selectivity and learned representations, aiding in a more fine-grained understanding of its inner workings. By identifying these patches, researchers can improve the model or validate existing hypotheses about its behavior.



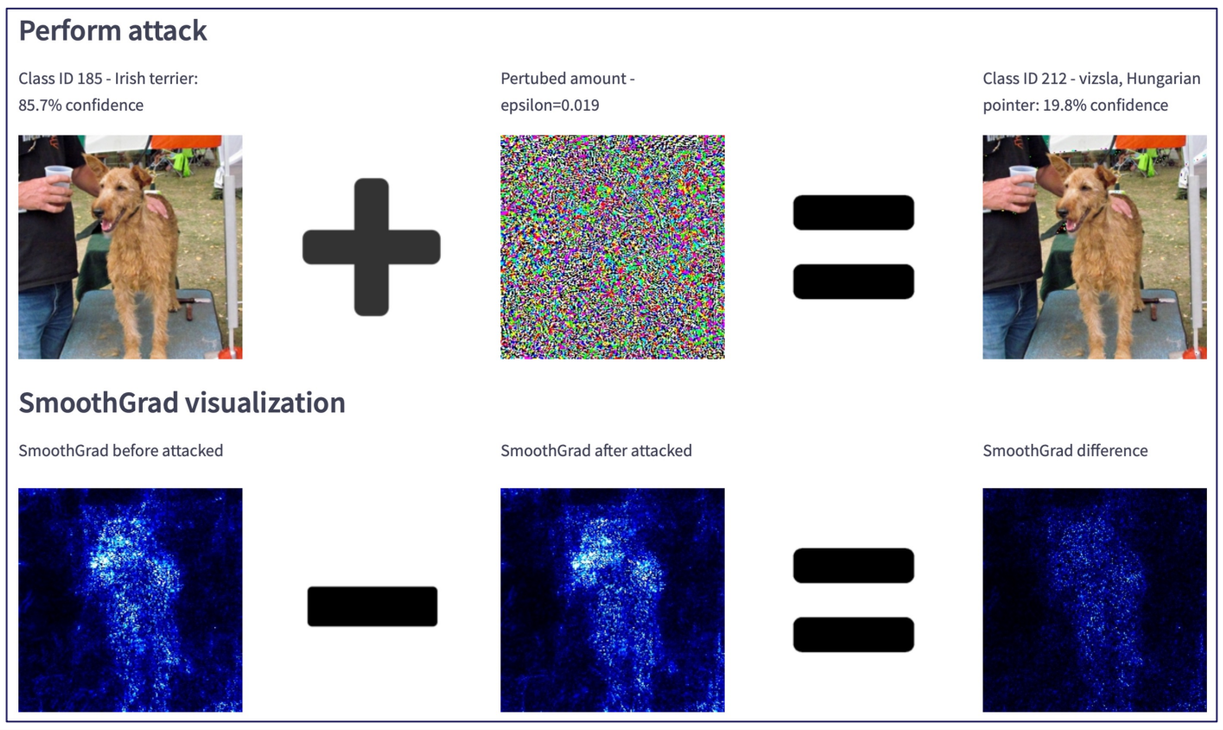
2. Feature Attribution Visualization

The SmoothGrad technique is used to identify input features that affect a model's prediction. It adds Gaussian noise to create smoother saliency maps from gradient information to address noise in the original maps.



3. Adversarial Perturbation Visualization

ConvNeXt model was attacked by FGSM method and then visualized with SmoothGrad technique to investigate the model's vulnerability.



[1] View at https://huggingface.co/spaces/taquynhnga/CNNs-interpretation-visualization