Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Adversarial Attacks on Machine Learning

Semester Thesis

Nastasia Morel

nmorel@ethz.ch

Distributed Computing Group
Computer Engineering and Networks Laboratory
ETH Zürich

Supervisors:
Andreas Plesner
Prof. Dr. Roger Wattenhofer

September 10, 2025

Acknowledgements

I would like to thank my supervisor Andreas Plesner for his support, guidance, and valuable feedback throughout this project. I am also grateful to Prof. Dr. Roger Wattenhofer for giving me the opportunity to explore such a fascinating topic.

Abstract

This semester project explores how adversarial attacks affect the internal neuron activations of deep convolutional neural networks. Using Projected Gradient Descent (PGD), we craft adversarial examples to perturb images from the ImageNet dataset and study their effects on neuron behavior in a ResNet-50 architecture.

We analyze fragile neurons, those whose activations change drastically under perturbations, and identify polysemantic neurons through visualizations generated using the Lucent library. Our experiments reveal that adversarial perturbations not only affect the output predictions but also distort internal representations, often shifting attention to irrelevant spatial regions.

Through activation plots, spatial heatmaps, and feature visualizations, we demonstrate how fragile and distributed internal representations can be. These findings highlight the importance of robustness-aware interpretability methods and open up new paths for studying the interaction between adversarial resilience and internal network structure.

Contents

Acknowledgements i

Abstract ii

1 Adversarial Attacks and Neuron Activation Analysis in Deep Neural Networks 1

1.1 Introduction 1

1.1.1 Motivation 1

1.1.2 Goals 1

1.2 Background 2

1.2.1 Adversarial Attacks 2

1.2.2 Neuron Interpretability 2

1.2.3 Related Work 3

1.3 Methodology 4

1.3.1 Dataset 4

1.3.2 Model Architecture 4

1.3.3 Attack Pipeline 5

1.3.4 Neuron Selection 5

1.3.5 Visualization Tools 6

1.4 Experiments 6

1.4.1 PGD Attacks on Selected Classes 6

1.4.2 Neuron Activation Comparisons 8

1.4.3 Lucent Visualizations 9

1.4.4 Heatmap Analysis 10

1.4.5 Neuron Distribution Across Dataset 11

1.5 Discussion 13

1.5.1 Observed Patterns 13

1.5.2 Polysemantic Neurons 13

1.5.3 Impact of Adversarial Perturbations 13
1.5.4 Limitations and Interpretability Concerns 14
1.6 Conclusion and Future Work 14
1.6.1 Summary of Findings 14
1.6.2 Future Directions 14

Bibliography 15

A ResNet-50 convolution layers A-1

B Neuron Comparison B-1

C Stacked Probability Charts C-1

Chapter 1

Adversarial Attacks and Neuron Activation Analysis in Deep Neural Networks

This semester project explores how adversarial attacks a ect internal neuron activations in deep neural networks. By combining Projected Gradient Descent (PGD) adversarial attacks with feature visualization tools (Lucent), we investigate fragile neurons, polysemantic activations and the spatial characteristics of neuron responses. Experiments are conducted on the ImageNet dataset, with a focus on class transformations (e.g., panda to weasel, truck to animal).

1.1 Introduction

1.1.1 Motivation

Understanding how individual neurons contribute to classi cation decisions can explain the internal mechanics of neural networks. This work extends prior research (e.g., Distill's interpretability articles) by introducing adversarial perturbations and tracking neuron behavior under these conditions.

1.1.2 Goals

- Identify neurons that are highly sensitive or "fragile" to input perturbations.
- Visualize features that maximally activate individual neurons.
- Generate spatial activation heatmaps to localize neuron responses.
- Compare neuron activations across classes (e.g., truck vs. animal) under adversarial transformation.

1.2 Background

1.2.1 Adversarial Attacks

Deep neural networks are vulnerable to adversarial examples, perturbations to input data that are often imperceptible to humans but drastically change the model's prediction. These perturbations are typically made up by maximizing the model's loss with respect to the input image, under a small perturbation budget.

Projected Gradient Descent (PGD)

PGD [1] is one of the most widely used methods for generating adversarial examples. Given an input image x and true label y, PGD iteratively performs gradient ascent on the loss $L(f(x); y)$ with respect to x, followed by a projection step to keep the perturbation within an ℓ_1 ball of radius . Formally:

$$x^{t+1} = \quad {}_{B\ (x)} \quad x^t + \quad \text{sign} \quad r \ {}_x L(f(x^t); y)$$

In our implementation, we applied PGD on images from the ImageNet dataset using a pre-trained ResNet-50 model, observing signi cant label ipping with minimal visible distortion. These adversarial examples serve as the basis for neuron activation comparisons and feature visualization in the remainder of this work.

1.2.2 Neuron Interpretability

To understand how neural networks internally process and represent data, we analyze the activations of individual neurons, either spatially or across channels. Neurons that respond strongly to semantically meaningful features can be interpreted as detectors for those features (e.g., fur texture, vehicle parts, etc.).

Lucent Feature Visualization

We used the Lucent library[1] to synthesize input images that maximally activate speci c neurons in a chosen layer. Lucent performs gradient ascent on a randomly initialized image to maximize a neuron's activation.

Through Lucent, we visualized neurons in the `layer3` and `layer4` blocks of ResNet-50. Some neurons appeared polysemantic, activating for visually distinct concepts like "fur" (weasel) and "metal edges" (truck). This observation aligns with the phenomenon described in the Distill article on feature visualization.

[1] https://github.com/greentfrapp/lucent

Fragile Neurons and Activation Patterns

We de ne fragile neurons as those whose activation changes dramatically when an adversarial perturbation is applied. In practice, we computed the di erence between neuron activations before and after PGD attacks. Some neurons displayed large activation drop-o s or spikes in the presence of adversarial noise, indicating their instability and importance in classi cation decisions.

Furthermore, we explored top-k neuron activations for speci c classes (e.g., weasel, panda, truck) and analyzed how these activations changed under adversarial transformation (e.g., panda! truck). These experiments reveal both robust and fragile patterns of class-speci c neuron behavior.

1.2.3 Related Work

Our work is situated at the intersection of adversarial robustness and neural network interpretability. Several foundational papers have informed our methodology and goals.

Adversarial Examples and Fragile Features. Szegedy et al. [2] were the rst to demonstrate the vulnerability of deep neural networks to adversarial perturbations: small, imperceptible changes that cause models to misclassify inputs. Building on this, Goodfellow et al. [3] introduced the Fast Gradient Sign Method (FGSM) and argued that adversarial examples exploit the excessive linearity of deep models in high-dimensional space. Later, Madry et al. [1] proposed Projected Gradient Descent (PGD) as a universal rst-order adversary and set the stage for adversarial training as a robust defense method.

Ilyas et al. [4] took this further by arguing that adversarial vulnerability is not due to bugs but to the exploitation of non-robust features, features predictive for the model but misaligned with human perception. These works motivate our investigation into how adversarial perturbations a ect internal representations and neuron activations.

Neuron Interpretability and Visualization. Feature visualization has been a key method for understanding internal representations. Olah et al. [5] and [6] presented techniques to synthesize inputs that maximally activate neurons, revealing polysemantic or distributed patterns of neuron behavior. In their follow-up work on curve detectors and circuits [6], they analyzed low-level building blocks of visual recognition models, helping to distinguish between interpretable and fragile units.

Network Dissection [7] o ered a complementary approach by aligning individual neurons with semantic concepts in a labeled dataset (Broden), thus quan-

tifying interpretability. These techniques inspired our use of Lucent to synthesize neuron-preferred features and interpret their spatial activations.

Attribution Graphs and Polysemanticity. Recent work by Anthropic and the Transformer Circuits team [8, 9] has explored the structure of knowledge and interpretability in large models. Although primarily applied to language models, their concepts of monosemantic vs. polysemantic neurons, as well as attribution graphs, aligns with our ndings about polysemantic neurons in vision models. In particular, the presence of neurons that activate for semantically distant concepts (e.g., fur and metal) aligns with their discoveries of neurons supporting multiple latent representations.

Hierarchical Concept Attribution. Wang et al. [10] introduced HINT, a framework for associating neurons with hierarchical concepts such as "dog" "animal." They showed how adversarial attacks can disrupt such associations, leading to misclassi cations. Our work similarly investigates fragile neurons and activation shifts, focusing on spatial localization and class transitions (e.g., panda ! truck) rather than hierarchical taxonomy.

Together, these prior works underscore the importance of understanding internal model representations, not just for interpretability, but also as a critical factor in adversarial robustness. Our project contributes by bridging these elds through systematic visual and activation-based analysis of neuron responses under adversarial perturbations.

1.3 Methodology

1.3.1 Dataset

We use the ImageNet dataset with a set of 1000 classes. To explore neuron behavior under adversarial conditions, we generate adversarial examples using PGD. Speci c attention is given to a selected subset of classes, notably weasel and trailer truck , which are used to study class-targeted adversarial transformations and neuron sensitivity.

1.3.2 Model Architecture

All experiments are performed using the ResNet-50 architecture, with both the standard and a robustly trained variant. Robust models are essential for comparing neuron behaviors under adversarial perturbations.

ResNet-50 is a deep convolutional neural network composed of an initial convolution and max-pooling layer, followed by four main stages of residual blocks

(layer1 layer4) that progressively increase the feature dimensionality (256, 512, 1024, 2048) 1.1. Each stage is made of bottleneck blocks that employ shortcut connections, allowing gradients to ow more easily through the network and enabling very deep training. The network ends with a global average pooling and a fully connected classi cation layer producing 1000 ImageNet outputs. A compact outline of all the convolution layers is provided in Appendix A.

Figure 1.1: Resnet-50's architecture

1.3.3 Attack Pipeline

Adversarial examples are generated using the PGD attack with an l_1 norm constraint. The pipeline is implemented in PyTorch and follows the standard iterative formulation. The core steps are:

- Baseline inference: Record original model predictions and internal activations.
- Adversarial generation: Craft PGD adversarial images that either reduce classi cation con dence or force the model to misclassify inputs as a speci c target class.
- Class swap attacks: Design transformations such as weasel truck and truck ! weasel to observe neuron-level e ects.
- Activation comparison: Compute and compare neuron activations before and after attack to identify fragile units.

1.3.4 Neuron Selection

Neuron activations are analyzed per convolutional layer, focusing on identifying:

- Top-k neurons: For each class (e.g., weasel), we identify the top 10% of neurons with the highest average activation across validation samples.

- Fragile neurons: Neurons that show a large activation drop or spike between clean and adversarial examples.
- Polysemantic neurons: Neurons that activate strongly for semantically distant classes (e.g., bothtruck and weasel), possibly indicating feature reuse or co-adaptation.

1.3.5 Visualization Tools

To interpret internal neuron representations and their robustness, we use a suite of visualization techniques:

- Lucent: Activation maximization is performed on selected neurons using the Lucent library. Synthetic images are generated to visualize the preferred input patterns of each neuron.
- Activation plots: Bar plots are created to compare average neuron activations across clean, adversarial, and target-class inputs. These allow for quick identi cation of fragile or polysemantic neurons.
- Heatmaps: Spatial activation maps are generated using feature maps to highlight where neurons re within an image. This is useful for comparing attention regions before and after attack.
- Activation histograms: Aggregated histograms across the dataset help assess neuron selectivity and class speci city.

All visualizations are created for both standard and adversarial inputs to allow direct comparisons. Activation maps are normalized layer-wise for consistency.

1.4 Experiments

This section presents a series of experiments performed on a ResNet-50 model trained on the ImageNet dataset. We investigate the e ect of adversarial perturbations on neuron activation patterns, visualizations, and feature localization. Experiments target speci c class transformations, focusing primarily onweasel and trailer truck .

1.4.1 PGD Attacks on Selected Classes

To study how adversarial attacks a ect the internal behavior of the network, we applied PGD to perturb input images and force class misclassi cation.