Evaluating Neural Network Robustness against FGSM and PGD Adversarial Attacks with L Norms Perturbations

PRESENTER: Ahmeed A. Yinusa
INSTRUCTOR: Dr. Joshua L. Phillips

AGENDA

INTRODUCTION

Artificial Intelligence (AI), with the integration of Deep Learning (DL) has resulted in notable and significant advancements in modern data-driven technologies, such as autonomous vehicle systems, image recognition and classification, and natural language processing,

Deep Learning Architectures

- Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks (RNNs)
- Long Short-Term Memory Networks (LSTMs)
- Transformers
- DenseNet
- Autoencoders, etc.

Adversarial Attacks

INTRODUCTION CONT'D

Attack Technique	Type
FGSM (Fast Gradient Sign Method)	White-box, Non-targeted
PGD (Projected Gradient Descent)	Iterative White-box
Carlini-Wagner	Optimization-based White-box
DeepFool	White-box
JSMA (Jacobian-based Saliency Map Attack)	White-box
One-Pixel Attack	White-box
Boundary Attack	White-box
Spatial Transformation Attack	White-box or Black-box
Universal Adversarial Perturbations	White-box or Black-box
Data Poisoning Attack	White-box or Black-box

Table 1: Adversarial Techniques and Types

INTRODUCTION CONT'D

Real-World Scenarios

Researchers tricked an autonomous driving AI into misreading fake signs and exceeding speed limits, showcasing adversarial attacks through physical means.

Figure 1: Autonomous Vehicle [1].

Microsoft's AI chatbot Tay began tweeting offensive content after a data-poisoning attack in 2016, revealing the vulnerability of AI to malicious inputs.

Figure 2: Attacked Microsoft Chatbot(Tay Bot) [2].

INTRODUCTION CONT'D

Problem Statement

- Neural networks show high error rates against adversarial attacks.
- Attacks can undermine safety-critical systems relying on AI.

Research Objectives

- Evaluate DenseNet-161 robustness against FGSM and PGD using L_1, L_2, L_∞ norms.
- Examine the impact of adversarial training on model resilience.

Research Questions

- How robust is DenseNet-161 against FGSM and PGD attacks with L norms?
- Does adversarial training improve resilience?

RELATED WORK

Section	Key Themes	Findings and Contributions
The Inception and Evolution of Adversarial Machine Learning	Goodfellow et al. (2014) [3] discuss neural networks' vulnerability to adversarial examples, attributing it to their linear nature. They suggest adversarial training as a method to reduce error rates.	Pioneered the understanding of linearity in neural networks and its role in adversarial vulnerability; Developed adversarial training techniques.
	Madry et al., (2017) [4] investigate the vulnerability of deep neural networks to adversarial examples and propose a robust optimization framework	Introduced a methodology to enhance adversarial robustness in deep learning models.
	Huang et al., (2022) [5] enhance PGD and C&W algorithms for adversarial attacks targeting tram object detectors, demonstrating quick and effective attacks	Advanced the understanding of adversarial threats in public transportation safety.
Norm-Based Perturbations: A Spectrum of Attacks	Research on L_∞ norm perturbations is predominant, but L1 and L2 norms offer different perspectives and attack strategies	Broadened the understanding of adversarial perturbations beyond the commonly focused L_∞ norm
The Quest for Robustness Across Norms	Focus on L_∞ norm defenses leaves gaps against L1 and L2 norm attacks, suggesting a false sense of security in models.	Identified a critical research gap in defending against diverse norm-based adversarial attacks
Gaps & Combined Approach	The research evaluates DenseNet161's defense against FGSM and PGD attacks (using L1, L2, and L_∞ norms) on the Stanford Dogs dataset.	The study aims to enhance the understanding of adversarial robustness in detailed image classification, helping to create broader defense strategies against various adversarial attacks.

Table 2: Related Work

METHODOLOGY

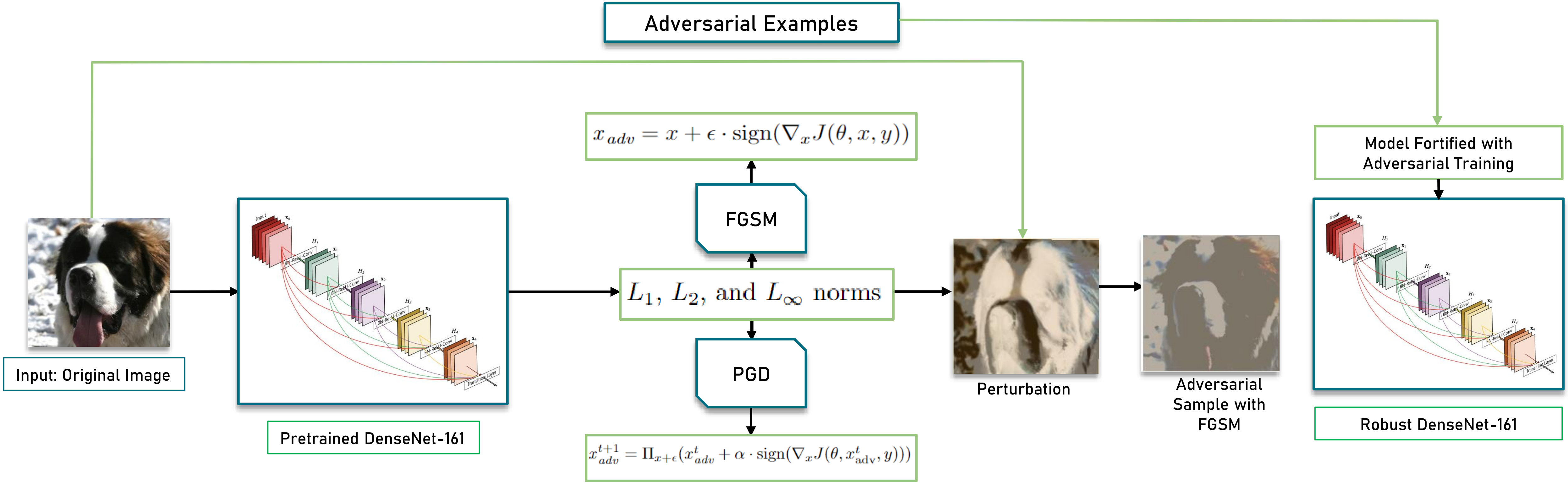


Figure 3: Research Process Design

DATA COLLECTION

Stanford Dogs Detail

The Stanford Dogs dataset provides a platform to evaluate the effectiveness of DenseNet-161 in a fine-grained visual categorization task and procedure.

Aspect	Detail
Dataset	Stanford Dogs
Number of Categories	120
Number of Images	20,580
Annotations	Class labels, Bounding boxes
Lists	Lists, with train/test splits (0.5MB)
Train Features	Train Features (1.2GB)
Test Features	Test Features (850MB)

Table 3: Stanford Dogs Dataset Detail.

Figure 4: Stanford Dogs Image Samples [6].

MODEL ARCHITECTURE

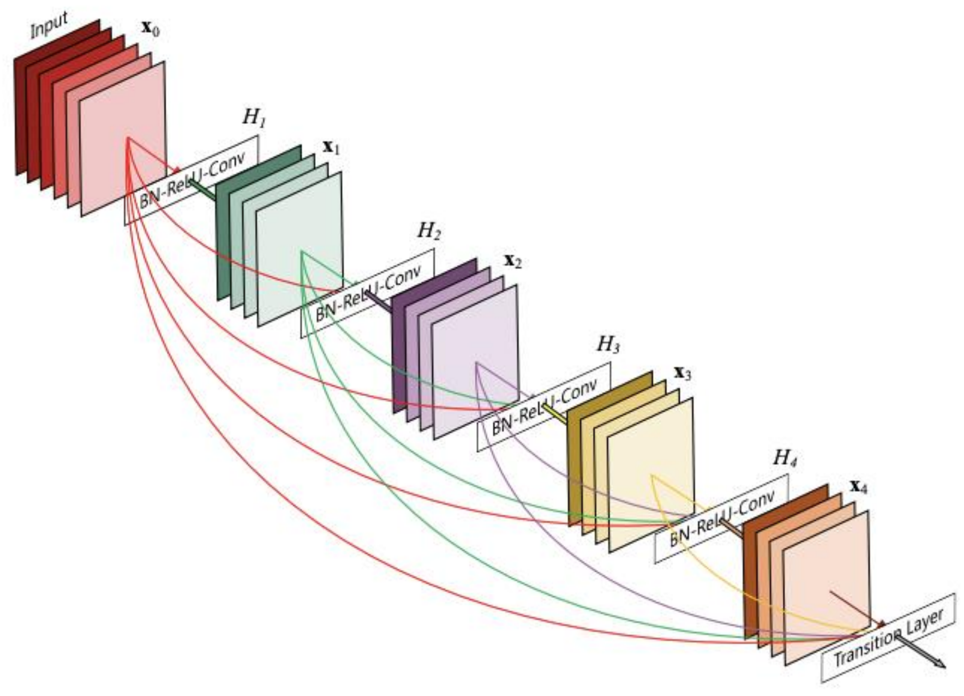
DenseNet-161: Architecture

DenseNet-161, a Dense Convolutional Network (DenseNet) variant, represents a significant advancement in deep learning architectures. DenseNet-161 is distinguished by its depth and complexity, incorporating 161 layers, and is specifically engineered to optimize parameter efficiency.

Features

- Dense Connectivity
- Efficient Feature Propagation
- Feature Concatenation
- Reduced Number of Parameters

MODEL HYPERPARAMETERS: Learning Rate = 0.001, Max-Epoch = 5, Batch-Size = 32, Optimizer = Adam



$$x_l = H_l([x_0, x_1, ..., x_{l-1}])$$

Here, x_l denotes the output feature maps of the lth layer.

Figure 5: A dense block comprising five layers [8].

ADVERSARIAL ATTACKS

Fast Gradient Sign Method (FGSM)

FGSM is an attack technique that generates adversarial examples by exploiting the gradients of the neural network. The goal is to create a new image, x_{adv}, that is visually similar to the original image, x, but is classified incorrectly by the network.

The adversarial image x_{adv} is computed as follows:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$$

where:

- x: Original input image.
- x_{adv}: Adversarial image.
- ϵ: Perturbation magnitude, controlling how much the input is altered.
- $\nabla_x J(\theta, x, y)$: Gradient of the model's loss function concerning the input image.
- J: Loss function used by the neural network.
- θ: Parameters (weights) of the model.
- y: True label of the input image.
- $\text{sign}(\cdot)$: Sign function that extracts the sign of the gradient.

The $\text{sign}(\cdot)$ function ensures that the perturbation is minimal but effective, causing the input to cross the decision boundary.

Projected Gradient Descent (PGD)

PGD is an iterative attack method that extends FGSM. PGD is often considered more powerful as it applies small perturbations iteratively, allowing for a more fine-grained search for adversarial examples. The adversarial image at each iteration $t + 1$ is calculated as:

$$x_{\text{adv}}^{t+1} = \Pi_{x+\epsilon}(x_{\text{adv}}^t + \alpha \cdot \text{sign}(\nabla_x J(\theta, x_{\text{adv}}^t, y)))$$

where:

- x_{adv}^{t+1}: Adversarial image at iteration $t + 1$.
- x_{adv}^t: Adversarial image at iteration t.
- α: Step size for each iteration.
- $\Pi_{x+\epsilon}(\cdot)$: Projection function ensuring the adversarial image remains within an ϵ-ball of the original image.

PARAMETERS USED: Epsilon = 0.003, Alpha = 0.01, Iteration = 10

THE L NORMS

The L norms, specifically L_1, L_2, and L_∞, are crucial in defining the nature of perturbations applied in adversarial attacks like FGSM and PGD. Each norm provides a different way to measure the perturbations' magnitude, thereby shaping the characteristics of the adversarial examples generated.

L_1 Norm (Manhattan Norm)

Mathematically: $||\delta||_1 = \sum_i |\delta_i|$

The L_1 norm of a perturbation is the sum of the absolute values of its vector elements. In the context of images, it represents the sum of absolute differences across all pixels.

L_2 Norm (Euclidean Norm)

Mathematically: $||\delta||_2 = \sqrt{\sum_i \delta_i^2}$

The L_2 norm is the square root of the sum of the squares of the vector elements, equivalent to the Euclidean distance from the origin.

L_∞ Norm (Maximum Norm)

Mathematically: $||\delta||_\infty = \max_i |\delta_i|$

The L_∞ norm is the maximum absolute value of the elements of the vector. In image perturbation, it limits the maximum change that can be applied to any pixel.

ADVERSARIAL TRAINING

Deep Learning Resilient Techniques

- Adversarial Training
- Defense Layers
- Randomization
- Ensemble Methods
- Feature Denoising
- Defensive Distillation
- Input Preprocessing
- Robust Optimizers

Adversarial Training Algorithm

Result: Train the model with enhanced robustness using adversarial training
Initialize: Model parameters;
while *training* **do**
  Get a batch of data (x, y);
  Forward Pass: Compute logits $logits = model(x)$;
  Compute Loss: $loss = CrossEntropyLoss(logits, y)$;
  Log training loss;
  Calculate and log training accuracy;
  if $current\ epoch \geq adv\_training\_start\_epoch$ **then**
    Adversarial Training:

1. Compute gradient w.r.t input data: $data\_grad = \nabla_x loss$;
2. Generate adversarial examples: $x_{adv} = fgsm\_attack(x, \epsilon, data\_grad)$;
3. Forward pass with adversarial examples: $logits_{adv} = model(x_{adv})$;
4. Compute loss for adversarial examples: $loss_{adv} = CrossEntropyLoss(logits_{adv}, y)$;
5. Log adversarial training loss;
6. Compute combined loss: $combined\_loss = loss + loss_{adv}$;
7. Backward pass and update model parameters using $combined\_loss$;

  else
    Standard Training:

1. Backward pass and update model parameters using $loss$;

  end
end

Algorithm 3: Adversarial Training for Dog Breed Classifier

RESULTS

The research shows DenseNet-161, trained on the Stanford Dogs dataset, achieving optimal performance in the second epoch with high precision and accuracy, indicating its strong capability in fine-grained image classification of dog breeds.

Metric Performance without Adversarial Attacks

Epoch	*Validation Loss*	*Accuracy*	*F1 Score*	*Precision*
1	0.7362	0.7840	0.7798	0.8127
2	0.6460	0.8020	0.7971	0.8230
3	0.6998	0.7932	0.7858	0.8206
4	0.6884	0.7928	0.7859	0.8180
5	0.6914	0.7959	0.7920	0.8159

Table 4: Model Metric Performance

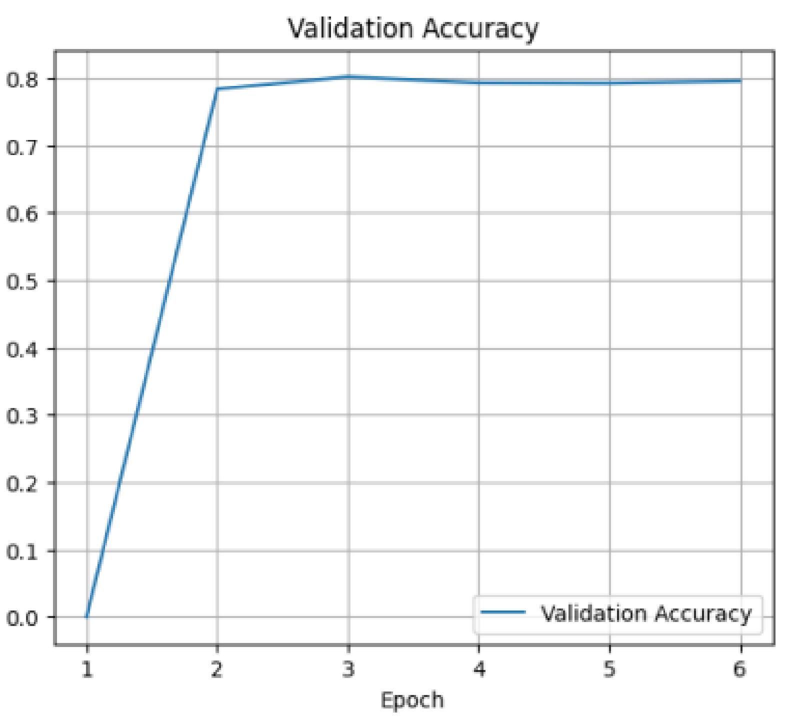
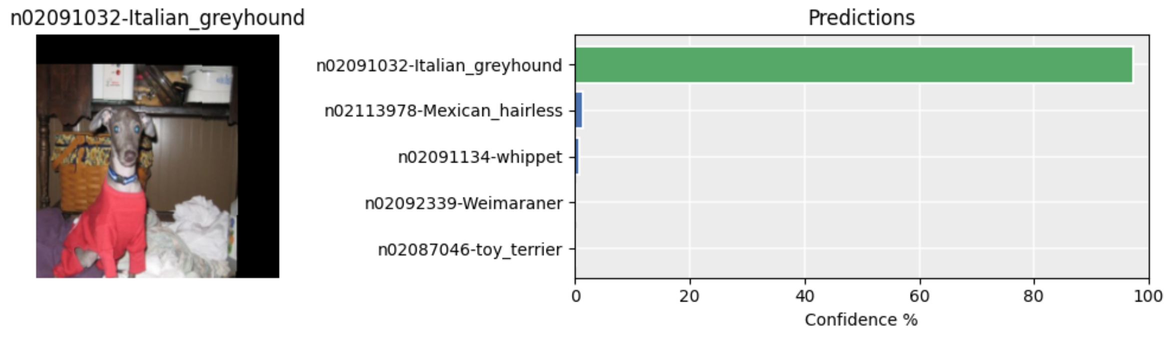
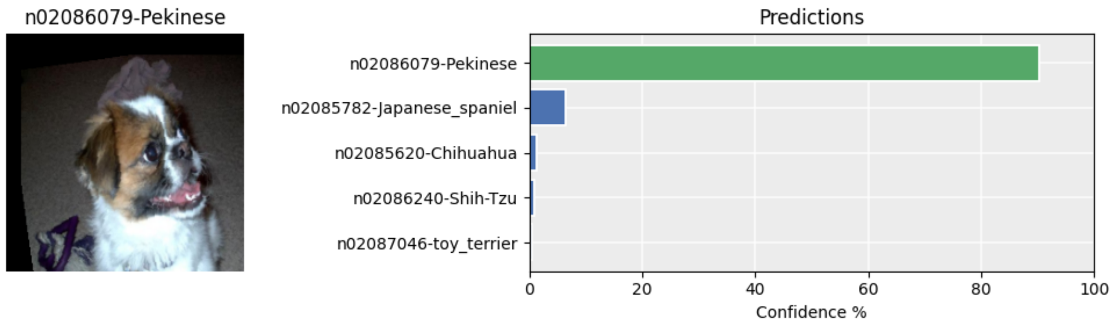


Figure 5: Graph of Model's Validation Accuracy

Model Prediction without Adversarial Attacks with Confidence Score in Percentage





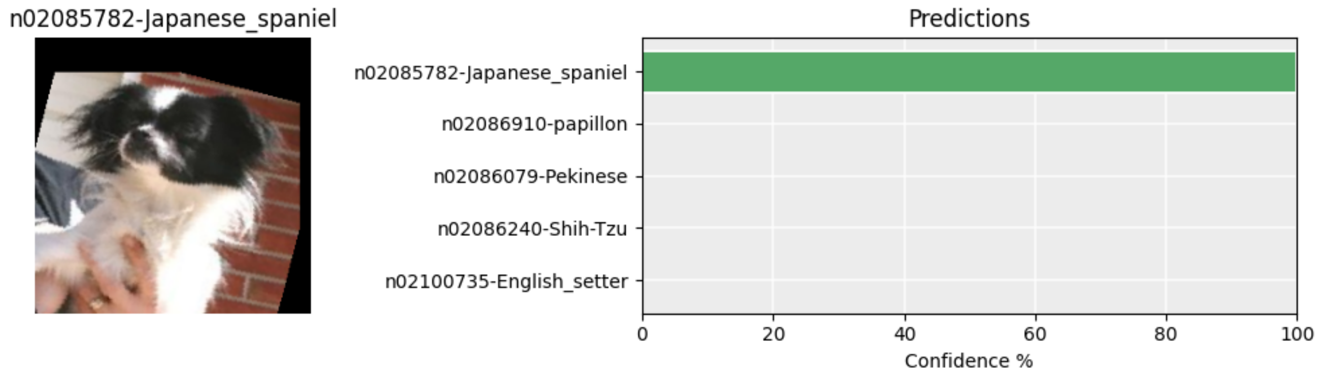


Figure 6 : Dog Image Prediction Confidence Scores

RESULTS CONT'D

Model Prediction with FGSM Adversarial Attacks with Confidence Score in Percentage

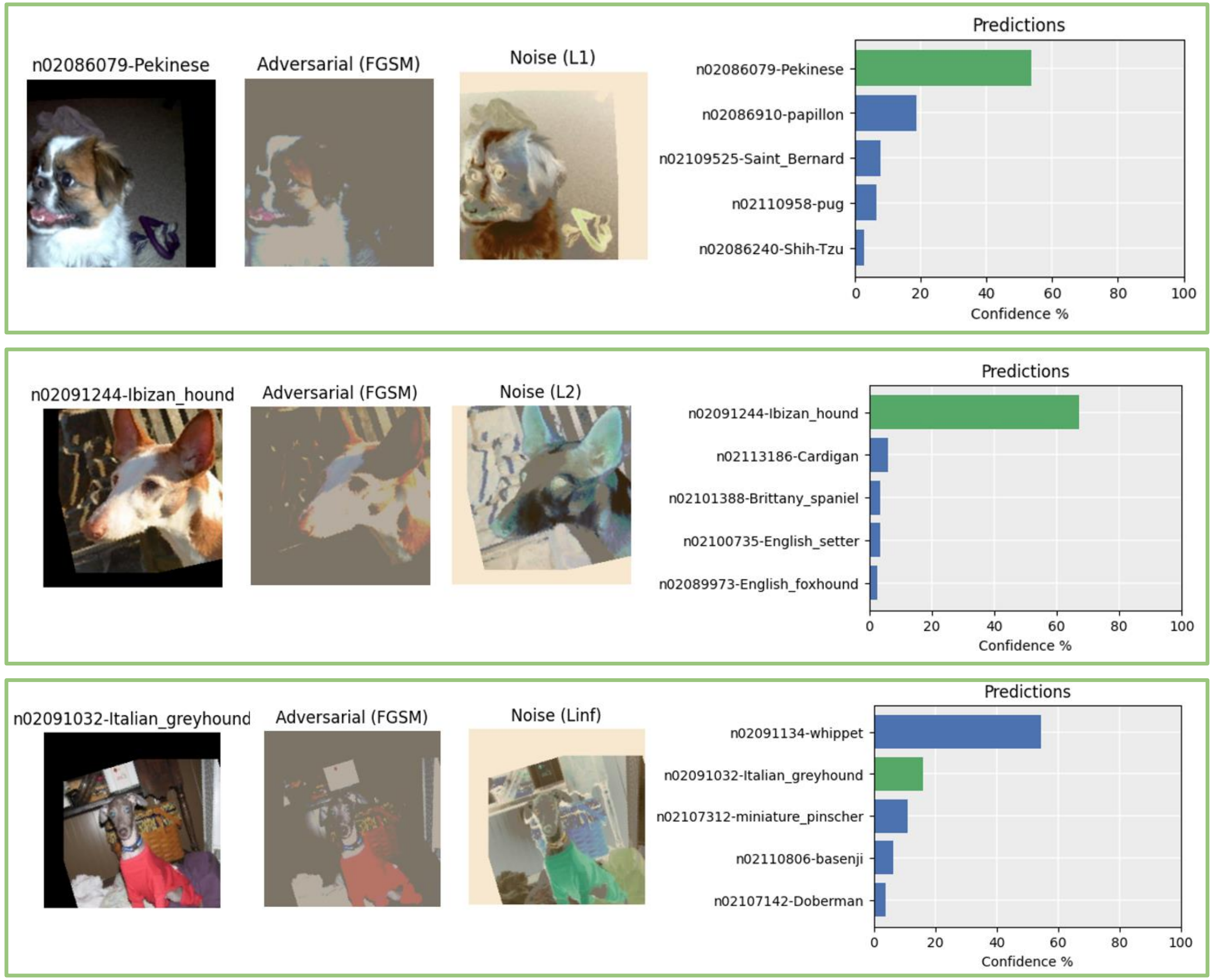


Figure 7 : Dog Image Prediction Confidence Scores with FGSM Adversarial Attacks

RESULTS CONT'D

Model Prediction with PGD Adversarial Attacks with Confidence Score in Percentage

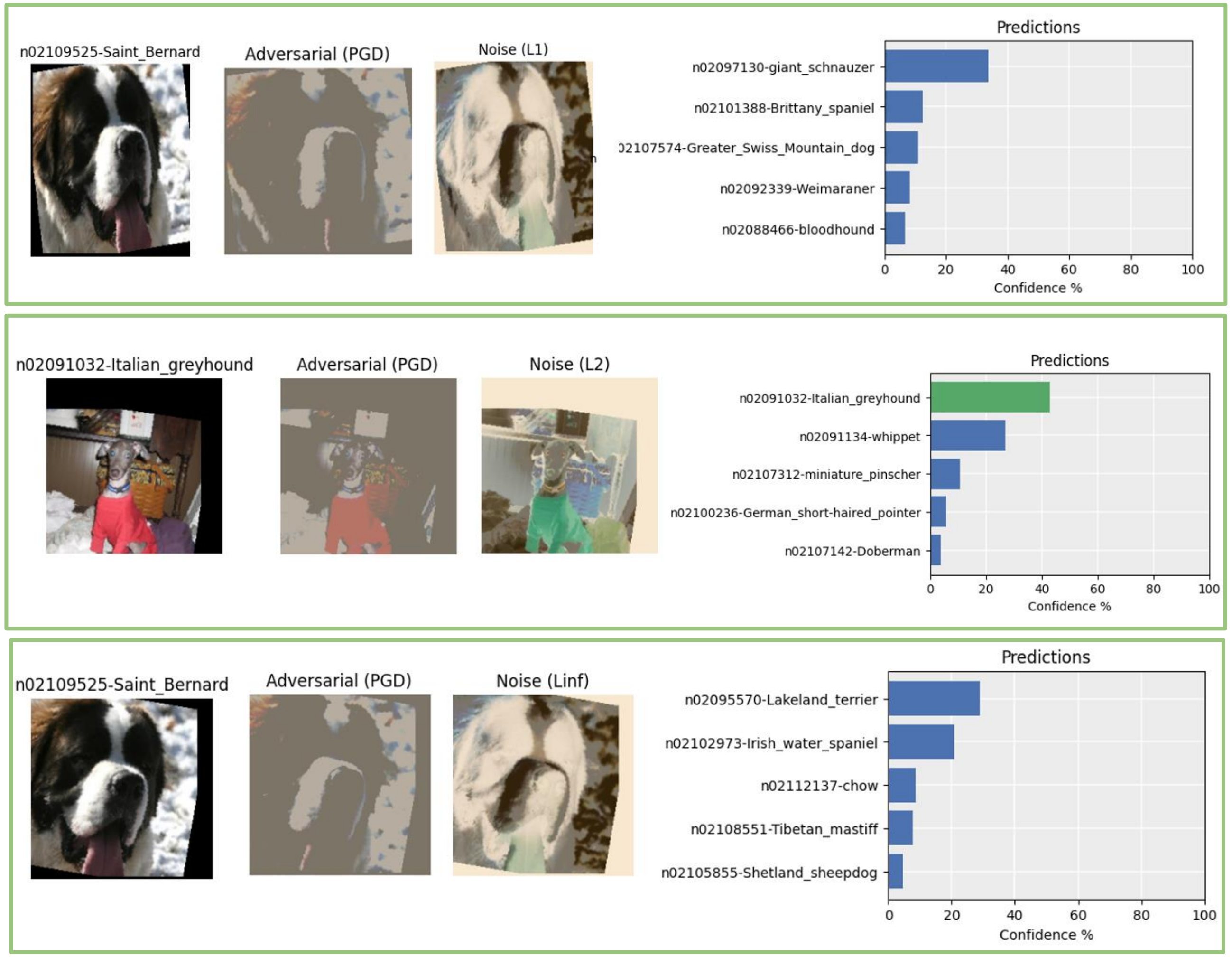


Figure 8 : Dog Image Prediction Confidence Scores with PGD Adversarial Attacks

RESULTS CONT'D

Model Metric Performance After Adversarial Training with Adversarial Attacks Implementation

Attack Type	Norm	Accuracy
Clean	None	79.37%
FGSM	L_∞	17.78%
PGD	L_1	19.61%
PGD	L_2	16.06%
PGD	L_∞	4.03%

Table 5 : FGSM & PGD Adversarial Attacks on Model Before Training

Attack Type	Norm	Accuracy
Clean	None	79.59%
FGSM	L_∞	31.05%
PGD	L_1	34.11%
PGD	L_2	29.32%
PGD	L_∞	9.79%

Table 6 : FGSM & PGD Adversarial Attacks on Model After Training

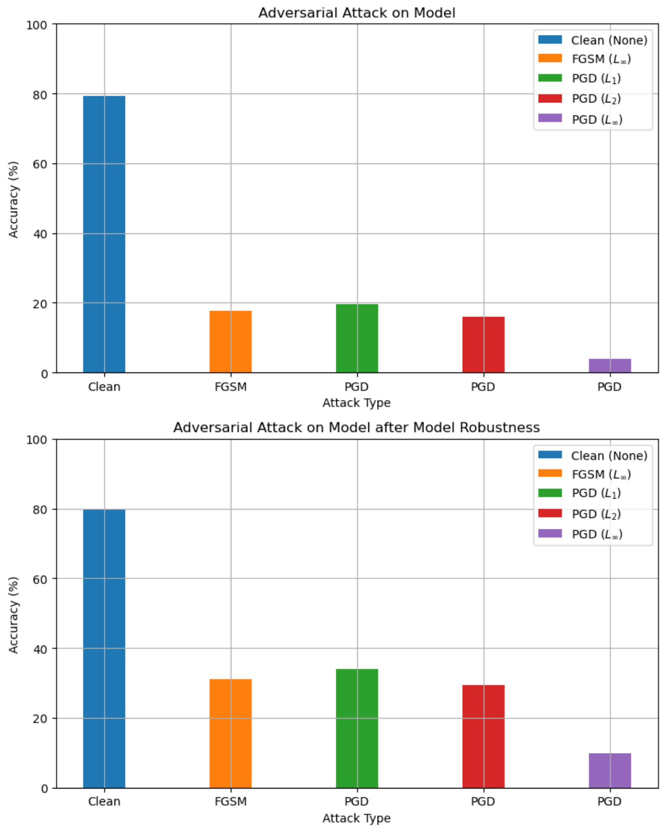


Figure 8 : Distribution of Adversarial Attacks Before and After Training

CONCLUSION

The study shows DenseNet161's vulnerability to adversarial attacks, notably PGD, with accuracy dropping from 79% to 4%. Adversarial training did make the model somewhat more robust, yet significant gaps remain, especially against L_∞ norm attacks, highlighting the urgent need for more advanced security in AI for real-world resilience.

REFERENCES

- Jamie. (2021, December 14). *Subrogating Automated Driving Systems & Vehicle Failures | MWL*. Matthiesen, Wickert & Lehrer S.C. https://www.mwl-law.com/subrogating-automated-driving-systems-and-autonomous-vehicle-failures/
- Schwartz, O. (2019, November 25). *In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation*. IEEE Spectrum. https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation
- J. Goodfellow, J. Shlens, C. Szegedy, Explaining and Harnessing Adversarial Examples, arXiv preprint arXiv:1412.6572, 2014
- Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards Deep Learning Models Resistant to Adversarial Attacks, arXiv preprint arXiv:1706.06083, 2017.
- G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely Connected Convolutional Networks, arXiv preprint arXiv:1608.06993, 2022.
- Khosla, N. Jayadevaprakash, B. Yao, L. Fei-Fei, Novel Dataset for Fine-Grained Image Categorization: Stanford Dogs, In Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC), 2011.
- Papers with Code - DenseNet, Paperswithcode.com, Available: https://paperswithcode.com/lib/torchvision/densenet
- Welcome to PyTorch Lightning — PyTorch Lightning 2.0.6 documentation, Lightning.ai, Available: https://lightning.ai/docs/pytorch/stable/

THANK YOU