I-FGSM Attack Analysis: Strengthening Adversarial Examples via Iterative Refinement

Project Portfolio 2

Michael Anggi G.A

I-FGSM Attack Analysis: Strengthening Adversarial Examples via Iterative Refinement

Background

- This project involved implementing I-FGSM adversarial attacks on a ResNet18 model trained from scratch on CIFAR-10 dataset.

- The goal was to bridge the gap between theoretical knowledge and hands-on implementation in AI security.

- **Following our Portfolio Project-1 on FGSM, this work extends the research by implementing the Iterative Fast Gradient Sign Method (I-FGSM) for enhanced adversarial attack generation**

I-FGSM Attack Analysis: Strengthening Adversarial Examples via Iterative Refinement

Basic Theory

- **While the Fast Gradient Sign Method (FGSM) provided an efficient approach for generating adversarial examples through single-step gradient computation, researchers found that its one-step nature limited the attack's effectiveness against more robust models.**

- **Building upon FGSM's foundation, Kurakin et al. (2017) introduced the Iterative Fast Gradient Sign Method (I-FGSM) to overcome these limitations by applying the gradient sign method iteratively with smaller step sizes.**

- **This iterative approach allows I-FGSM to generate more potent adversarial examples by refining perturbations through multiple iterations, resulting in higher attack success rates while maintaining the computational efficiency that made FGSM popular among researchers and practitioners.**

I-FGSM Attack Analysis: Strengthening Adversarial Examples via Iterative Refinement

Reference

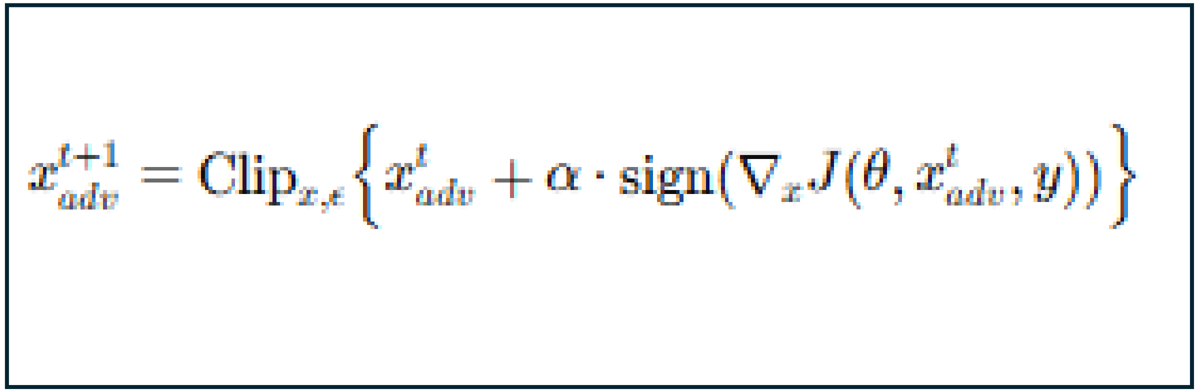
- *"The basic iterative method is a natural extension of FGSM. Instead of taking one large step in the direction of the gradient, we take many smaller steps and clip the intermediate result after each step to ensure that it stays within the ε-neighborhood of the original image." — (Kurakin et al., 2017, p. 8)*

- *"We found that iterative attacks are more successful than single-step attacks when the adversarial examples are crafted on one model and tested on a different model (i.e., in the transfer setting)." — (Kurakin et al., 2017, p. 1)*

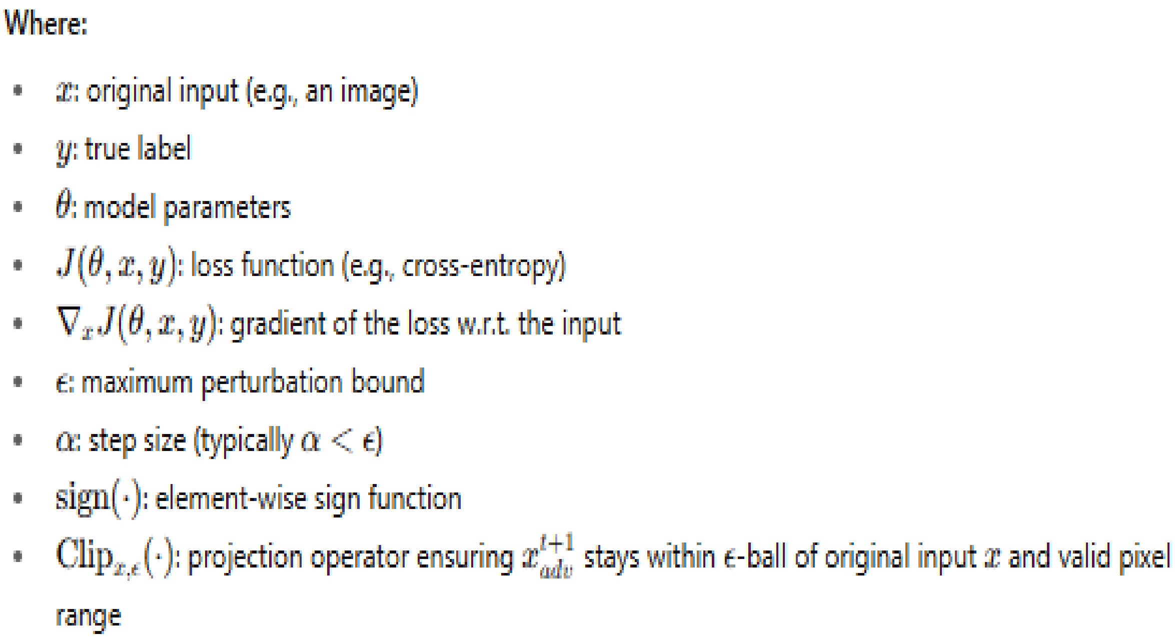
I-FGSM Attack Analysis: Strengthening Adversarial Examples via Iterative Refinement

Basic Equation

The I-FGSM (also known as Basic Iterative Method, BIM) extends FGSM by applying the perturbation multiple times with a small step size. After each iteration, the adversarial example is clipped to remain within a specified perturbation bound. This iterative approach refines the perturbation, often making the attack stronger than a single-step FGSM.



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



Where:

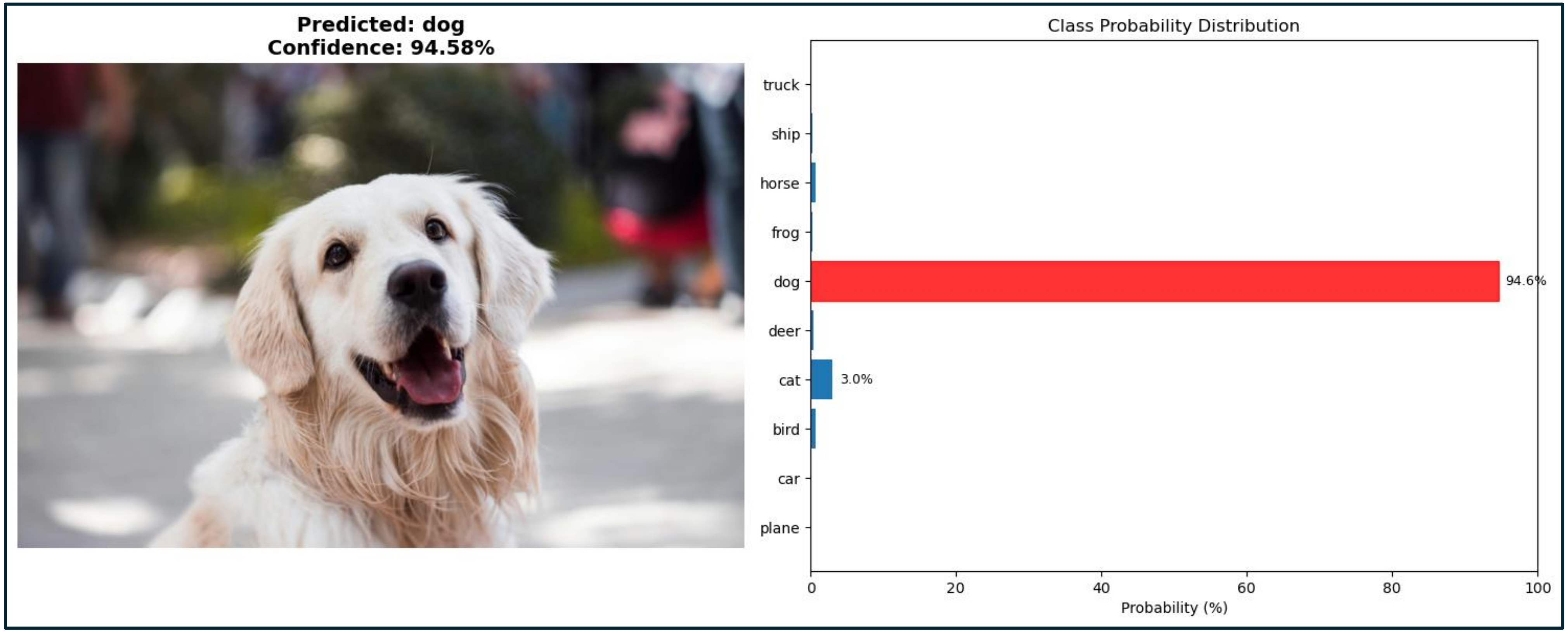
- x: original input (e.g., an image)
- y: true label
- θ: model parameters
- $J(\theta, x, y)$: loss function (e.g., cross-entropy)
- $\nabla_x J(\theta, x, y)$: gradient of the loss w.r.t. the input
- ϵ: maximum perturbation bound
- α: step size (typically $\alpha < \epsilon$)
- $\text{sign}(\cdot)$: element-wise sign function
- $\text{Clip}_{x,\epsilon}(\cdot)$: projection operator ensuring x_{adv}^{t+1} stays within ϵ-ball of original input x and valid pixel range

I-FGSM Attack Analysis: Strengthening Adversarial Examples via Iterative Refinement

Michael Anggi G.A.

Basic Model

A ResNet18 model was trained on the CIFAR-10 dataset, and the input image was classified as a ‘dog’ with 94.6% confidence.



I-FGSM Attack Analysis: Strengthening Adversarial Examples via Iterative Refinement

Michael Anggi G.A.

I- FGSM Attack

- The I-FGSM attack achieves a remarkable 94% success rate (n=16) with ε=0.10, demonstrating significant improvement over smaller epsilon values where success rates remained at 0%.

- The iterative approach shows consistent performance across 10, 20, and 40 iterations with a stable success rate of 17% (n=24) for each iteration count, indicating that the attack converges quickly and additional iterations beyond 10 may not provide substantial benefits.

- The alpha ratio analysis reveals that I-FGSM maintains consistent attack effectiveness across different step sizes (α/ε ratios of 0.10, 0.20, 0.25, and 0.50), all achieving 17% success rate (n=24), suggesting robustness to hyperparameter variations.

I-FGSM Attack Analysis: Strengthening Adversarial Examples via Iterative Refinement

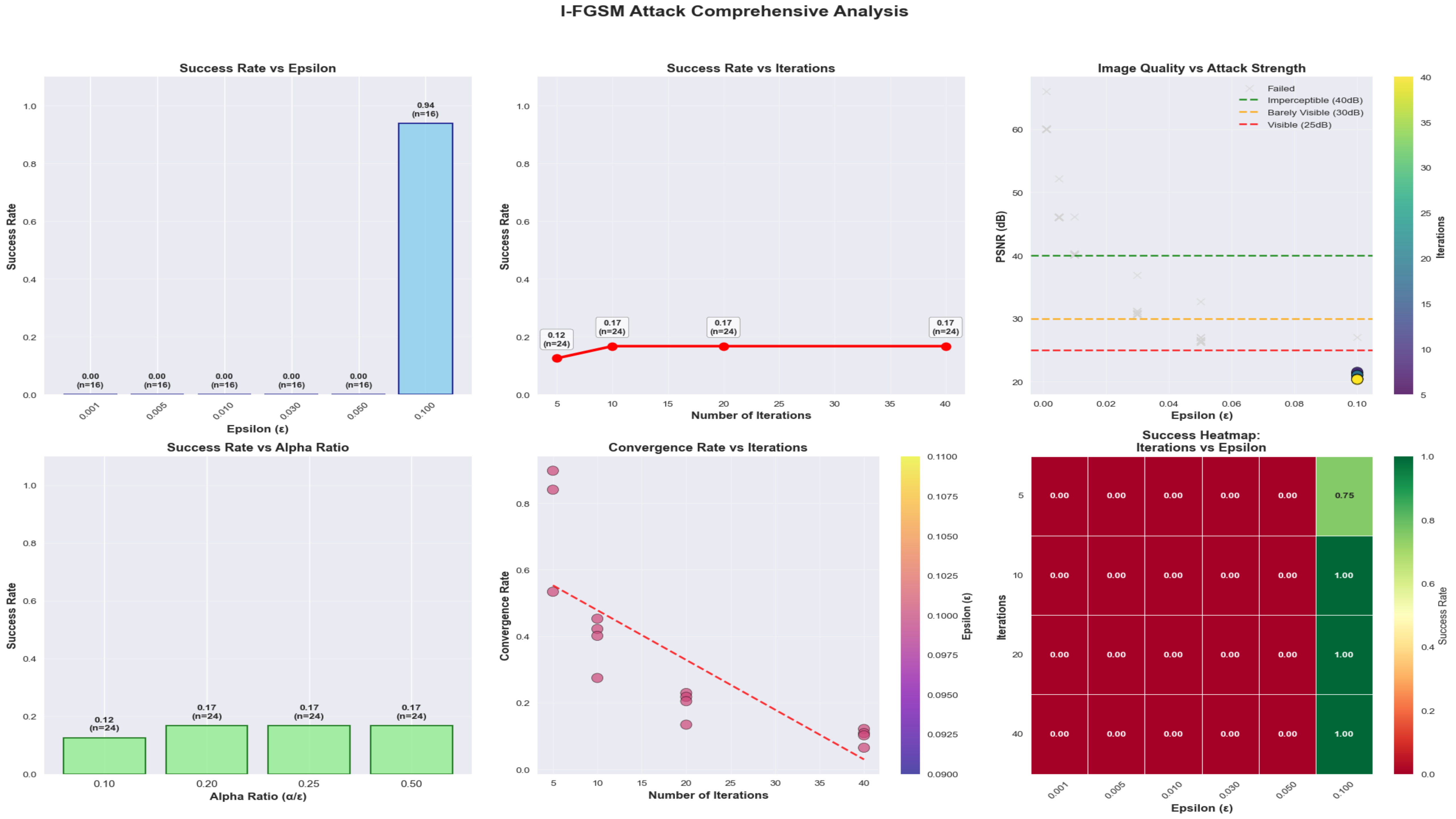
Michael Anggi G.A.

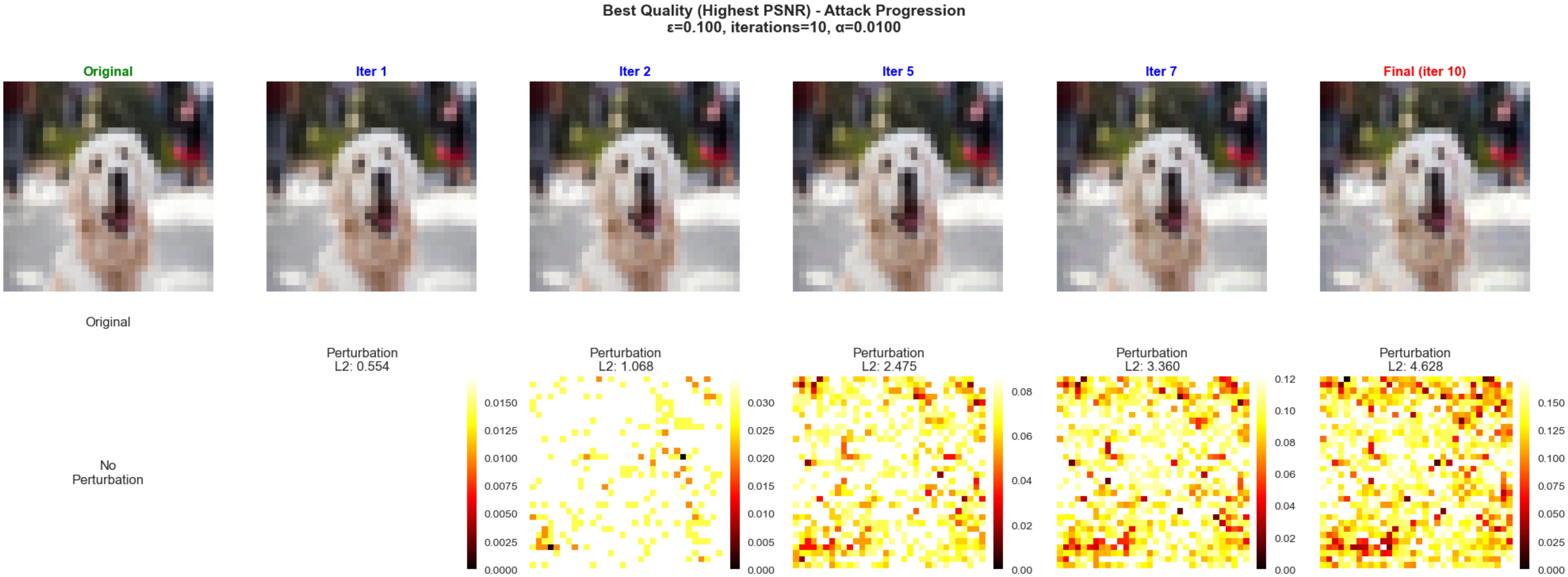
I- FGSM Attack

- The convergence analysis demonstrates that the attack reaches maximum effectiveness within the first 10-15 iterations, with the convergence rate stabilizing around 0.10-0.11 and showing diminishing returns in subsequent iterations.

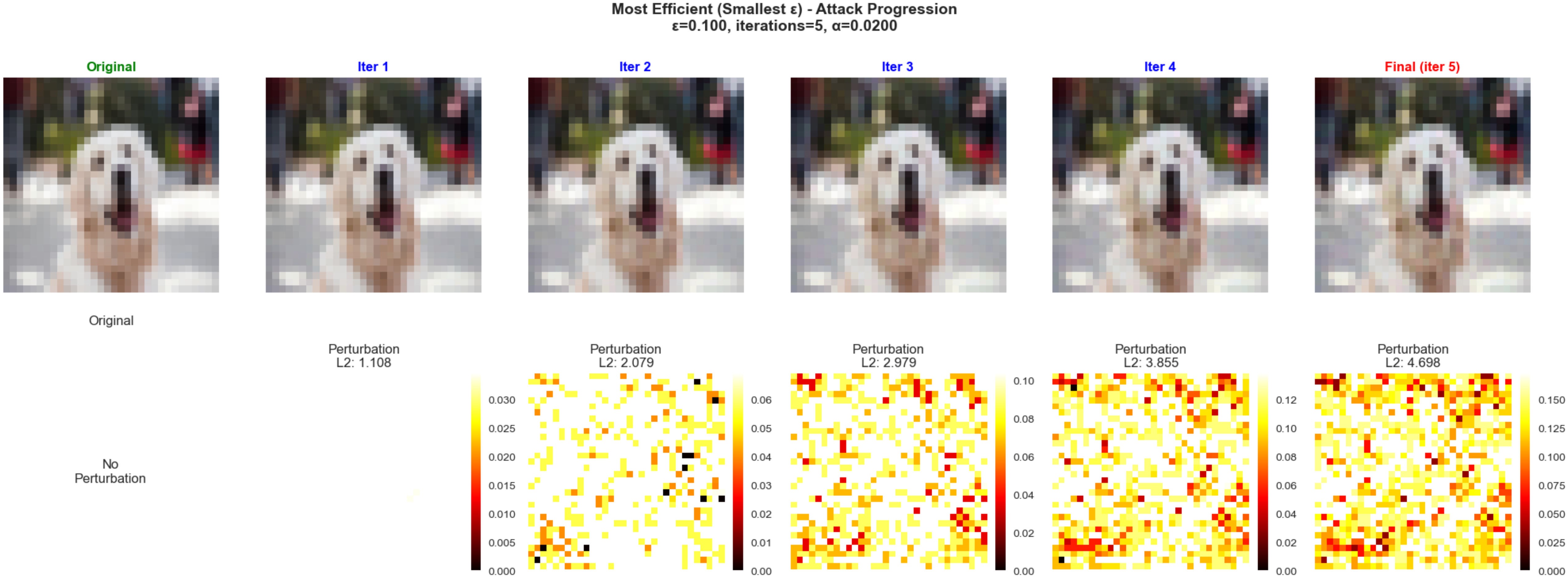
- The success heatmap clearly illustrates that I-FGSM requires a minimum epsilon threshold ($\varepsilon \geq 0.08$) to achieve meaningful attack success, with complete failure at lower perturbation budgets ($\varepsilon < 0.08$) regardless of iteration count.

- Image quality metrics show that successful attacks maintain reasonable PSNR values around 20-25 dB, balancing attack effectiveness with visual imperceptibility constraints.

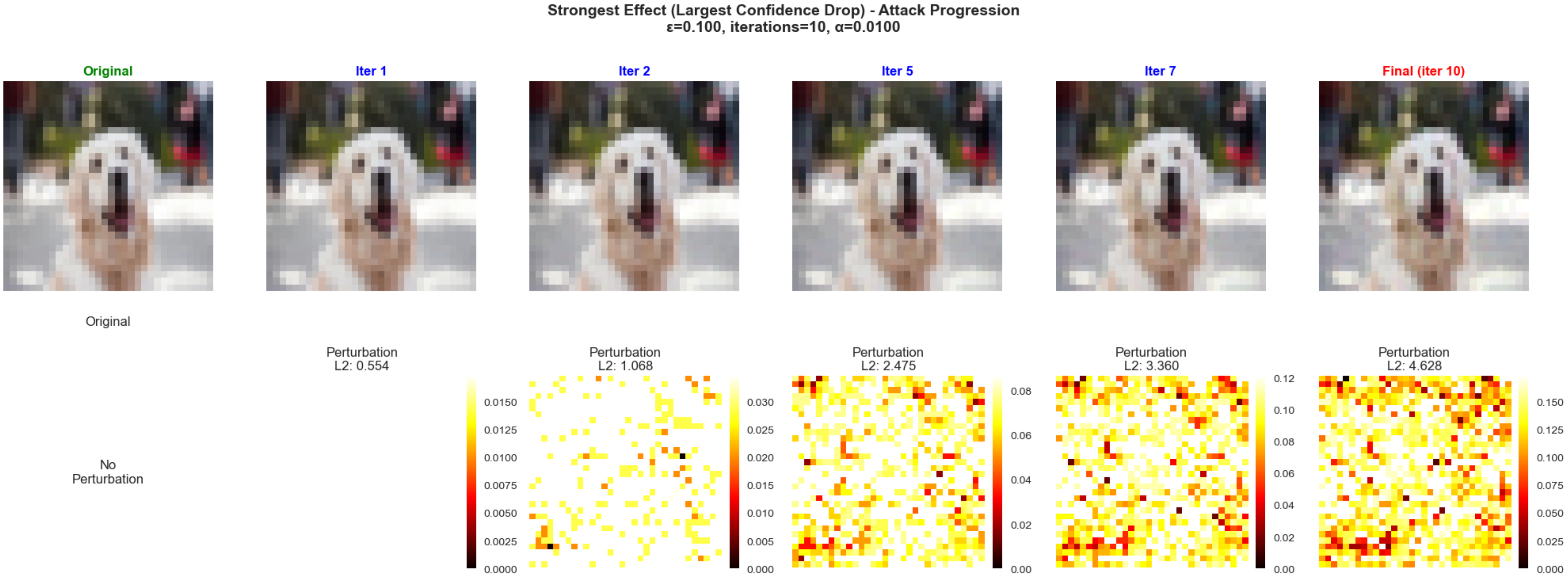

I-FGSM Attack Comprehensive Analysis
Success Rate vs Epsilon
Success Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.00 (n=16)
0.00 (n=16)
0.00 (n=16)
0.00 (n=16)
0.00 (n=16)
0.94 (n=16)
0.001
0.005
0.010
0.030
0.050
0.100
Epsilon (ε)
Success Rate vs Iterations
Success Rate
0.12 (n=24)
0.17 (n=24)
0.17 (n=24)
0.17 (n=24)
5
10
15
20
25
30
35
40
Number of Iterations
Image Quality vs Attack Strength
PSNR (dB)
Failed
Imperceptible (40dB)
Barely Visible (30dB)
Visible (25dB)
20
30
40
50
60
0.00
0.02
0.04
0.06
0.08
0.10
Epsilon (ε)
Iterations
5
10
15
20
25
30
35
40
Success Rate vs Alpha Ratio
Success Rate
0.12 (n=24)
0.17 (n=24)
0.17 (n=24)
0.17 (n=24)
0.10
0.20
0.25
0.50
Alpha Ratio (α/ε)
Convergence Rate vs Iterations
Convergence Rate
0.0
0.2
0.4
0.6
0.8
5
10
15
20
25
30
35
40
Number of Iterations
Epsilon (ε)
0.0900
0.0925
0.0950
0.0975
0.1000
0.1025
0.1050
0.1075
0.1100
Success Heatmap: Iterations vs Epsilon
Iterations
5: 0.00 0.00 0.00 0.00 0.00 0.75
10: 0.00 0.00 0.00 0.00 0.00 1.00
20: 0.00 0.00 0.00 0.00 0.00 1.00
40: 0.00 0.00 0.00 0.00 0.00 1.00
0.001
0.005
0.010
0.030
0.050
0.100
Epsilon (ε)
Success Rate
0.0
0.2
0.4
0.6
0.8
1.0



This visualization demonstrates the iterative progression of the I-FGSM attack (ε=0.100, 10 iterations, α=0.0100) showing how perturbations gradually accumulate from L2=0.554 in iteration 1 to L2=4.628 in the final iteration, with the perturbation heatmaps revealing increasingly concentrated adversarial noise while the adversarial images remain visually similar to the original throughout the attack process.



This visualization shows the "Most Efficient" I-FGSM attack progression (ε=0.100, 5 iterations, α=0.0200) where perturbations accumulate more rapidly from L2=1.108 in iteration 1 to L2=4.698 in the final iteration, demonstrating how larger step sizes (α=0.0200) achieve similar perturbation magnitudes in fewer iterations compared to the previous configuration, while the perturbation heatmaps show concentrated adversarial noise developing across the image.



Strongest Effect (Largest Confidence Drop) - Attack Progression
ε=0.100, iterations=10, α=0.0100

This visualization depicts the "Strongest Effect" I-FGSM attack progression (ε=0.100, 10 iterations, α=0.0100) that achieves the largest confidence drop, showing the systematic accumulation of perturbations from L2=0.554 in iteration 1 to L2=4.628 in the final iteration, with the perturbation heatmaps revealing how the attack strategically targets specific regions of the image to maximize the model's misclassification confidence.

I-FGSM Attack Analysis: Strengthening Adversarial Examples via Iterative Refinement

Michael Anggi G.A.

FGSM Attack Summary

- The attack success rate exhibits a binary behavior pattern, remaining completely ineffective at ε<0.08 but achieving significant success rates once the minimum perturbation budget is met, highlighting the discrete nature of adversarial vulnerability.

- Alpha ratio analysis reveals robustness to hyperparameter variations, with consistent 17% success rates across different step sizes (0.10-0.50), indicating that I-FGSM performance is relatively stable across reasonable parameter ranges.

- The progressive perturbation accumulation through iterations allows for controlled trade-offs between attack strength and image quality, with PSNR values maintaining reasonable perceptual bounds around 20-25 dB.

I-FGSM Attack Analysis: Strengthening Adversarial Examples via Iterative Refinement

Michael Anggi G.A.

Conclusion

- The Iterative Fast Gradient Sign Method (I-FGSM) demonstrates superior attack effectiveness compared to single-step FGSM through its iterative refinement approach, achieving 94% success rate at ε=0.10 while maintaining reasonable image quality with PSNR values around 20-25 dB.

- I-FGSM exhibits a critical epsilon threshold behavior where attacks completely fail below ε=0.08 but achieve significant success rates once this threshold is exceeded, revealing the discrete nature of neural network vulnerability and the importance of perturbation budget selection in adversarial attack strategies.

- The iterative approach converges rapidly within 10-15 iterations and shows robustness to hyperparameter variations across different alpha ratios (0.10-0.50), making I-FGSM a practical and reliable method for generating adversarial examples that balance attack effectiveness with computational efficiency and visual imperceptibility.