Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30289

Comparison of Projected Gradient Descent Attack Effects on ResNet18 and VGG16 Models

Xueting Zeng

School of Computer Engineering and Science, Shanghai University, Shanghai, China
zxt2941@outlook.com

Abstract. In the field of AI image classification, traditional research has a clear focus. It mainly studies model classification accuracy. But it does not pay enough attention to the model's robustness to outside changes. In real use, outside changes can have a big effect. Even small ones can make the model give wrong results. These mistakes cause risks. They affect the reliability and safety of the system. Residual Network (ResNet)-18 and Visual Geometry Group Network (VGG)-16 are classic models. They are used in deep learning image classification. They are often used in many tasks. Projected Gradient Descent (PGD) is an adversarial attack method. It works in steps and is very effective. This paper uses PGD attacks as a test. In the cat-dog classification task, ResNet18 and VGG16 models are trained on a standard dataset. Tests are done on both normal samples and adversarial ones made by PGD, studying how stable both models are when facing outside changes. This research provides theoretical and practical evidence for improving the security and reliability of image classification models in practical applications.

Keywords: Image classification, Projected gradient descent (PGD), Adversarial attack, Model robustness

1. Introduction

In traditional research in the field of image classification in artificial intelligence, researchers often focus on the accuracy of model classification and are committed to pursuing higher classification and recognition accuracy by improving training strategies, optimizing network structures, etc., but to a certain extent ignore the model's anti-interference ability. However, in practical application scenarios, image classification tasks often face many uncontrollable external interference factors. Although these factors are difficult for humans to detect, they may cause the model to make completely wrong predictions, seriously threatening the reliability and safety of related systems. In safety-critical fields such as medical imaging diagnosis and autonomous driving image recognition, such misjudgments caused by adversarial disturbances can have serious consequences, such as misclassifying malignant tumor MRI images as benign or misidentifying traffic lights as other objects [1].

Since the breakthrough of AlexNet, deep learning has promoted the development of image classification. Residual Network (ResNet)-18 and Visual Geometry Group Network (VGG)-16, as two representative deep learning image classification models, have been widely used in

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ30289

classification tasks. VGG16 has achieved excellent results in image classification tasks with its concise and regular network architecture, which fully extracts local features of images by stacking multiple convolutional layers and pooling layers [2].

ResNet18, with its unique residual connection structure, effectively solves the gradient vanishing and gradient exploding problems in the deep network training process, and can reduce the training difficulty and computational cost of the model while ensuring the performance of the model [3]. As a powerful iterative adversarial attack method, Projected Gradient Descent (PGD) has the characteristics of good attack effect and strong versatility, and can effectively attack a variety of deep learning models [4].

In recent years, some new research works have combined the concept of enhanced learning with model robustness, and carried out attack and defense experiments on classic models such as VGG16 and ResNet [5]. These results show that the introduction of adversarial mechanisms can not only improve the generalization ability of the model, but also further highlight the important value of adversarial experiments in revealing the vulnerabilities of different network structures. However, there is very limited research on adversarial attacks on the VGG16 model trained on the cat and dog binary classification dataset [6].

With this in mind, this paper uses Projected Gradient Descent (PGD) attacks, a powerful iterative adversarial attack method, as a benchmark. This paper select the representative and practical use case of cat and dog classification, and examine two classic deep learning models, ResNet18 and VGG16. After training the models using classic datasets, this paper conduct performance tests on both original samples and adversarial examples generated by PGD attacks to examine the stability of the two models' classification performance in the face of external perturbations. Ultimately, this study provides theoretical and practical evidence for improving the security and reliability of image classification models in real-world applications.

2. Methods

This experiment utilized the following methods, models, and parameters.

2.1. PGD

PGD is a classic white-box iterative adversarial attack method. Its core principle is to calculate the gradient of the model's loss function with respect to the input sample through multiple rounds of iterations, apply small perturbations along the gradient (with a step size of α), and constrain the sample to a preset perturbation range (ε neighborhood) through a projection operation after each iteration, ensuring that the perturbation is imperceptible and consistent with data rationality. It consists of 16 convolutional layers and 3 fully connected layers. It is a pretrained model whose parameters can be directly loaded and used.

2.2. VGG16 and ResNet18

VGGNet and ResNet are both classic architectures in the field of convolutional neural networks. They both improve the capabilities of the model by increasing the depth of the network.

VGG16's core design concept focuses on enhancing feature extraction capabilities through network deepening and the stacking of small convolution kernels with the network comprising 16 layers total 13 convolutional layers and 3 fully connected layers its structure remains straightforward as all convolutional layers adopt 3×3 small convolution kernels stacking multiple small convolutions

DOI: 10.54254/2755-2721/2026.TJ30289

instead of using large ones reduces parameter counts while preserving the receptive field and introducing additional nonlinearity to strengthen the model's fitting performance.

Too many network layers can cause the vanishing gradient problem. This problem makes it hard to update the parameters in deep layers and causes the model training to stop improving. ResNet's main advantage is its residual connection design. It solves the vanishing gradient problem in deep networks and makes it possible to train deeper models without losing performance.

ResNet18 has 18 layers. These include 17 convolutional layers and 1 fully connected layer. Its overall structure consists of three parts. First is an initial convolution and pooling layer. Second are four stages composed of stacked residual blocks. Each residual block contains two 3×3 convolutions. Third is the output part: it outputs classification results through global average pooling and a fully connected layer.

ResNet18 adopts a lightweight deep design. This design ensures strong feature extraction capabilities. Meanwhile, it maintains low computational complexity. Thus, ResNet18 is widely used in image classification and transfer learning.

2.3. Evaluation criteria

First, the accuracy of the two models after the attack will be tested, and the attack success rate will be obtained.

$$Attacksuccessrate = accuracybeforeattack - accuracyafterattac \tag{1}$$

Then, this experiment will test the L2 and SSIM values. L2 is the perturbation amplitude, which measures the pixel difference between the original image and the adversarial sample. The calculation formula is:

$$L2(x, x_{adv}) = \sqrt{\sum_{i=1}^{H\times W\times C}(x_i - x_{adv,i})^2} \tag{2}$$

Where H, W, C is the height, width, and number of channels of the image, x_i is the original image pixel value, and $x_{adv,i}$ is the adversarial sample pixel value.

The SSIM value measures image similarity from three dimensions: brightness, contrast, and structure. The value range is [0,1]. The calculation formula is:

$$SSIM(x, x_{adv}) = \frac{(2\mu_x\mu_{adv}+C_1)(2\sigma_{x,adv}+C_2)}{(\mu_x^2+\mu_{adv}^2+C_1)(\sigma_x^2+\sigma_{adv}^2+C_2)} \tag{3}$$

Among them, μ_x and μ_{adv} are means, σ_x^2 and σ_{adv}^2 are variances, $\sigma_{x,adv}$ is the covariance, and C_1 and C_2 are stability constants. The closer the SSIM value is to 1, the smaller the visual difference is.

3. Experiment

3.1. Overall experimental process

First, the original cat and dog dataset was preprocessed to provide a data foundation for subsequent model training and testing. Second, the VGG16 and ResNet18 models were trained using the preprocessed dataset and tested on original images to obtain baseline performance. Next, the PGD attack method was used to generate adversarial examples, and the VGG16 and ResNet18 models

DOI: 10.54254/2755-2721/2026.TJ30289

were tested on these adversarial examples to evaluate their robustness under adversarial attacks. Finally, a comparative analysis of the original image test results and the adversarial example test results was performed to compare the effectiveness of the PGD attack on the VGG16 and ResNet18 models.

3.2. Experimental dataset and experimental configuration

3.2.1. Experimental dataset

The dataset used in this experiment is the Kaggle Cats vs Dogs dataset. It is a large dataset for classifying cat and dog images. The training set makes up 80%, and the test set makes up 20%. Each part has 50% cat images and 50% dog images.

The data preprocessing in this experiment had three steps. The first step was resizing. All images were changed to the same size of 224x224 pixels. The second step was converting all images into PyTorch tensors and normalizing the pixel values to the range [0, 1]. The third step was standardizing the images using the dataset's mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225].

3.2.2. Experimental configuration

During training, ResNet18 uses the Stochastic Gradient Descent (SGD) optimizer, while VGG16 uses the Adam optimizer. The training parameters for the VGG16 and ResNet18 models also differ to accommodate the convergence characteristics of each model. See the Table 1 for specific configurations.

Table 1. Specific configurations

Parameters	VGG16	ResNet18
training epochs	10	20
initial learning rate	0.001	0.0001
learning rate schedule	5 rounds × 0.1 (attenuation factor)	5 rounds × 0.1 (attenuation factor)
batch size	32	32

The loss function used is cross-entropy loss, and the training process includes model training and validation in each round. The final evaluation metrics include training loss, validation loss, and accuracy, and a visualization is generated.

When using PGD attack, the test is divided into multiple groups, and the parameters are as Table 2.

Table 2. Attack parameters

Step size	0.001
Number of iterations	10
Perturbation intensity	0.003,0.01,0.0314,0.05,0.07

This paper only generate adversarial samples for samples that were originally predicted correctly, and save four types of image files: original image, adversarial sample, perturbed image, and

DOI: 10.54254/2755-2721/2026.TJ30289

comparison image. The final evaluation metrics include accuracy, attack success rate, L2, and SSIM value.

3.3. Experimental results and analysis

3.3.1. Training results

Table 3. Initial accuracy

Models	Accuracy
VGG16	98.20%
ResNet18	97.40%

Table 3 shows the initial accuracy of the two models during training. It can be observed that both models have high initial accuracy, indicating that both models have been fully trained and have good feature extraction and discrimination capabilities. The accuracy of the VGG16 model is slightly higher than that of the ResNet18 model.

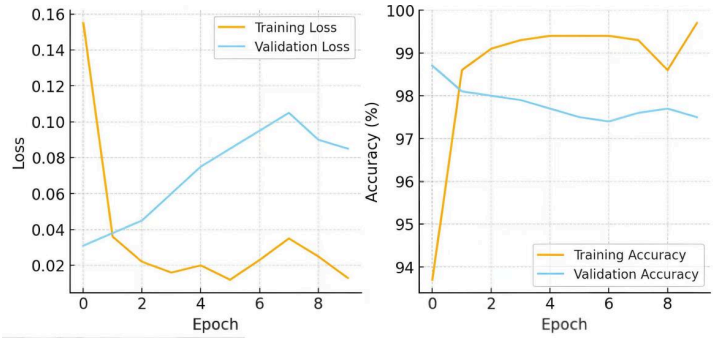


Figure 1. VGG16's training and validation loss and training and validation accuracy. (picture credit: original)

Figure 1 illustrates the training and validation loss, as well as the training and validation accuracy, of the VGG16 model. In this experiment, the training loss rapidly decreased from an initial high value, then fluctuated slightly within a low loss range. This indicates that VGG16's ability to fit the training set continued to improve with each epoch, stabilizing in the later stages, and the model's error on the training data gradually converged. Its validation loss initially increased, then decreased somewhat later, but remained higher than the training loss overall, indicating a certain degree of overfitting.

Furthermore, VGG16's training accuracy rapidly increased and remained near 100%, indicating strong performance on the training set. Meanwhile, its validation accuracy, after initially increasing, gradually declined, ultimately falling somewhat short of the training accuracy. Combined with the overfitting characteristics of the loss curve, and can infer that VGG16's classification performance on the validation set is slightly inferior to that on the training set due to overfitting to the training set, but remains at a high level.

DOI: 10.54254/2755-2721/2026.TJ30289

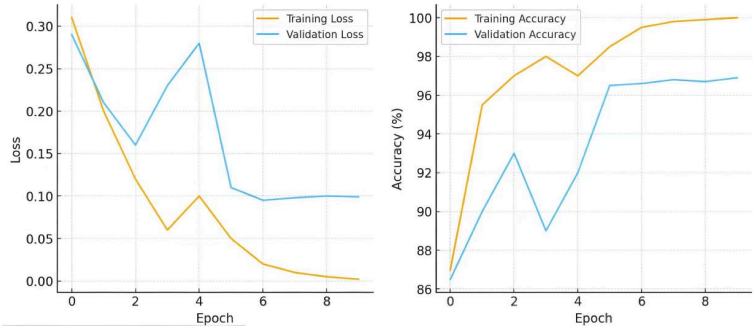


Figure 2. ResNet18's training and validation loss and training and validation accuracy. (picture credit: original)

Figure 2 shows the ResNet18's Training and Validation Loss and Training and Validation Accuracy. For the ResNet18 model in this experiment, its training loss gradually decreases from a high value and approaches 0 in the later stage, indicating that ResNet18 fits the training set very well and can fully learn the characteristics of cats and dogs in the training data.

In addition, the validation loss of ResNet18 fluctuates greatly in the early stage, but gradually decreases and stabilizes in the later stage, and the gap with the training loss becomes smaller.

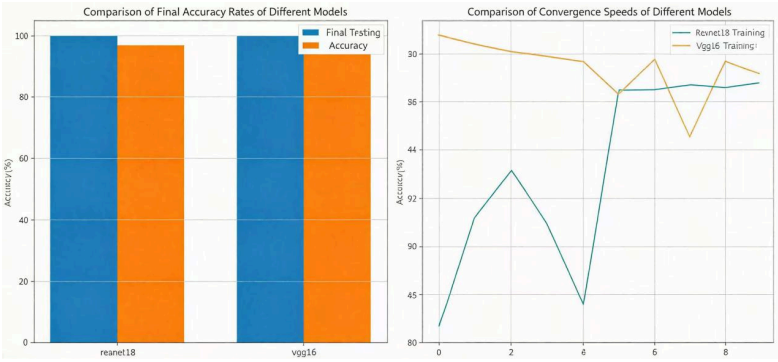


Figure 3. The initial accuracy and convergence speed of the two models. (picture credit: original)

Figure 3 compares the initial accuracy and convergence speed of the two models. ResNet18 and VGG16 both achieve very high final training accuracy, approaching 100%. This result shows that both models can effectively learn training data.

ResNet18's final validation accuracy is slightly lower than its training accuracy, but it is still very high. VGG16's final validation accuracy is lower than that of ResNet18. This finding further proves that ResNet-18 has better generalization ability. It also has a smaller gap between training and validation accuracy.

In terms of convergence speed, VGG16 has a relatively high starting point for its validation accuracy curve. But the curve fluctuates and becomes unstable afterward. ResNet18's validation accuracy curve starts at a low level. It rises rapidly, though with some fluctuations in the middle. Finally, it stabilizes at a high level.

This phenomenon indicates two points. First, ResNet18 converges faster during training. Second, VGG16 has a more erratic convergence process.

DOI: 10.54254/2755-2721/2026.TJ30289

3.3.2. Example attack results

The following is an example of the attack effect, showing the original image test results with a perturbation amplitude of 0.05, the test results of the attacked image, and the attack mask (Figure 4).



(a) original, result = cats (b) original, result = cats (c) epilson = 0.05

Figure 4. The original image test results with a perturbation amplitude of 0.05, the test results of the attacked image, and the attack mask. (picture credit: original)

3.3.3. PGD attack effect indicators

Table 4. Accuracy after attack and attack success rate

Perturbation intensity	VGG16 accuracy	ResNet18 accuracy	VGG16 attack success rate	ResNet18 attack success rate
0.003	36.25%	35.75%	61.95%	61.65%
0.01	33.10%	32.58%	65.10%	64.82%
0.0314	31.08%	30.45%	67.12%	66.95%
0.05	29.88%	29.48%	68.32%	67.92%
0.07	27.62%	27.52%	70.58%	69.88%

It can be observed (Table 4) that after introducing the PGD attack, the performance of both models decreased significantly. As the perturbation strength ε gradually increased, the model classification accuracy showed a monotonically decreasing trend, with a significant decrease of nearly 70 percentage points. Meanwhile, the attack success rate kept rising as ε increased. This result shows that the PGD attack worked effectively in this experiment. It undermined the model's discrimination boundary. Samples that would have been correctly classified were misclassified under minor perturbations. This reflects the model's high sensitivity to adversarial perturbations.

The experiment also compared the attack resistance of the two models. With the same perturbation strength, VGG16 gets higher accuracy than Residual Network ResNet18. Its attack success rate is a little lower than ResNet18's. The performance difference between the two models is within 1%. This may be because of the structural differences between the models. VGG16 uses more convolutional layers to refine features step by step. ResNet18 uses residual connections to reduce the vanishing gradient problem in deep networks. The introduction of residual units ensures more stable gradient propagation in deep networks. This helps counteract the cumulative amplification effect of perturbation propagation.

In contrast, VGG16's purely stacked convolutional layer structure is more likely to be driven into local minima when subjected to progressively improved perturbations, resulting in greater impact on its prediction signal. Overall, different feature learning mechanisms lead to subtle differences in their sensitivity to PGD perturbations, but in the simple case of binary classification, this structural difference has not yet significantly impacted robustness.

DOI: 10.54254/2755-2721/2026.TJ30289

This result is consistent with the conclusions of the analysis of model adversarial robustness in other studies. For example, some researchers pointed out that the VGG series models are more sensitive to adversarial perturbations than the ResNet series [7]; and in the Badnets attack, models with different network structures also have different resistance to attacks, which is closely related to the feature extraction mechanism of the model [8]. The results of this experiment are also highly consistent with the conclusions of Engstrom et al. in the robustness review. They pointed out that the residual connection structure of the ResNet series can alleviate the gradient vanishing problem in deep networks, so that the model can maintain the stability of feature extraction during the perturbation propagation process, while the stacked convolutional layers of the VGG series are prone to perturbation accumulation, thereby exacerbating the prediction error. It can be seen that the gradient transfer characteristics of the network are one of the core factors affecting its adversarial robustness [9].

3.3.4. Adversarial sample characteristics

Table 5. The values of L2 and SSIM

Disturbance intensity	The L2 value of VGG16	The L2 value of VGG16	The SSIM value of VGG16	The SSIM value of ResNet18
0.003	0.85	0.78	0.9990	0.9992
0.01	3.20	3.05	0.9905	0.9955
0.0314	10.80	10.20	0.9632	0.9624
0.05	17.50	16.90	0.8804	0.8846
0.07	24.60	23.90	0.7792	0.7910

To further analyze the characteristics of the model's adversarial examples, this experiment measured the L2 and the SSIM at different perturbation intensities. The results are shown in the Table5. It can be observed that as ε increases, the perturbation amplitude of both VGG16 and ResNet18 shows a linear upward trend, indicating that the input perturbation energy continues to increase. Simultaneously, the SSIM value gradually decreases, indicating that the perturbation gradually damages the image's visual structure. However, even at the maximum perturbation of ε = 0.07, the SSIM value remains above 0.78, indicating that the generated adversarial examples still have a high visual similarity to the original image, and the attack is relatively "stealthy." Comparing the two models, it is found that the perturbation amplitude of ResNet18 is slightly smaller at the same ε, while the SSIM value is slightly higher. This further confirms its relatively stronger stability and structure preservation capabilities. Specifically, it requires less perturbation and less visual distortion to achieve the same attack target.

In summary, from the comprehensive analysis of the four dimensions of classification accuracy, attack success rate, perturbation amplitude and SSIM, PGD attack can significantly reduce model performance [10], but under the same attack intensity, ResNet18 always shows higher robustness and perceptual fidelity than VGG16. The residual structure increases the model’s ability to handle adversarial changes. It helps the model respond smoothly to input changes and keeps the classification performance stable.

DOI: 10.54254/2755-2721/2026.TJ30289

4. Conclusion

This study looks at the robustness of deep learning image classification models against adversarial attacks. It focuses on two convolutional neural networks, VGG16 and ResNet18. The PGD method was used to carry out attack experiments. The evaluation included classification accuracy, attack success rate, perturbation size, and SSIM value. Results show that PGD attacks reduce model performance a lot. When perturbation size goes up, model accuracy goes down and attack success rate goes up. This shows that deep neural networks are vulnerable to adversarial noise. Between the two models, ResNet18 is more robust because of its residual connections. VGG16 is good at feature extraction but is more sensitive to changes. The study has some limits. It uses a binary classification task with simple data and only uses the PGD method without other attacks or defenses. Future work should test models on more complex multi-class datasets, use multiple attack methods for comparison, and add adversarial training, regularization, and design improvements to strengthen defense.

References

[1] Anonymous. (2024, July 25). A multi-layered defense against adversarial attacks in brain tumor classification using ensemble adversarial training and feature squeezing. PubMed Central.
[2] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556.
[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778).
[4] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations.
[5] Singh, A., Singh, M. P., & Singh, A. K. (2025). Adversarially enhanced learning (AEL): Robust lightweight deep learning approach for radiology image classification against adversarial attacks. Image and Vision Computing, 154, 105405. https: //doi.org/10.1016/j.imavis.2024.105405
[6] Wang, I. H., Lee, K. C., & Chang, S. L. (2020). Proceedings of the IEEE Eurasia Conference on IoT, Communication and Engineering (ECICE) (pp. 231–232). IEEE.
[7] Shen, B. (2024). The investigation of robustness differences for variously-sized models in adversarial attacks and defense. In Proceedings of the International Conference on Signal Processing and Machine Learning.
[8] Gu, T., Liu, K., Dolan-Gavitt, B., & Garg, S. (2019). Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 7, 47230–47244.
[9] Carlini, N., & Wagner, D. (2017). Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (pp. 3–14).
[10] Zhang, Y., Li, Y., Li, Y., & Guo, Z. (2023). A review of adversarial attacks in computer vision. arXiv preprint arXiv: 2308.07673.