# Adversarial attack study on VGG16 for cat and dog image classification task

**Chenrui Cui**

School of Software Engineering, Tongji University, Shanghai, 201804, China

2152614@tongji.edu.cn

**Abstract.** Contemporarily, adversarial attacks on deep learning models have garnered significant attention. With this in mind, this study delves into the effectiveness of adversarial attacks specifically targeted at the VGG16 model in the context of cat and dog image classification. Employing the Fast Gradient Sign Method (FGSM) for attack, the experimental findings reveal that, within a certain perturbation range, FGSM attacks can indeed reduce the model's average confidence, albeit with relatively minor impacts on accuracy. According to the analysis, the accuracy drops (decreased from 88.5% to 88.2%) is not significant, possibly due to limited classes. With small ε, perturbation results in a notable confidence drop. However, at higher ε, perturbation impact lessens, averaging around 50% confidence for cat and dog classes, indicating a 2-class scenario's upper limit in non-targeted FGSM attacks. Additionally, this research underscores the need for further exploration into various adversarial attack methods and model interpretability within the realm of image classification. Overall, these results shed light on guiding further exploration of adversarial attack defense strategies, holding significant potential for real-world applications in enhancing the robustness of AI systems against adversarial attacks.

**Keywords:** VGG16, adversarial attacks, image classification, deep learning, model robustness.

## 1. Introduction

Image classification, a fundamental task in computer vision, has evolved significantly over time. Early efforts relied on handcrafted features and traditional machine learning. The breakthrough came with deep learning and convolutional neural networks (CNNs), e. g., VGG16, which revolutionized accuracy. However, the impact of adversarial attacks on image classification tasks is well-recognized. Adversarial attacks involve making subtle modifications to input data with the aim of deceiving machine learning models or deep neural networks, causing them to produce incorrect outputs even when the input appears normal. Image classification stands as a pivotal task in the field of computer vision, with the emergence of deep neural network models like VGG16 leading to remarkable advancements in classification accuracy. VGG16's popularity is attributed to its relatively simple yet powerful architecture, which excels in various image classification tasks.

However, VGG16 models are susceptible to adversarial attacks in certain scenarios, such as the recognition of traffic signs in autonomous driving applications [1] and the classification of tobacco diseases and pests [2]. While current research predominantly focuses on the general properties of adversarial attacks, encompassing evaluations across different datasets and models [3], it is important to note that the effectiveness of adversarial attacks varies across different tasks and models. Moreover, the assessment of adversarial effects on the CIFAR-10 dataset, which is commonly used for testing attack effectiveness, may differ from real-world application scenarios.

Currently, there is limited research on adversarial attacks specifically targeting VGG16 models trained on the "Cat vs. Dog" dataset [4]. Therefore, this paper aims to investigate the performance of VGG16 models trained for cat and dog classification under adversarial attacks, particularly those generated using the Fast Gradient Sign Method (FGSM). This research not only contributes to a deeper understanding of the vulnerabilities of VGG16 when facing adversarial attacks but also holds significance in enhancing the security and reliability of image classification models and addressing potential adversarial threats.

## 2. Method

### 2.1. VGG16 network

As shown in Fig. 1, VGG16 is a straightforward and deep convolutional neural network architecture. It consists of 16 convolutional layers and 3 fully connected layers [5]. VGG16 is a pretrained model with parameters that can be loaded directly.

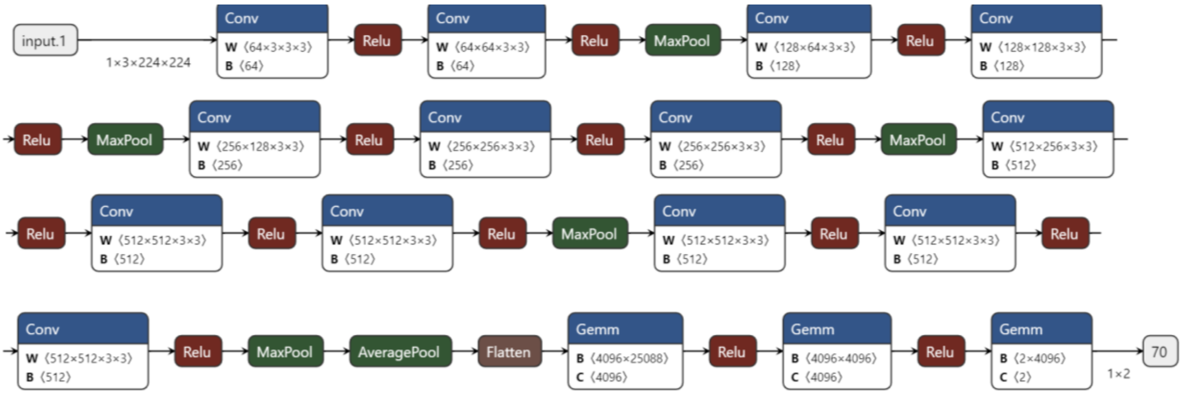

**Figure 1.** The structure of VGG16. Each layer is arranged sequentially from left to right and from top to bottom (Photo/Picture credit: Original).

### 2.2. FGSM

White-box attacks require knowledge of the target model's architecture, parameters, and the image to be attacked. FGSM is one form of white-box attack [6]. The FGSM method generates adversarial perturbations by computing the gradients of the target model [7]. Loss function is denoted as $Loss$, the initial input as $x$,perturbation as $\delta$, and parameters as $\theta$, the neural network is represented as $F(x;\theta)$. An adversarial example is denoted as $x+\delta$. The classifier is represented as $C(x)$. After training, the classifier ensures that the input $x$ yields the correct classification target [8], which can be represented as $C(x)=y$. Non-targeted attacks aim to cause the adversarial example to be classified incorrectly, which can be represented as $C(x+\delta)\neq y$. The objective during training is $min_{\theta}\ Loss(F(x;\theta),y)$, which means seeking a set of parameters "θ" that minimize the loss function. As for Fast Gradient Sign Method (FGSM), the attack objective is to $max_{\delta}Loss(F(x+\delta;\theta),y)$ under the condition $\|\delta\|_{\infty}\leq\varepsilon$, meaning to find a $\delta$ within a perturbation range $\varepsilon$ that maximizes the loss function [9]. FGSM needs to solve the equation: $x+\delta = x+\varepsilon\cdot sign(\nabla_{x}Loss(F(x;\theta),y))$, which requires computing gradients and selecting an appropriate $\varepsilon$ value [10]. However, for certain complex models and tasks, FGSM may require larger $\varepsilon$ values for successful attacks, potentially making the adversarial samples visually conspicuous and compromising attack stealthies.

### 2.3. *Attack method*

Before training the model, it is necessary to preprocess the images to conform to the input requirements of the VGG16 network. To conduct attacks on the VGG16 model, first, performed five rounds of iterative training with randomly shuffled training data locally, resulting in a VGG16 model tailored for the cat-dog classification task on the "Cat vs. Dog" dataset. Subsequently, a non-targeted Fast Gradient Sign Method (FGSM) attack was applied to a specific image to evaluate the attack's effectiveness. Finally, the test dataset underwent FGSM attacks to assess changes in confidence scores and determine the overall attack impact. The attack process is shown in Fig. 2.

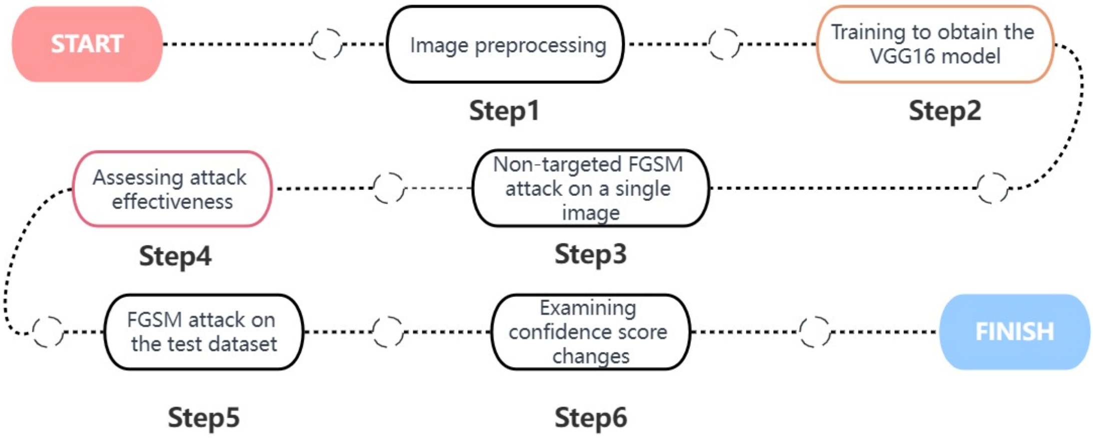
START
Image preprocessing
Step1
Training to obtain the VGG16 model
Step2
Assessing attack effectiveness
Step4
Non-targeted FGSM attack on a single image
Step3
FGSM attack on the test dataset
Step5
Examining confidence score changes
Step6
FINISH

**Figure 2.** The attack process (Photo/Picture credit: Original).

## 3. Results and discussion

The experimental dataset employed the Dogs vs. Cats dataset [11], which comprises thousands of images of cats and dogs gathered from the internet, constituting a binary classification task. During classifier training, a total of 2,000 images were utilized, evenly split between 1,000 cat images and 1,000 dog images. Due to the VGG16 network's requirement for input images of size 224x224, image preprocessing was necessary. In each training iteration, out of the 2,000 images, 80% were used for training, and the remaining 20% for validation. A total of 5 training iterations were conducted, with the image order randomized in each iteration. In the binary classification task of cat vs. dog images, the VGG16 classifier produces a confidence score indicating whether an input image is a cat or a dog. In the experiments, the class with the higher confidence score was chosen as the classification result. For instance, if the confidence score for a given image indicated 94% for cat and 6% for dog, the image is classified as a cat. The test dataset consisted of 500 images different from the training and validation sets, and accuracy and confidence scores were observed after testing. During the 5 iterations of training, the model's loss and accuracy on the validation set changed as shown in Fig. 3 and Fig. 4. The third iteration of training achieved the highest accuracy on the validation set (88.50%) with the lowest total loss (118.45). After the third iteration, both accuracy and loss had essentially converged. Therefore, the model from the third iteration of training was selected for testing. On the test dataset, the model achieved an accuracy of 85.80%. The average confidence score for classifying cats was 82.28%, and for classifying dogs, it was 83.76%. When a specific image (as shown in Fig. 5) is used as input, the classifier outputs the following probabilities: the probability of the image being a cat is 93.22%, and the probability of it being a dog is 6.78%. A set of perturbation (ε) were tested, including 1%, 5%, 10%, 20%, 30%, and 40%. The results of FGSM attacks on this image at different perturbation sizes are as follows. Due to the requirement of VGG16 to input images of size 224x224, the images underwent resizing and colour domain distortion during data preprocessing, resulting in some distortion of the images. To validate the effectiveness of the attacks across the entire dataset, the attacks were tested on the entire test dataset. The changes in accuracy and the average confidence scores for both cats and dogs under different perturbations are given in Table 1.

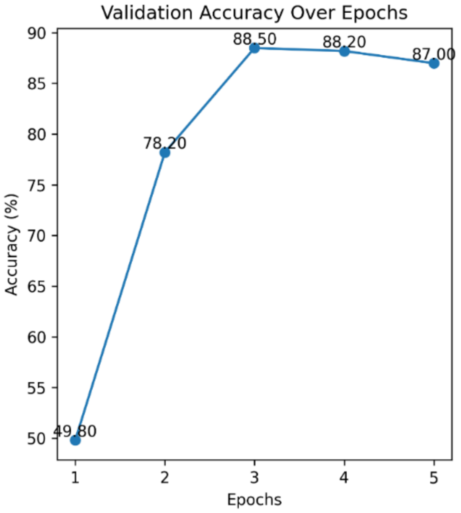

**Figure 3.** Validation Accuracy Over Epochs (Photo/Picture credit: Original).

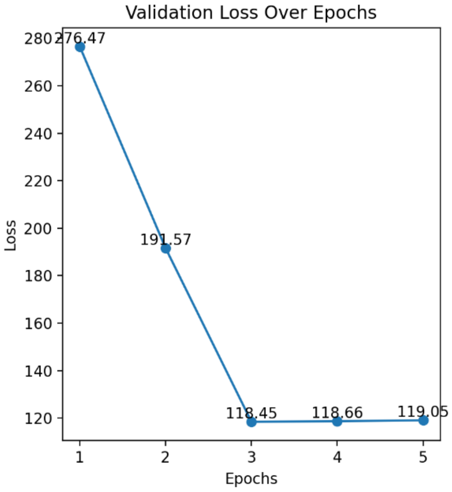

**Figure 4.** Validation Loss Over Epochs (Photo/Picture credit: Original).

**Table 1.** Changes in Accuracy Due to Attacks. "Perturbation Size" is $\varepsilon$, where the first row with Perturbation Size of 0 represents the state before the attack.

| Perturbation Size | Accuracy (%) |
|---|---|
| 0 | 88.50 |
| 0.01 | 88.40 |
| 0.05 | 88.40 |
| 0.1 | 88.40 |
| 0.2 | 88.40 |
| 0.3 | 88.40 |
| 0.4 | 88.20 |

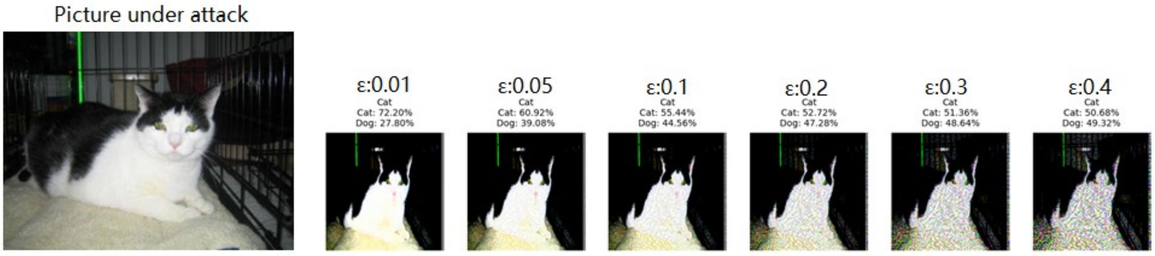

**Figure 5.** Attack results on a single image (Photo/Picture credit: Original).

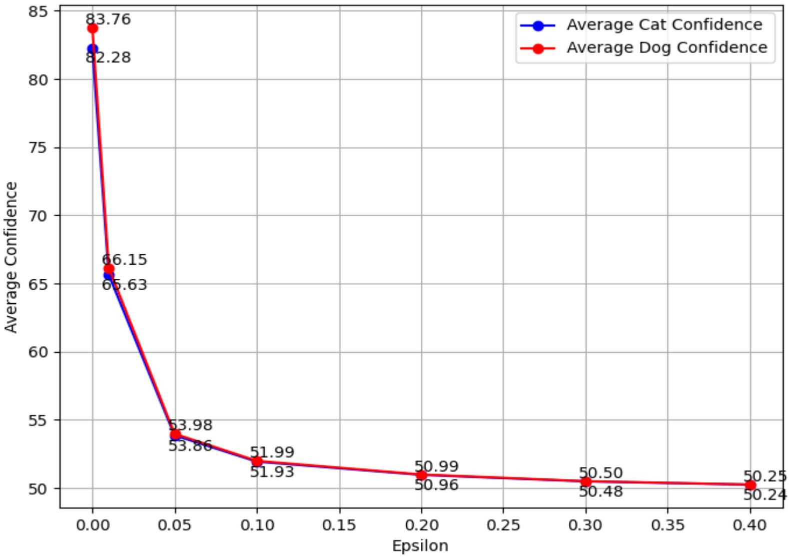

**Figure 6.** Average confidence with $\varepsilon$, which shows the reduction in confidence on the test set. “$\varepsilon$” is denoted as “Epsilon” at the horizontal axis in the diagram (Photo/Picture credit: Original).

**Table 2.** Attack-Induced Changes in Average Confidence. "Perturbation Size" is $\varepsilon$. The first row with Perturbation Size of 0 represents the state before the attack.

| Perturbation Size | Average Confidence in Identifying as a Dog (%) | Average Confidence in Identifying as a Cat (%) |
|---|---|---|
| 0 | 83.76 | 82.28 |
| 0.01 | 66.15 | 65.63 |
| 0.05 | 53.98 | 53.86 |
| 0.1 | 51.99 | 51.93 |
| 0.2 | 50.99 | 50.96 |
| 0.3 | 50.50 | 50.48 |
| 0.4 | 50.25 | 50.24 |

According to Table 1, the decrease in accuracy is not significant, which may be due to the small number of classes in the classification. However, in Table 2, from the perspective of confidence, FGSM attacks result in a decrease in the model's average confidence. When the perturbation ($\boldsymbol{\varepsilon}$) is small, the reduction in perturbation leads to a significant decrease in the classification model's confidence. However, when the perturbation is high, the reduction in perturbation has little impact on the classifier's average confidence, and at this point, it is also more noticeable that applying adversarial samples leads to changes in the images (seen from Fig. 6). In non-targeted attack tasks, as the perturbation increases, the curves of the average confidence for both cat and dog classes converge to around 50%. This indicates that in a cat vs. dog binary classification scenario, FGSM non-targeted attacks can at most reduce the classifier's average confidence to around 50%, which corresponds to 100% divided by the number of classes.

## 4. Conclusion

To sum up, the primary focus of this study was to investigate the impact of FGSM and other adversarial attacks on the VGG16 model in the context of cat and dog image classification. The experiments confirmed that in this scenario, FGSM attacks exhibit significant effectiveness, especially when the perturbation is small, as even minor changes in the perturbation can significantly reduce the model's average confidence. The limitations of this study are evident in the relatively small number of classes, as it only explored the classification of cat and dog images. It is possible that this limited class set contributed to the almost non-existent misclassification due to attacks. In future research, additional categories can be added from other datasets, such as pigs, cows, sheep, and more. Further study will explore more adversarial attack methods like BIM and PGD, or combine them with relevant knowledge of model interpretability to investigate whether targeting specific pixels or locations can achieve better attack results. The study provides valuable insights into the application of adversarial attacks in image classification and offers useful guidance for future research exploring a wider range of categories and attack methods.

## References

[1] Ye B, Yin H, Yan J and Ge W 2021 IEEE International Intelligent Transportation Systems Conference (ITSC) p 15.
[2] Swasono D I, Tjandrasa H and Fathicah C 2019 12th international conference on information & communication technology and system (ICTS) pp. 176-181.
[3] Dong Y, Fu Q A, Yang X, et al. 2020 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp 321-331.
[4] Wang I H, Lee K C and Chang S L 2020 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE) pp 230-233.
[5] Simonyan K and Zisserman A 2014 arxiv preprint arxiv:1409.1556.
[6] Li F, Wang C, Chen L, et al. 2020 Journal of Computer Research and Development vol 57(10) pp 2066.
[7] Kurakin A, Goodfellow I J and Bengio S 2018 Artificial intelligence safety and security pp 99-112.
[8] Jmour N, Zayen S and Abdelkrim A 2018 international conference on advanced systems and electric technologies (IC_ASET) pp 397-402.
[9] Chakraborty A, Alam M, Dey V, Chattopadhyay A and Mukhopadhyay D 2018 arXiv preprint arXiv:1810.00069.
[10] Goodfellow I J, Shlens J and Szegedy C 2014 arXiv preprint arXiv:1412.6572.
[11] Parkhi O M, Vedaldi A, Zisserman A and Jawahar C V 2012 IEEE conference on computer vision and pattern recognition pp 3498-3505.