Multi-target Adversarial Attack on Image Classification Deep Neural Networks

Xiang Siqi 1004875   Kishen 1005885   Luah Shi Hui 1005512   Liu Yu 1005621

1 INTRODUCTION

Deep learning models, especially those applied in the domain of image classification, have demonstrated extraordinary capabilities and have been applied to various fields, including healthcare and autonomous driving. However, the robustness of these models is challenged by adversarial attacks. These attacks involve intentionally perturbing inputs to mislead models into making incorrect predictions, raising significant concerns about their deployment in safety-critical applications.

Traditional adversarial attack methodologies on image classification tasks have primarily focused on single-target prediction tasks, where the aim is to deceive the model into misclassifying an image as an incorrect label. While effective at exploiting vulnerabilities in deep learning models, this traditional approach does not fully capture the complexity of real-world applications, where decisions are neither binary nor singular. In contrast, multi-target classification tasks, prevalent in sectors such as medical imaging and multi-class object detection, require the model to discern among multiple correct categories, adding more complexity to the classification challenge.

To this end, we introduce the Multi-Targeted Iterative Fast Gradient Sign Method (MT-IFGSM), an innovative adversarial attack methodology designed specifically for multi-targeted image classification tasks. Our experiments demonstrate that our methodology outperforms the traditional Iterative Fast Gradient Sign Method (ITFGSM) on multi-targeted image classification tasks in effectiveness and stealthiness. Additionally, the adaptability of our methodology across different deep neural network architectures showcases its broad applicability and potential to serve as a benchmark for evaluating the robustness of multi-target classification models against adversarial attacks.

Our code is available at: https://github.com/TsukiSky/multi-targeted-itfgsm-on-image-classification.

2 METHODOLOGY

As existing adversarial attacks for image classification do not fully address the complexities and unique challenges in multi-targeted classification tasks, we propose MT-IFGSM (**M**ulti-**T**argeted **I**terative **F**ast **G**radient **S**ign **M**ethod), an untargeted adversarial attack that is designed for multi-targeted classifications. This section first gives an overview of this attack. Subsequent sections explain the problem definition and outline the implementation of MT-IFGSM.

2.1 Overview

Figure 1 illustrates the overview of the proposed method. By identifying the vulnerable classes that a given model and sample pair is most likely to misclassify into, MT-IFGSM introduces subtle perturbations to images that deceive models into misclassifying the sample into those vulnerable classes. By this means, fewer perturbations can be added to the sample to mislead the model into making wrong classification decisions.

2.2 Problem Definition

In multi-targeted classification tasks, models are required to identify among multiple possible correct labels for each input sample. This presents a unique set of challenges for adversarial attacks, as the goal shifts from inducing a single incorrect classification to achieving misclassification across several specific target labels. The problem is compounded by the need for adversarial examples to remain stealthy to human observers, necessitating a careful balance between effectiveness and stealthiness. MT-IFGSM addresses these challenges by introducing a method that systematically identifies the most vulnerable targets within the multi-class task and crafts perturbations that are both effective in deceiving the model and subtle enough to evade both human-eye and defense algorithms.

2.3 Implementation

The Multi-Targeted Iterative Fast Gradient Sign Method (MT-IFGSM) involves several steps, aiming to generate adversarial examples that are tailored to exploit the vulnerabilities of models in multi-target classification tasks. This section describes each step in MT-IFGSM.

2.3.1 Input Samples and Victim Models

As a general untargeted attack approach, MT-IFGSM adapts to most deep-learning models. The input sample is typically an image that is unseen by the victim model. MT-IFGSM takes both and generates adversarial samples.

2.3.2 Identifying Vulnerable Labels

To obtain a baseline prediction, MT-IFGSM first feeds through to get the model's confidence in classifying into each available category. It therefore identifies those classes that exhibit relatively larger differences in model confidence to the actual class and considers them as candidates for the attack. These candidates represent areas where the model's decision boundary may be more easily perturbed in adversarial attacks.

2.3.3 Adversarial Targeted Perturbation

With the vulnerable labels identified, MT-IFGSM proceeds to generate perturbed samples. This phase is designed to exploit the weaknesses revealed in the previous step, manipulating the input in a way that maximizes the likelihood of misclassification from the original classes toward the vulnerable labels identified. Similar to the traditional ITFGSM, MT-IFGSM employs an iterative refinement process to ensure the perturbation remains undetectable to human observers while still effectively misleading. In each iteration, a small, calculated perturbation is applied to the input image, gradually nudging its output towards the desired misclassification.

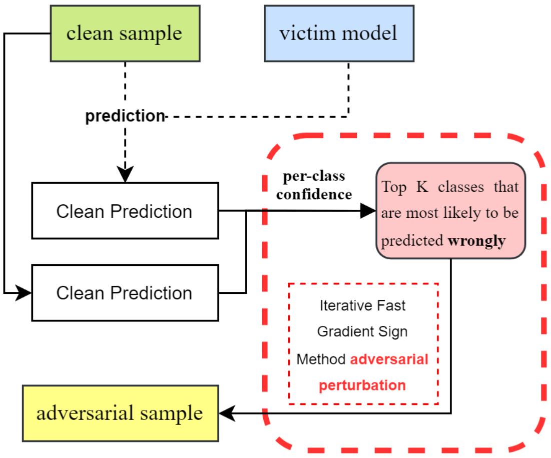
Compared to ITFGSM, the MT-IFGSM method perturbs the original image towards the nearest incorrect label rather than drastically altering its classification. This targeted approach efficiently reduces the extent of perturbation required for misclassification, thereby preserving the original appearance of the image to the greatest extent.

3 EXPERIMENT

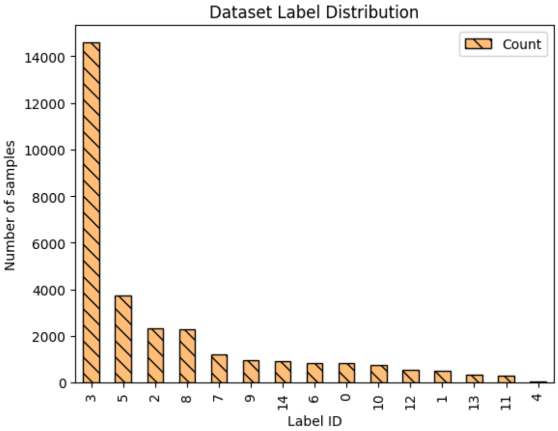
This section presents the experimental setup and results of applying the Multi-Targeted Iterative Fast Gradient Sign Method (MT-IFGSM) on a subset of the NIH Chest X-rays dataset [1]. We also implemented three deep-learning models to evaluate the performance of MT-IFGSM.

3.1 Dataset

The NIH Chest X-ray dataset is comprised of chest X-ray images labelled with diseases identified to train the victim models. Due to time and computation power limitations, we randomly sampled 25000 images from the NIH Chest X-rays dataset to use in our experiment. The class distribution of the new dataset is shown in Figure 2.


clean sample
victim model
prediction
Clean Prediction
Clean Prediction
per-class confidence
Top K classes that are most likely to be predicted **wrongly**
Iterative Fast Gradient Sign Method adversarial perturbation
adversarial sample

[Figure 1]. Process of MT-IFGSM

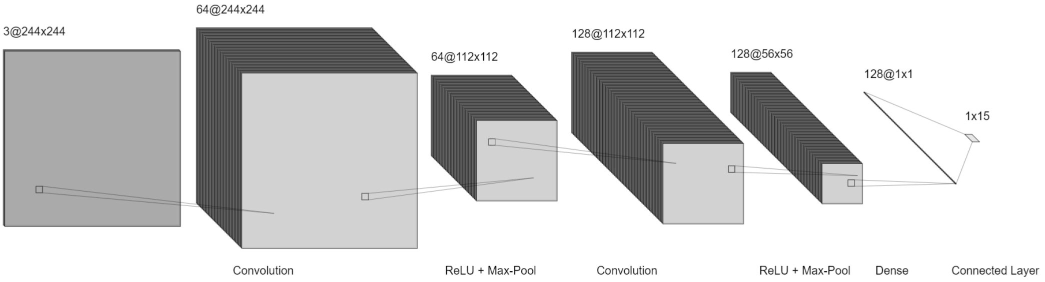


[Figure 2]. Dataset label distribution

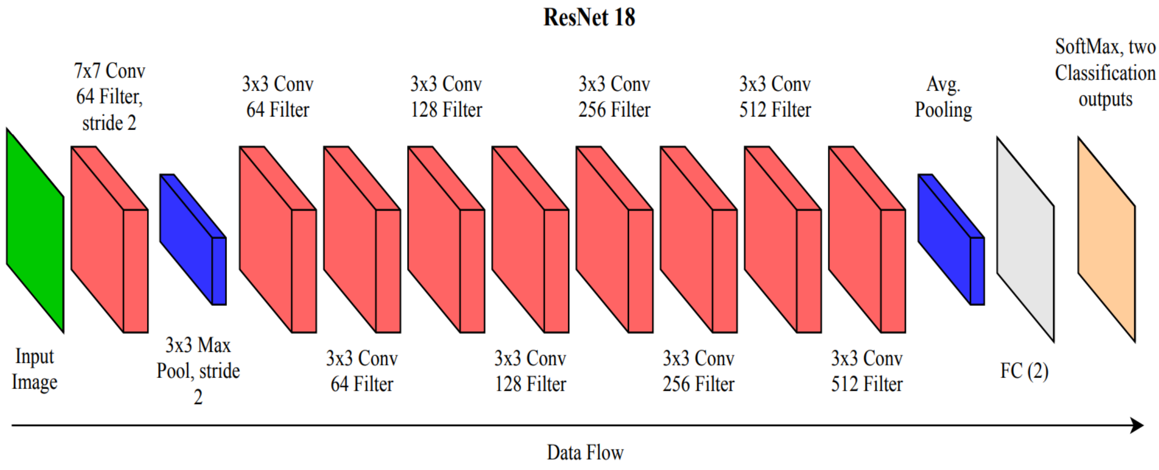
3.2 Victim Models

To evaluate the performance of the MT-IFGSM attack, we trained three victim models. Firstly, we manually implemented a 2-layer CNN model and a simple Vision Transformer (ViT) model. Other than that, we used the ResNet model provided by the torchvision API. Given the multi-target nature of our classification task, we appended a sigmoid activation layer to the output of each model. This layer transforms the raw output scores into probabilities between 0 and 1. A threshold of 0.5 is used to determine the presence of each class, where probabilities greater than or equal to 0.5 are interpreted as an indication of the presence of the associated condition. Figure 3 to Figure 5 illustrate the structures of these models.

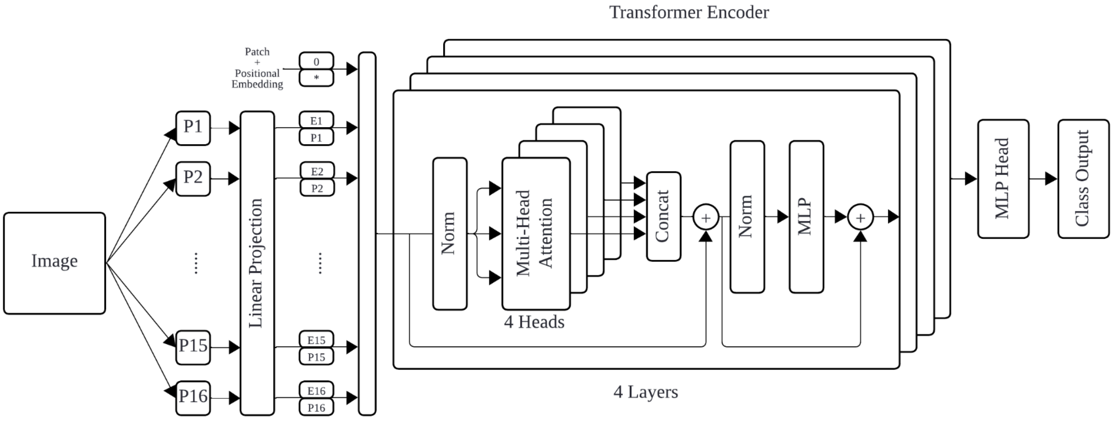
- **2-layer CNN.** A straightforward Convolutional Neural Network with two convolutional layers followed by a fully connected layer. This model is a baseline for evaluating the attacking performance of MT-IFGSM.
- **ResNet.** A complex CNN-architecture model with residual connections [2, 3]. We directly load the architecture of this model provided by the *torchvision.models* API.
- **Simple ViT.** Our implementation of a simplified Vision Transformer model. In our experiment, we adopted a patch size of 16 and designed an encoder consisting of 4 layers with 4 attention heads.



[Figure 3]. 2-layer CNN



[Figure 4]. ResNet18



[Figure 5]. Simple ViT

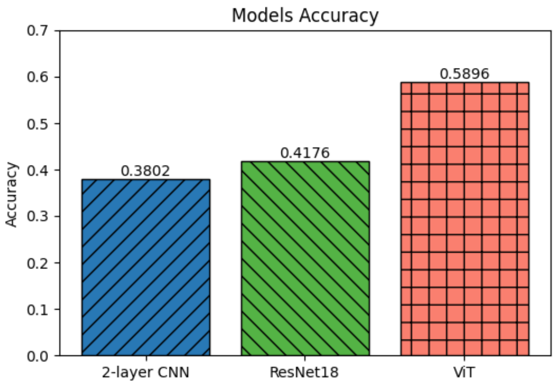
3.3 Evaluation

In this section, we evaluate the attack performance of MT-IFGSM by applying it to the victim models. The dataset was divided into training and testing subsets at a ratio of 4:1. Therefore, our testing dataset comprises 5000 samples, each annotated with one or multiple disease labels.

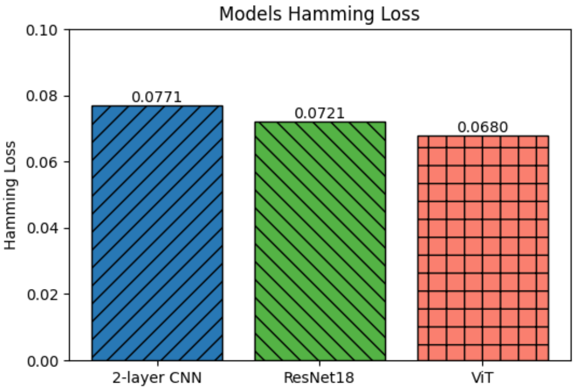
3.3.1 Attack Effectiveness

Evaluation Metrics. we evaluate the performance using accuracies, and hamming losses. The hamming loss, in particular, measures the fraction of the wrong labels to the total number of labels.

Results. Figure 6 and Figure 7 illustrate the performance of the model without adversarial attacks. The test dataset consists of 5000 samples, each annotated with one or more classes. A model's prediction is considered correct if it accurately identifies all the classes associated with a sample.



[Figure 6]. Model Accuracy



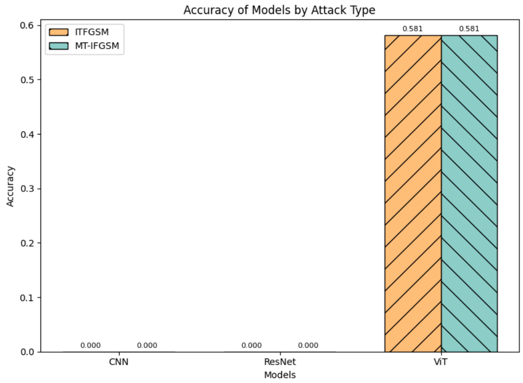
[Figure 7]. Model Hamming Loss

The 2-layer CNN achieved an accuracy of 38.02%, indicative of its basic architecture's capabilities. The ResNet18, with its deeper and more complex structure, secured a higher accuracy of 41.76%. The Vision Transformer (ViT) model outperformed both, achieving the highest accuracy of 58.96%, which can be attributed to its attention-based mechanisms that might be better suited for the intricacies of multi-label chest X-ray image classification.

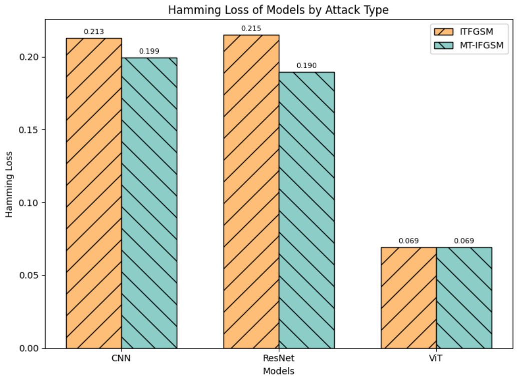
Regarding the Hamming loss, the 2-layer CNN obtained a Hamming loss of 0.0771, suggesting that it had the highest rate of individual label errors among the three models. The ResNet model showed a slight improvement with a Hamming Loss of 0.0721, reflecting its ability to reduce label misclassification errors compared to the simpler CNN. The Vision Transformer (ViT) model demonstrated the lowest Hamming Loss at 0.0680, indicating its outstanding performance in accurately classifying individual labels across the multi-label dataset.

We then subjected a subset of the test dataset, consisting of 1,000 samples, to the ITFGSM and MT-IFGSM attacks to create adversarial samples. These perturbed images were subsequently inputted into the models for attack evaluation. The outcomes of the attacks are shown in Figure 8 and Figure 9.

Figure 8 shows that the accuracy of 2-layer CNN and ResNet drops to 0% when confronted with adversarial samples generated by both ITFGSM and MT-IFGSM. However, ViT is still able to maintain an accuracy of 58.1%, only marginally lower than its performance with clean samples. This demonstrates that both ITFGSM and MT-IFGSM archive high attack success rates against simple neural networks, but are not competent in attacking more resilience models that possess attention mechanisms. Figure 9 illustrates the hamming loss of models on adversarial samples. Generally, the loss obtained by ITFGSM is higher than MT-IFGSM. This is due to the targeted perturbing mechanism of MT-IFGSM, which introduces more subtle yet effective alterations.



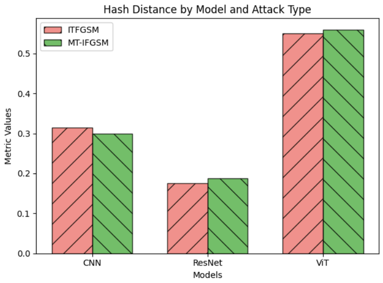
[Figure 8]. Models' Accuracy on Adversarial Samples



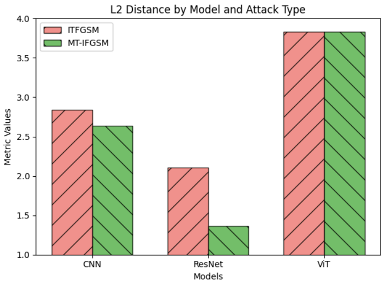
[Figure 9]. Models' Accuracy on Adversarial Samples

3.3.2 Stealthiness

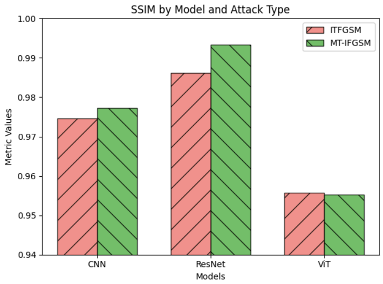
Evaluation Metrics. To assess the stealthiness of our attacks, we evaluate the perturbed images using three common metrics that quantify the subtle changes made: hash-sensitive difference, L2 distance, and Structural Similarity Index Measure (SSIM) compared to the clean samples. More stealthy attacks are identified by lower values in hash-sensitive difference and L2 distance. An SSIM value closer to 1 indicates minimal perceptible differences between the perturbed and clean images.



[Figure 10]. Hash Distance



[Figure 11]. L2 Distance



[Figure 12] SSIM

Results. The results are illustrated in Figure 10, 11, and 12. In Figure 10, the hash distance metrics show little differences between ITFGSM and MT-IFGSM attacks across all models. This indicates that both attacks manage to alter the images from the original images. Figure 11 shows that MT-IFGSM outperforms ITFGSM on CNN and ResNet models in terms of L2 distance. The slight difference in the ViT model can be attributed to its use of patch embedding. This architecture causes gradient-based attacks like ITFGSM and MT-IFGSM to perturb the image on a patch level, which is more noticeable and potentially contributes to a more significant alteration detected by the L2 distance.

In Figure 12, MT-IFGSM shows an improvement over ITFGSM, particularly on the ResNet and 2-layer CNN model, where the SSIM for MT-IFGSM is closer to the ideal value of 1.0. This suggests that MT-IFGSM can create adversarial examples that maintain a higher degree of visual similarity to the original images. This demonstrates the stealthiness of the MT-IFGSM attack in preserving image quality while still misleading the model.

5 CONCLUSION

In this project, we present the Multi-Targeted Iterative Fast Gradient Sign Method (MT-IFGSM), an adversarial attack designed to exploit vulnerabilities in deep learning models in multi-targeted image classification tasks. Our experiment, conducted on a subset of the NIH Chest X-ray dataset and tested on three distinct deep-learning architectures, demonstrated the effectiveness and stealthiness of the MT-IFGSM attack in comparison with the traditional IFTGSM attack.

The 2-layer CNN, ResNet18, and Vision Transformer (ViT) models all exhibited varying degrees of vulnerability to MT-IFGSM, with the ViT model showing the highest resilience in terms of accuracy but not impervious to attacks. Our experiments also show that the stealthiness of the adversarial examples generated by MT-IFGSM was superior compared to those produced by the traditional ITFGSM. These results emphasise the importance of rigorous security measures and the development of robust defense mechanisms in multi-targeted classification tasks, as the deployment of such models in sensitive fields demands a high degree of reliability and resistance to adversarial threats.

We hope that our findings will inspire the community to further explore the intersection of adversarial attack and multi-target classification in the near future.

6 REFERENCE

[1]. NIH Chest X-rays. (2018, February 21). Kaggle. https://www.kaggle.com/datasets/nih-chest-xrays/data

[2]. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[3]. Ghorakavi, Ram Srivatsav. "TBNet: pulmonary tuberculosis diagnosing system using deep neural networks." arXiv preprint arXiv:1902.08897 (2019).