VIETNAM NATIONAL UNIVERSITY, HANOI	**SOCIALIST REPUBLIC OF VIETNAM**
VNU UNIVERSITY OF ENGINEERING AND TECHNOLOGY	**Independence – Freedom – Happiness**

INFORMATION ON DOCTORAL THESIS

1. Full name : Nguyen Duc Anh....................... 2. Sex: Male

3. Date of birth: 19/10/1993 .... 4. Place of birth: Nam Dinh

5. Admission decision number: 1344/ QĐ-CTSV Dated 25/11/2019, VNU University of Engineering and Technology

6. Changes in academic process:

Decision number 1146 dated 19/5/2021 regarding the change of advisors.

Before change: Assoc. Prof. Pham Ngoc Hung, VNU University of Engineering and Technology

After change:

+ Main supervisor: Assoc. Prof. Pham Ngoc Hung, VNU University of Engineering and Technology

+ Co-supervisor: Prof. Nguyễn Lê Minh, Japan Advanced Institute of Science and Technology (JAIST)

7. Official thesis title: Methods for ensuring the robustness of deep neural networks

8. Major: Software Engineering ....

9. Code: 9480103.01..

10. Supervisors:

+ Main supervisor: Assoc. Prof. Pham Ngoc Hung, VNU University of Engineering and Technology

+ Co-supervisor: Prof. Nguyễn Lê Minh, Japan Advanced Institute of Science and Technology (JAIST)

11. Summary of the new findings of the thesis:

To ensure the quality of DNNs for image classification, various metrics have been proposed. However, despite careful testing of networks using these metrics, recent research has shown that networks can be easily attacked adversarially. The robustness of DNNs is the ability to accurately recognize the label of an input image in the presence of adversarial perturbation. Therefore, improving the robustness is considered one of the crucial solutions to enhance the quality of DNNs. Specifically, the thesis has achieved four main results as follows.

Firstly, the thesis proposes the HA4FNN method to improve the low success rate and performance of DeepCheck when testing feedforward neural networks. HA4FNN propose a heuristic solver and ignore the usage of maintaining neuron activation states. Experiments show that for the case of adding adversarial perturbation to a pixel, the average success rate of DeepCheck is 0.7%. In contrast, the average success rate of HA4FNN is 54.3%. Moreover, when adding adversarial perturbation to some correctly predicted images, DeepCheck may take tens of minutes, while HA4FNN only takes a maximum of a few seconds.

Secondly, the thesis proposes the PatternAttack method to enhance the diversity and quality of adversarial examples generated by Adversarial Transformation Networks (ATN). PatternAttack introduces a generalized ATN to add adversarial perturbation to input images using various patterns, and then employs a greedy algorithm to remove redundant adversarial perturbation. Experiments show that PatternAttack can generate adversarial examples satisfying multiple patterns. Especially for the pattern modifying all pixels, most attacks achieve over 99% average success rate. In terms of adversarial example quality, according to the L0 criterion, PatternAttack can remove hundreds of adversarial pixels to obtain a single adversarial pixel.

Thirdly, the thesis proposes the QI4AE method to enhance the quality of adversarial examples against all adversarial attacks. The main idea of QI4AE is to combine a greedy algorithm with an autoencoder. Experiments show that the L_0 reduction rate of adversarial perturbation can be reduced by 82% - 95% and L_2 can be reduced by 56% - 81%.

Additionally, QI4AE can improve the quality of adversarial examples with low computational cost. Experiments show that it takes only a few seconds to enhance the quality of 1,000 adversarial examples.

Finally, to enhance the robustness of DNNs, the thesis proposes the SCADefender method to remove adversarial perturbation from adversarial examples. A part of the training data of SCADefender consists of adversarial examples generated by various adversarial attack methods. Experiments show that SCADefender can achieve an average detection rate of 97.78% for MNIST, 90.43% for Fashion-MNIST, and 80.64% for CIFAR-10. Without using the autoencoder defense network, the detection rate of the testing network for this set of adversarial examples is 0%.

12. Practical applicability, if any:

The studies presented in the thesis are significant not only in theoretical aspects but also contribute to making the testing the robustness of DNNs more practical. This is particularly meaningful for DNNs with high requirements for resilience against external attacks, including adversarial attacks. Furthermore, the tools developed in the thesis have been implemented and applied at TSDV Vietnam and receiving positive feedback.

13. Further research directions, if any:

Despite the specific contributions of the research as presented above, these results still have limitations that need to be addressed. The subsequent research of the thesis aims to tackle these limitations. Specifically, the limitations and further research directions of the thesis are as follows.

In the first study, HA4FNN has two limitations. The first limitation is that HA4FNN does not support targeted adversarial attacks for feedforward neural networks. In HA4FNN, the predictor adds adversarial perturbation to the correctly predicted image to induce misclassification. One limitation of the predictor is that it does not generate adversarial perturbation for a correctly predicted image to produce an adversarial example classified as a specific target label. The second limitation is that HA4FNN does not support convolutional neural network attacks. Although PatternAttack is proposed to attack convolutional neural networks and is considered a better solution than HA4FNN, the two methods have different approaches. PatternAttack defines a target function and minimizes this function by using gradients to generate an adversarial example. In contrast, HA4FNN constructs the C program from the target network and applies program analysis techniques, symbolic execution, and the predictor to generate an adversarial example. Since convolutional networks have complex architectures, converting the network to source code is not very efficient. Additionally, using symbolic execution on source code corresponding to convolutional networks can be computationally expensive due to the potentially large number of statements. In the future, the thesis will further improve HA4FNN to address these limitations.

In the second study, although PatternAttack can generate high-quality and diverse adversarial examples, it has the following three limitations. First, PatternAttack has not been experimented with larger images such as those from ImageNet. Currently, PatternAttack has only been experimented on small images such as 28 × 28 × 1 or 28 × 28 × 3. Second, selecting a suitable autoencoder network architecture in Formula 4.1 is challenging. The autoencoder network architecture determines the network's ability to add adversarial perturbation to correctly predicted images. Third, selecting the weights among the components in Formula 4.1 affects the convergence speed of the autoencoder learning process. With appropriate weight selection, the autoencoder learning process avoids concentrating on minimizing one component and neglecting the other components. In summary, the thesis will continue researching different autoencoder network architectures, self-weight adjustment techniques, and conducting experiments on larger image datasets.

In the fourth study, although SCADefender can remove a considerable amount of adversarial perturbation from adversarial examples, it has two limitations. First, SCADefender may misclassify a correctly recognized input image. The reason is that when deployed in practice, SCADefender cannot accurately distinguish actual noisy images from adversarial ones. Instead, all images reaching the target network must go through the defensive autoencoder. Second, the impact of the autoencoder network architecture has not been fully explored in the experiments. Third, the dataset for training the defensive autoencoder can be large if it includes adversarial examples from various undirected attacks, making the learning process increasingly complex with the introduction of new adversarial attacks. Thus, generalizing the characteristics of adversarial perturbation is crucial. In conclusion, the thesis will investigate solutions to minimize the chances of image correction from correct to incorrect classifications, explore various autoencoder network architectures, and delve deeper into the nature of adversarial perturbation.

14. Thesis-related publications:

- **Duc-Anh Nguyen**, Kha Do Minh, Ngoc Nguyen Nhu, Pham Ngoc Hung (2023). SCADefender: An Autoencoder-based Defense for CNN-based Image Classifiers. In International Journal of Pattern Recognition and Artificial Intelligence (Q3 - accepted)
- **Duc-Anh Nguyen**, Kha Do Minh, Khoi Nguyen Le, Minh Nguyen Le, Pham Ngoc Hung (2022). Improving Diversity and Quality of Adversarial Examples in Adversarial Transformation Network. In Soft Computing (Q2)
- **Duc-Anh Nguyen**, Kha Do Minh, Pham Ngoc Hung, Nguyen Le Minh (2022). A Symbolic Execution-based Method to Perform Untargeted Attack on Feed-forward Neural Networks. In Automated Software Engineering (Q2)
- **Duc-Anh Nguyen**, Kha Do Minh, Duc-Anh Pham, Pham Ngoc Hung (2022). Method for Improving Quality of Adversarial Examples. In the 14th International Conference on Agents and Artificial Intelligence (ICAART - rank B)
- **Duc-Anh Nguyen**, Do Minh Kha, Pham Thi To Nga, Pham Ngoc Hung (2021). An Autoencoder-based Method for Targeted Attack on Deep Neural Network Models. In the 15th IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF - The best paper award)

Date: ……………………..

Signature: …………………

Full name: ………………

Date: ……………………..

Signature: …………………

Full name: ………………