The 38th ACM/SIGAPP Symposium On Applied Computing (Tallinn Estonia, March 27 – March 31, 2023)

# Security Verification Software Platform of Data-efficient Image Transformer Based on Fast Gradient Sign Method

In-pyo Hong, Gyu-ho Choi, Pan-Koo Kim and Chang Choi*

Presenter : In-pyo Hong (Dept. of Computer Engineering, Gachon University, Republic of Korea)

가천대학교
Gachon University

# • Introduction

- **Knowledge distillation** in AI has been actively conducted.
- Representive Image transformer model of **knowledge distillation**

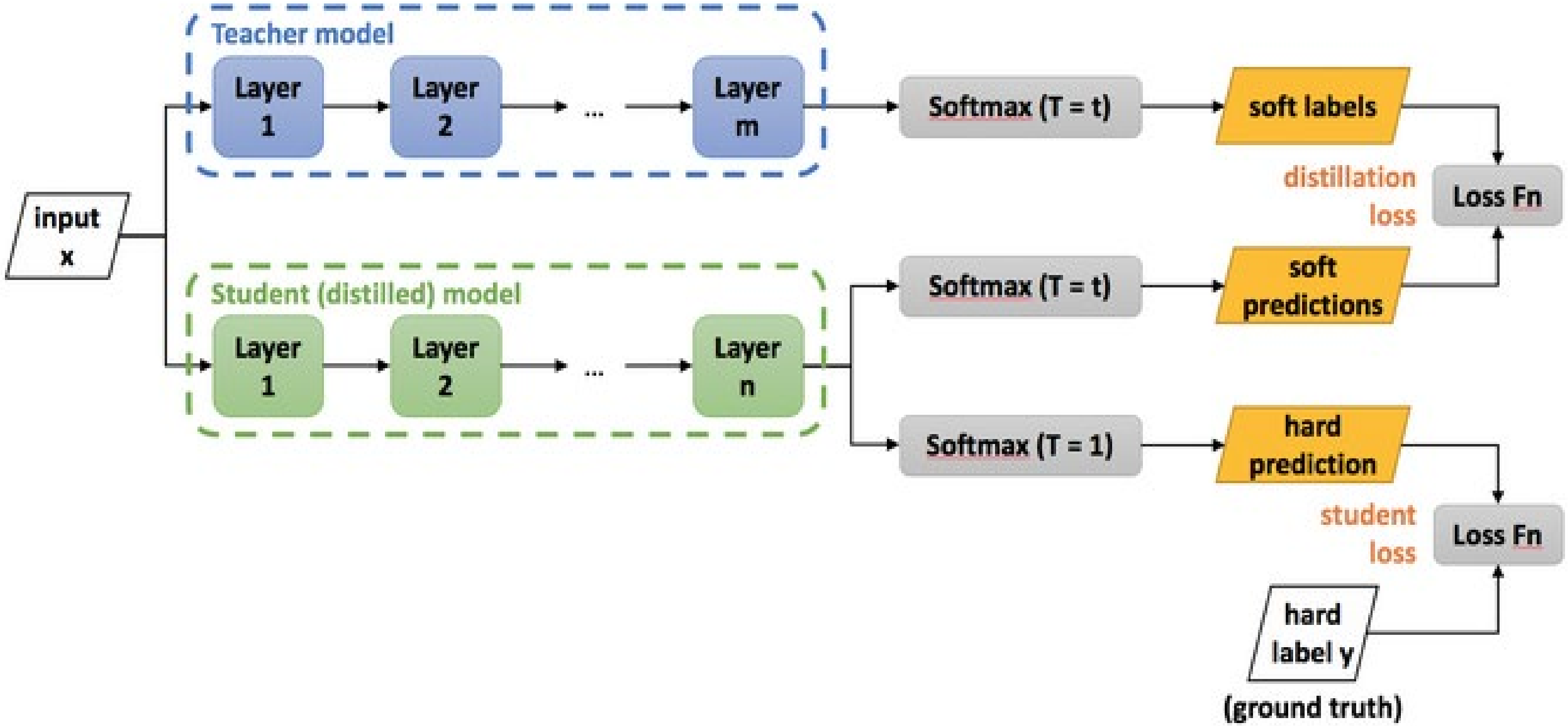
  ➔ **DeiT (Data-efficient Image Transformer)**
- DeiT has not been verified as safe against the **AI security threat**

  **(especially adversarial attack)**

Research Purpose

- *DeiT's adversarial attack vulnerability analysis*

# • Background #knowledge Distillation




<Fig1> Knowledge Distillation Structure

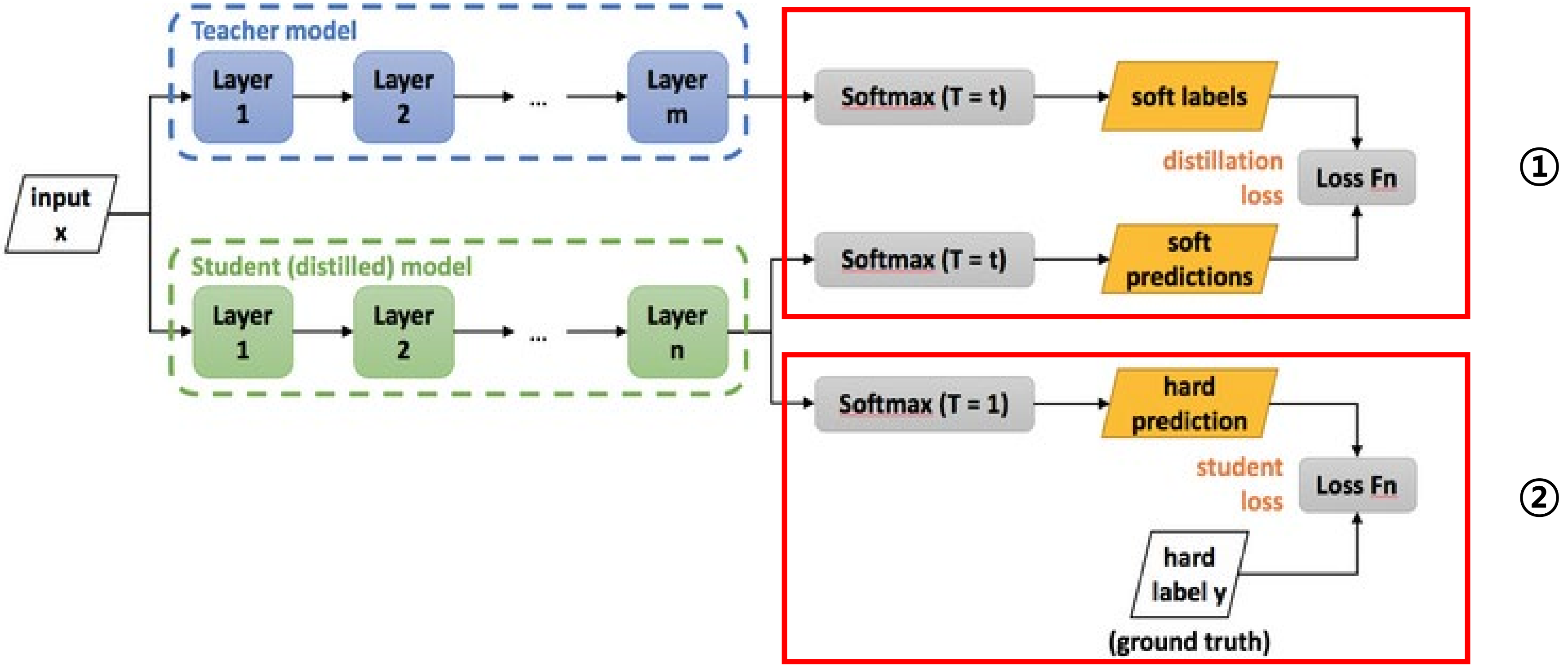
Concept of knowledge distillation : Using **2 training Models**

Purpose of knowledge distillation : Maximizing the performance of the **Student Model**



Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 2.7 (2015).

# • Background #knowledge Distillation




<Fig1> Knowledge Distillation Structure

Learning methods of the Student Model : Twice Training (① ②)

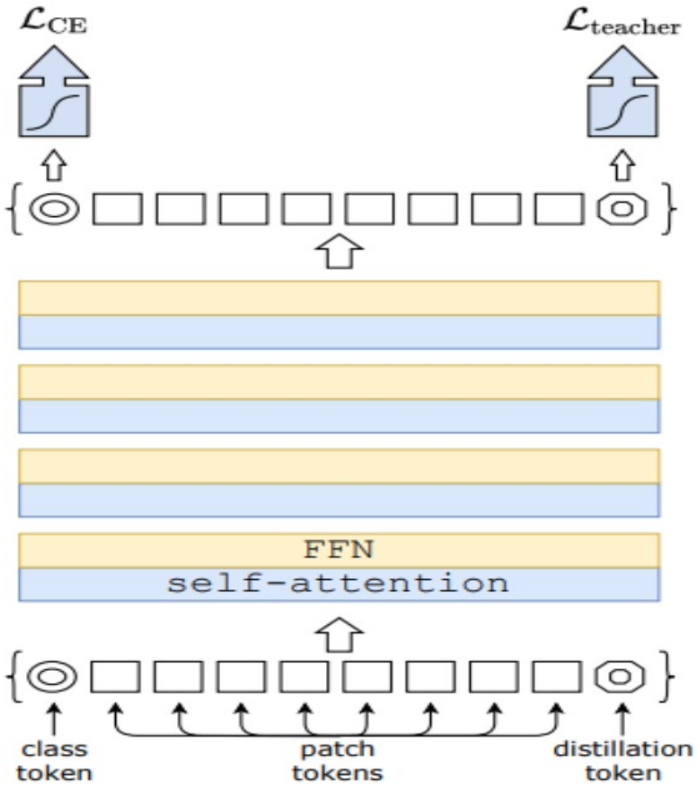
① : Training using the **soft labeled output of the teacher model** ➔ Learning **the weights of the teacher model**

② : Basic supervised learning with ground truth

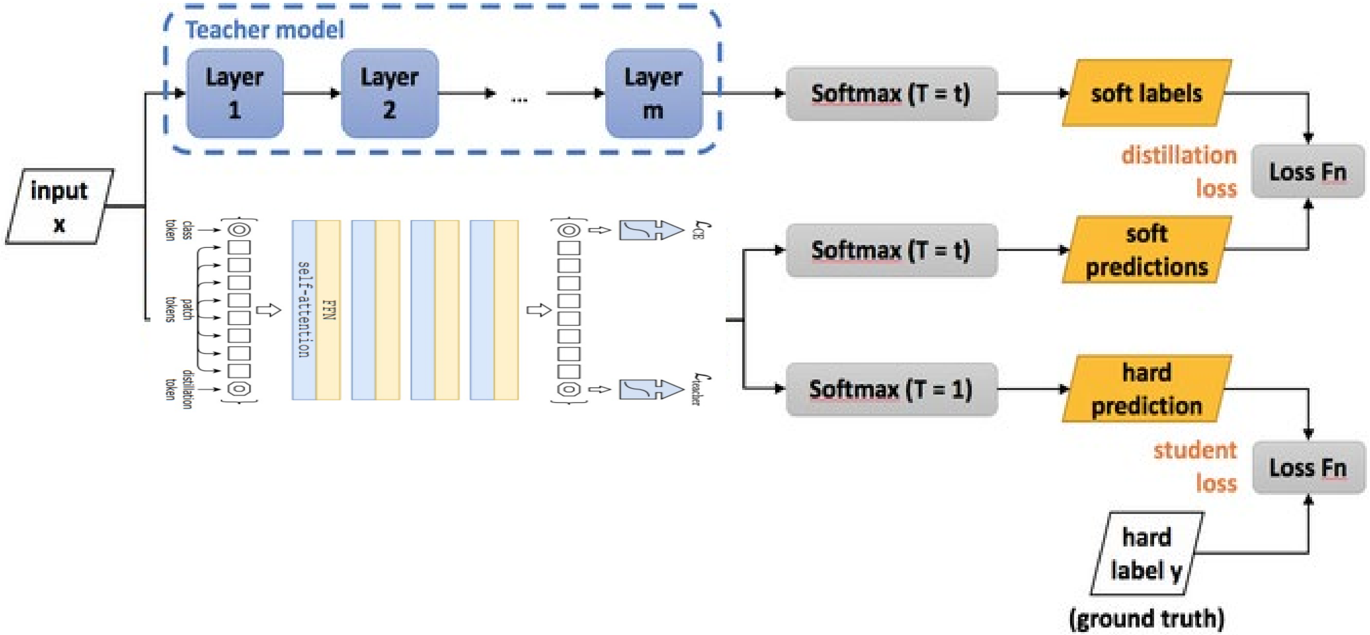


Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 2.7 (2015).

# • Background #DeiT




<Fig2> DeiT Model Structure



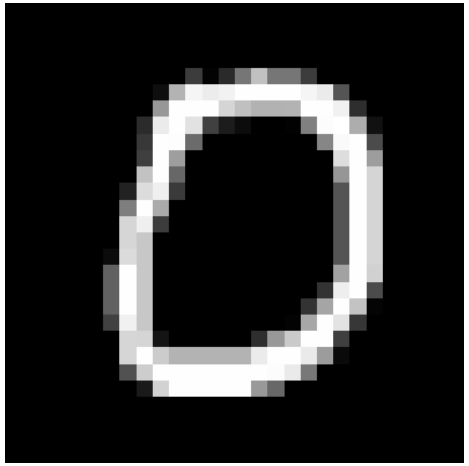

<Fig3> DeiT applied knowledge distillation structure



Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." International Conference on Machine Learning. PMLR, 2021.

# • Background #Adversarial Attack

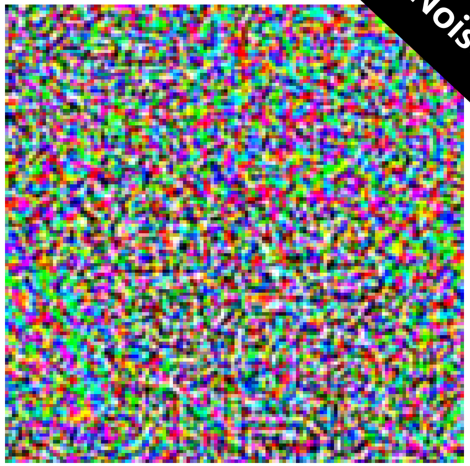
Original Image → Adversarial attack → Adversarial example



"0"
97.2% Confidence
(Correct classification)

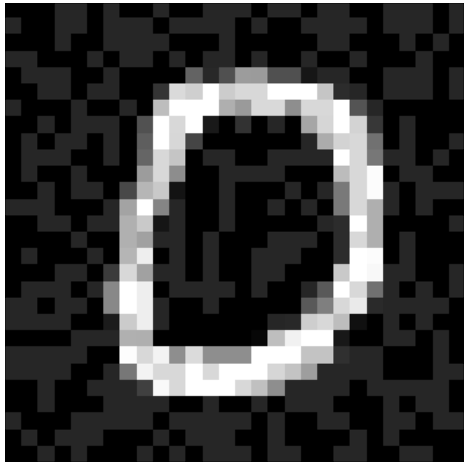
Epsilon
+ 0.15 x

Noise



Perturbation

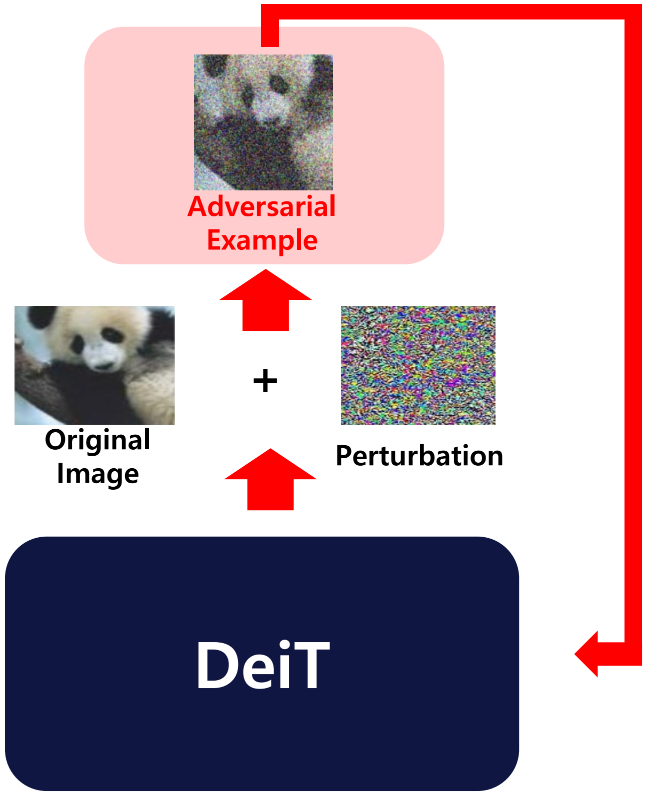
=



Predicted = "9"
99.4% Confidence
(Misclassification)

<Fig4> Adversarial Attack Example

# • Background #Adversarial Attack




<Fig5> Adversarial Attack Process

**DeiT model Attack**

Inducing

1) Model's **confidence decrease**

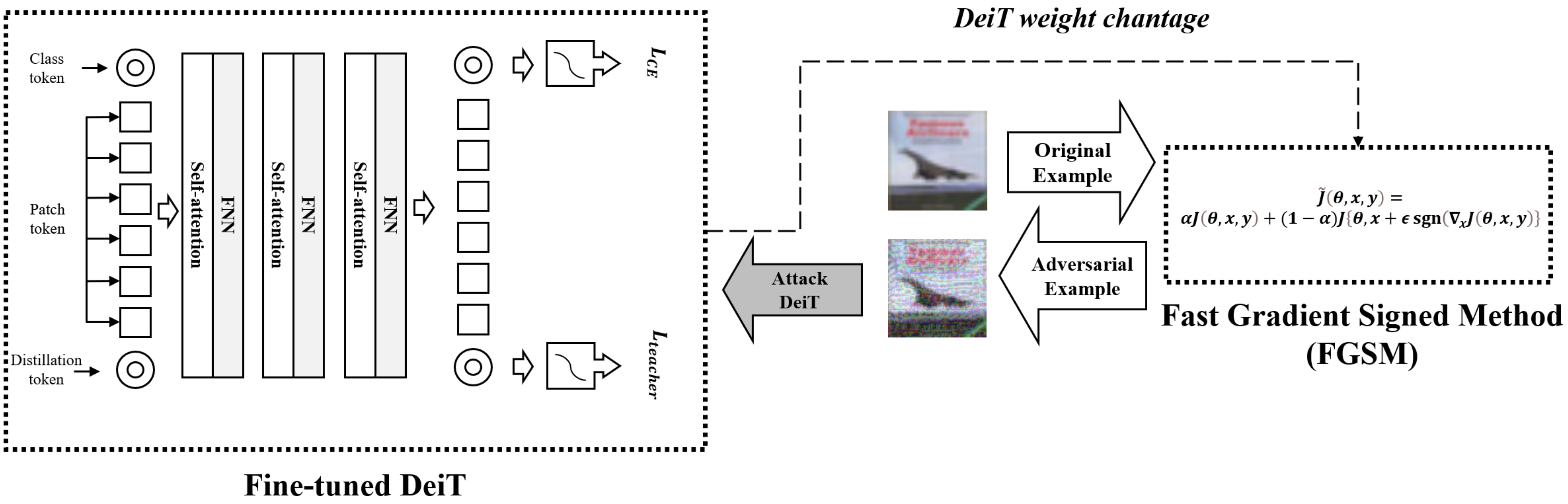
2) **Misclassification**

- The **most dangerous attack** in the field of **AI security**

- **Typical adversarial attack technique**

  1) **FGSM**
  2) PGD
  3) C&W Attack

# • Method



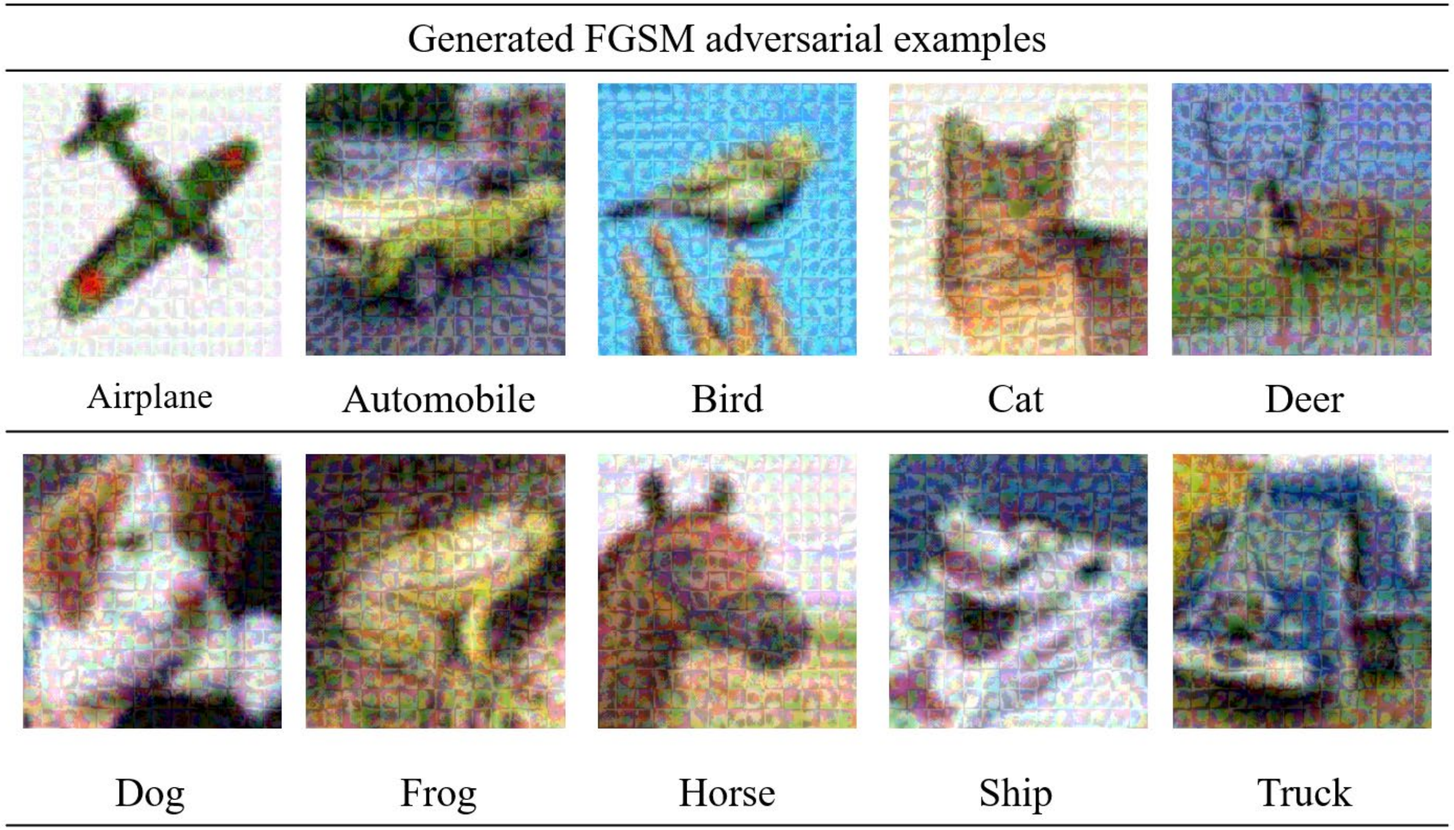

<Fig6> Experiment Overview

## DeiT's security verification method

1) **Fine-tune** of DeiT
2) **Weight extract** of fine-tuned DeiT
3) **Generate of FGSM adversarial examples** through DeiT weights
4) **Insert of generated adversarial examples** into fine-tuned DeiT and **performance analysis**

# • Experiment & Evaluation

- **Generating adversarial examples through FGSM**




<Fig7> Generated FGSM adversarial examples

- Generated by FGSM adversarial examples are composed of 100 images per class (A total of 1,000 images)

# • Experiment & Evaluation

▪ **DeiT performance verification (Accuracy)**

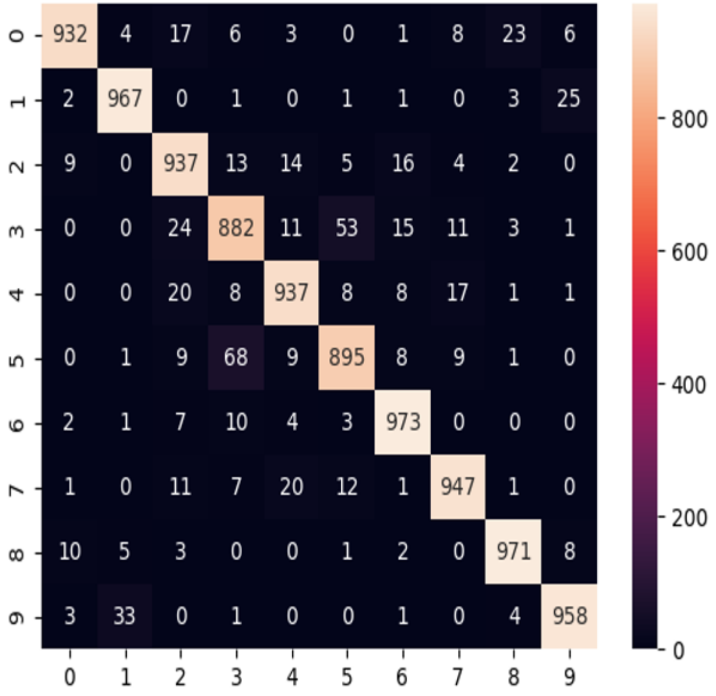
| Model | Attack | Dataset | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|---|---|
| DeiT | X | Cifar-10 | 0.9399 | 0.9399 | 0.9399 | **0.9399** |
| | O | Adversarial examples | 0.1050 | 0.1050 | 0.1049 | **0.1050** |

<Table 1> DeiT Verification Results

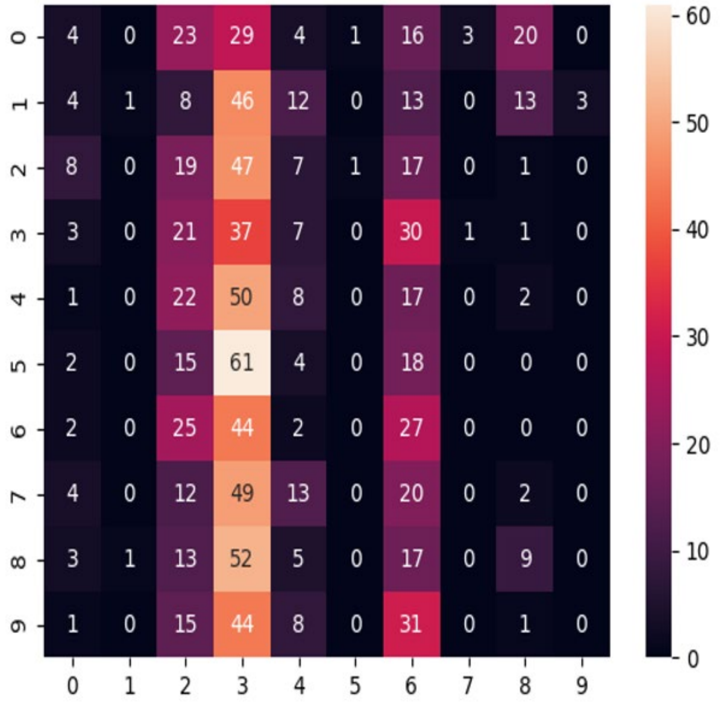
- **83.49% drop** compared to the normal dataset in case of adversarial attack
- We verified that **adversarial attack on DeiT is fatal**

# • Experiment & Evaluation

## ▪ **DeiT performance verification (Confusion Matrices)**




<Fig8> Original Dataset's Verification (DeiT)




<Fig9> Adversarial Examples Verification (DeiT)

- Adversarial example confusion matrix in Figure 9 : **Proper classification was not achieved.**

# • Conclusion & Future Work

- Currently, performance-oriented research is actively in progress in the field of AI.
- However, as new deep learning models develop, research from a security perspective must also be conducted.
- Through this experiment, we suggest that DeiT's security problem exists.

  **(DeiT accuracy drop of 83.49% in FGSM Attack)**

- **We suggest**

  1) *DeiT's defensive limitations exist*
  2) *Need to address DeiT's security vulnerabilties*

- **Future Work**

  1) *Create a robust DeiT Model (against adversarial attack)*

# Thank you

Inpyo Hong

hip9863@gachon.ac.kr
Dept. of Computer Engineering, Gachon University

가천대학교
Gachon University