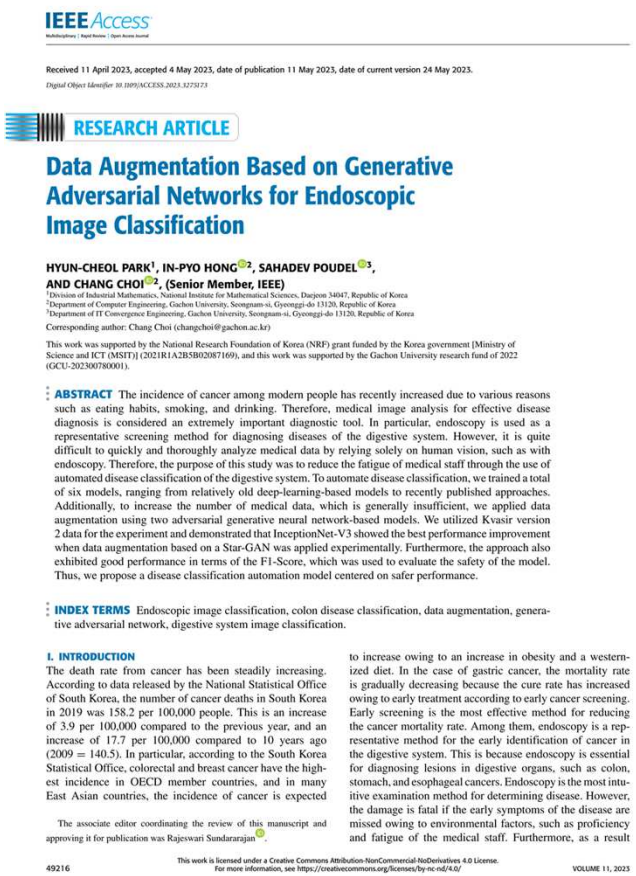
Research on Adversarial Defense Approach Based on Denoising and Distillation

가천대학교 IT융합공학과 (컴퓨터공학 전공) 석사과정
지도교수 : 최 창 교수님
홍인표

2023.11.23

Background



IEEE Access

Received 11 April 2023, accepted 4 May 2023, date of publication 11 May 2023, date of current version 24 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3275173

RESEARCH ARTICLE

Data Augmentation Based on Generative Adversarial Networks for Endoscopic Image Classification

HYUN-CHEOL PARK[1], IN-PYO HONG[2], SAHADEV POUDEL[3], AND CHANG CHOI[2], (Senior Member, IEEE)

[1]Division of Industrial Mathematics, National Institute for Mathematical Sciences, Daejeon 34047, Republic of Korea
[2]Department of Computer Engineering, Gachon University, Seongnam-si, Gyeonggi-do 13120, Republic of Korea
[3]Department of IT Convergence Engineering, Gachon University, Seongnam-si, Gyeonggi-do 13120, Republic of Korea

Corresponding author: Chang Choi (changchoi@gachon.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government [Ministry of Science and ICT (MSIT)] (2021R1A2B5B02087169), and this work was supported by the Gachon University research fund of 2022 (GCU-202300780001).

ABSTRACT The incidence of cancer among modern people has recently increased due to various reasons such as eating habits, smoking, and drinking. Therefore, medical image analysis for effective disease diagnosis is considered an extremely important diagnostic tool. In particular, endoscopy is used as a representative screening method for diagnosing diseases of the digestive system. However, it is quite difficult to quickly and thoroughly analyze medical data by relying solely on human vision, such as with endoscopy. Therefore, the purpose of this study was to reduce the fatigue of medical staff through the use of automated disease classification of the digestive system. To automate disease classification, we trained a total of six models, ranging from relatively old deep-learning-based models to recently published approaches. Additionally, to increase the number of medical data, which is generally insufficient, we applied data augmentation using two adversarial generative neural network-based models. We utilized Kvasir version 2 data for the experiment and demonstrated that InceptionNet-V3 showed the best performance improvement when data augmentation based on a Star-GAN was applied experimentally. Furthermore, the approach also exhibited good performance in terms of the F1-Score, which was used to evaluate the safety of the model. Thus, we propose a disease classification automation model centered on safer performance.

INDEX TERMS Endoscopic image classification, colon disease classification, data augmentation, generative adversarial network, digestive system image classification.

I. INTRODUCTION

The death rate from cancer has been steadily increasing. According to data released by the National Statistical Office of South Korea, the number of cancer deaths in South Korea in 2019 was 158.2 per 100,000 people. This is an increase of 3.9 per 100,000 compared to the previous year, and an increase of 17.7 per 100,000 compared to 10 years ago (2009 = 140.5). In particular, according to the South Korea Statistical Office, colorectal and breast cancer have the highest incidence in OECD member countries, and in many East Asian countries, the incidence of cancer is expected to increase owing to an increase in obesity and a westernized diet. In the case of gastric cancer, the mortality rate is gradually decreasing because the cure rate has increased owing to early treatment according to early cancer screening. Early screening is the most effective method for reducing the cancer mortality rate. Among them, endoscopy is a representative method for the early identification of cancer in the digestive system. This is because endoscopy is essential for diagnosing lesions in digestive organs, such as colon, stomach, and esophageal cancers. Endoscopy is the most intuitive examination method for determining disease. However, the damage is fatal if the early symptoms of the disease are missed owing to environmental factors, such as proficiency and fatigue of the medical staff. Furthermore, as a result

The associate editor coordinating the review of this manuscript and approving it for publication was Rajeswari Sundararajan.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

Title:

Data Augmentaion based on Generative Adversarial Networks for Endoscopic Image Classification

Authors :

Hyun-Cheol Park*, In-Pyo Hong*, Sahadeve Poudel and Changchoi
(* = Equal Contribution)

Journals :

IEEE Access (IF : 3.9)

Keywords:

Medical AI, GAN, Data Augmentation

Background

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

In-pyo Hong*
Gachon University
Seongnam-si, Gyeonggi-do
Republic of Korea
hip9863@gachon.ac.kr

Gyu-ho Choi†
Gachon University
Seongnam-si, Gyeonggi-do
Republic of Korea
ghchoi@gachon.ac.kr

Pan-koo Kim‡
Chosun University
Gwangju, Republic of Korea
pkkim@chosun.ac.kr

Chang Choi§
Gachon University
Seongnam-si, Gyeonggi-do
Republic of Korea
changchoi@gachon.ac.kr

ABSTRACT

Recently, research using knowledge distillation in artificial intelligence (AI) has been actively conducted. In particular, data-efficient image transformer (DeiT) is a representative transformer model using knowledge distillation in image classification. However, DeiT's safety against the patch unit's adversarial attacks was not verified. Furthermore, existing DeiT research did not prove security robustness against adversarial attacks. In order to verify the vulnerability of adversarial attacks, we conducted an attack using the fast gradient sign method (FGSM) targeting the DeiT model based on knowledge distillation. As a result of the experiment, an accuracy of 93.99% was shown in DeiT verification based on Normal data (Cifar-10). In contrast, when verified with abnormal data based on FGSM (adversarial examples), the accuracy decreased by 83.49% to 10.50%. By analyzing the vulnerability pattern related to adversarial attacks, we confirmed that FGSM showed successful attack performance through weight control of DeiT. Moreover, we verified that DeiT has security limitations for practical application.

CCS CONCEPTS

• **Security and privacy** → *Penetration testing.*

KEYWORDS

Deit, Knowledge distillation, Adversarial attacks, FGSM

ACM Reference Format:

In-pyo Hong, Gyu-ho Choi, Pan-koo Kim, and Chang Choi. 2023. Security Verification Software Platform of Data-efficient Image Transformer Based on Fast Gradient Sign Method. In *The 38th ACM/SIGAPP Symposium on Applied Computing (SAC '23), March 27- March 31, 2023, Tallinn, Estonia.* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3555776.3577731

*Conceptualization, Methodology, Writing - Editing
†Software, Writing - Review
‡Investigation
§Corresponding author, Supervision, Project administration, Funding acquisition

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SAC '23, March 27-March 31, 2023, Tallinn, Estonia

© 2023 Association for Computing Machinery.

ACM ISBN 978-1-4503-9517-5/23/03…$15.00
https://doi.org/10.1145/3555776.3577731

1 INTRODUCTION

In the field of computer vision, many deep learning models based on convolutional neural networks (CNN) have been developed with excellent performance[9]. However, research on computer vision processing based on a transformer being actively carried out as the transformer model recently showed excellent performance in vision tasks[13]. The most representative transformer model used in computer vision is a vision transformer (ViT). ViT outperforms existing CNN models by using encoder and class token[13]. However, ViT requires a very large amount of data for pre-training. Therefore, ViT has a limit that individual users cannot construct pre-training properly[12]. To address this limitation, a model called data-efficient image transformer (DeiT) was proposed[12]. The DeiT model is based on a transformer that uses knowledge distillation. DeiT uses the student model and teacher model. The student model efficiently learns from the teacher model by output smoothing. When training DeiT, the student model used ViT, and the teacher model used CNN models for solving the low inductive bias pointed out as a limitation of the existing ViT[12]. However, there may still be limitations in that DeiT using ViT makes it vulnerable to adversarial attacks. Furthermore, knowledge distillation's security limitations have not been verified. If that limitation exists, there is a significant difficulty in the practical use of DeiT as an image classification model. Thus, knowledge distillation's vulnerability verification is required for practical use. In this paper, we verified the DeiT model's security rather than the performance verification of DeiT to solve the safety problem of the unverified DeiT. We also analyzed whether ViT vulnerabilities still exist in DeiT. For the security verification of DeiT, we used an adversarial attack[10]. An adversarial attack is considered the most representative attack in AI security currently. Among them, we used fast gradient sign method (FGSM) technique to create adversarial examples since FGSM is a baseline method in the adversarial attack[3]. As a result of verifying DeiT with the generated adversarial example, it induced a performance drop of 83.49% and confirmed that DeiT has a security vulnerability.

1669

Title:

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

Authors :

In-pyo Hong, Gyu-ho Choi, Pan-koo Kim and Chang choi

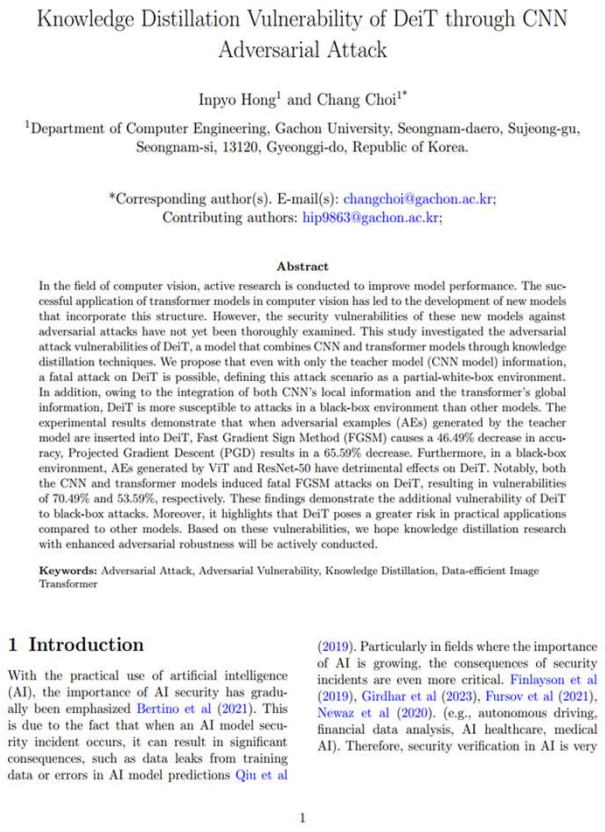
Conference:

ACM SAC '23 : Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing *(BK21 우수학회)*

Keywords:

Security for AI, Knowledge Distillation, DeiT

Background



Knowledge Distillation Vulnerability of DeiT through CNN Adversarial Attack

Inpyo Hong[1] and Chang Choi[1*]

[1]Department of Computer Engineering, Gachon University, Seongnam-daero, Sujeong-gu, Seongnam-si, 13120, Gyeonggi-do, Republic of Korea.

*Corresponding author(s). E-mail(s): changchoi@gachon.ac.kr;
Contributing authors: hip9863@gachon.ac.kr;

Abstract

In the field of computer vision, active research is conducted to improve model performance. The successful application of transformer models in computer vision has led to the development of new models that incorporate this structure. However, the security vulnerabilities of these new models against adversarial attacks have not yet been thoroughly examined. This study investigated the adversarial attack vulnerabilities of DeiT, a model that combines CNN and transformer models through knowledge distillation techniques. We propose that even with only the teacher model (CNN model) information, a fatal attack on DeiT is possible, defining this attack scenario as a partial-white-box environment. In addition, owing to the integration of both CNN's local information and the transformer's global information, DeiT is more susceptible to attacks in a black-box environment than other models. The experimental results demonstrate that when adversarial examples (AEs) generated by the teacher model are inserted into DeiT, Fast Gradient Sign Method (FGSM) causes a 46.49% decrease in accuracy, Projected Gradient Descent (PGD) results in a 65.59% decrease. Furthermore, in a black-box environment, AEs generated by ViT and ResNet-50 have detrimental effects on DeiT. Notably, both the CNN and transformer models induced fatal FGSM attacks on DeiT, resulting in vulnerabilities of 70.49% and 53.59%, respectively. These findings demonstrate the additional vulnerability of DeiT to black-box attacks. Moreover, it highlights that DeiT poses a greater risk in practical applications compared to other models. Based on these vulnerabilities, we hope knowledge distillation research with enhanced adversarial robustness will be actively conducted.

Keywords: Adversarial Attack, Adversarial Vulnerability, Knowledge Distillation, Data-efficient Image Transformer

1 Introduction

With the practical use of artificial intelligence (AI), the importance of AI security has gradually been emphasized Bertino et al (2021). This is due to the fact that when an AI model security incident occurs, it can result in significant consequences, such as data leaks from training data or errors in AI model predictions Qiu et al (2019). Particularly in fields where the importance of AI is growing, the consequences of security incidents are even more critical. Finlayson et al (2019), Girdhar et al (2023), Fursov et al (2021), Newaz et al (2020). (e.g., autonomous driving, financial data analysis, AI healthcare, medical AI). Therefore, security verification in AI is very

1

Title:

Knowledge Distillation Vulnerability of DeiT through CNN Adversarial Attack

Authors :

Inpyo Hong and Chang choi

Journal:

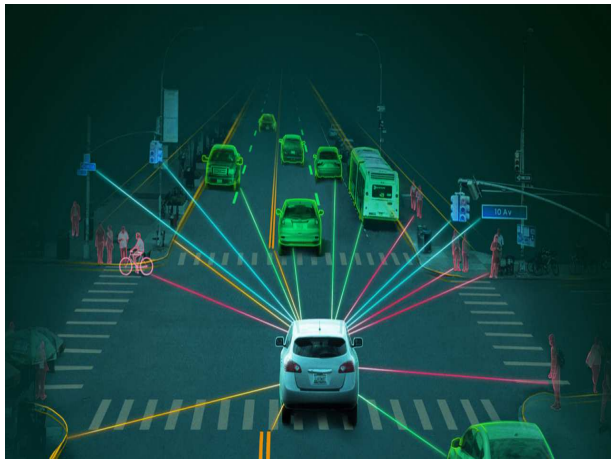
Neural Computing and Applications (IF: 6.0) - Accepted

Keywords:

Security for AI, Knowledge Distillation, DeiT, Partial-white box Attack

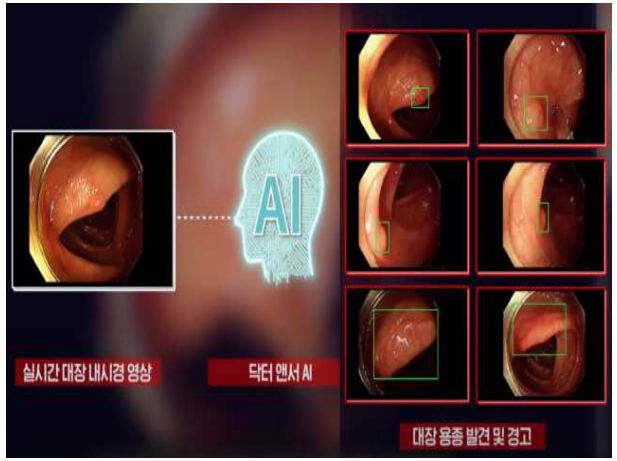
01 Introduction

최근, AI 기술의 발전에 따라 다양한 분야에서 AI기술이 상용화되어 활용되고 있음



<자율주행 기술>

<자연어처리 기술>



<의료 AI>

이에 따른 보안 문제도 중요한 고려사항으로 부상하고 있으며, AI 의 안전한 상용화를 위한 security for AI 연구는 필수적임

01 Introduction

현존하는 모든 AI 모델은 적대적 공격을 통해 제대로 된 의사결정을 하지 못한다는 보안적 결함이 존재함

<Fig1. AI 오분류를 위한 물리적 공격 예시 [1] >

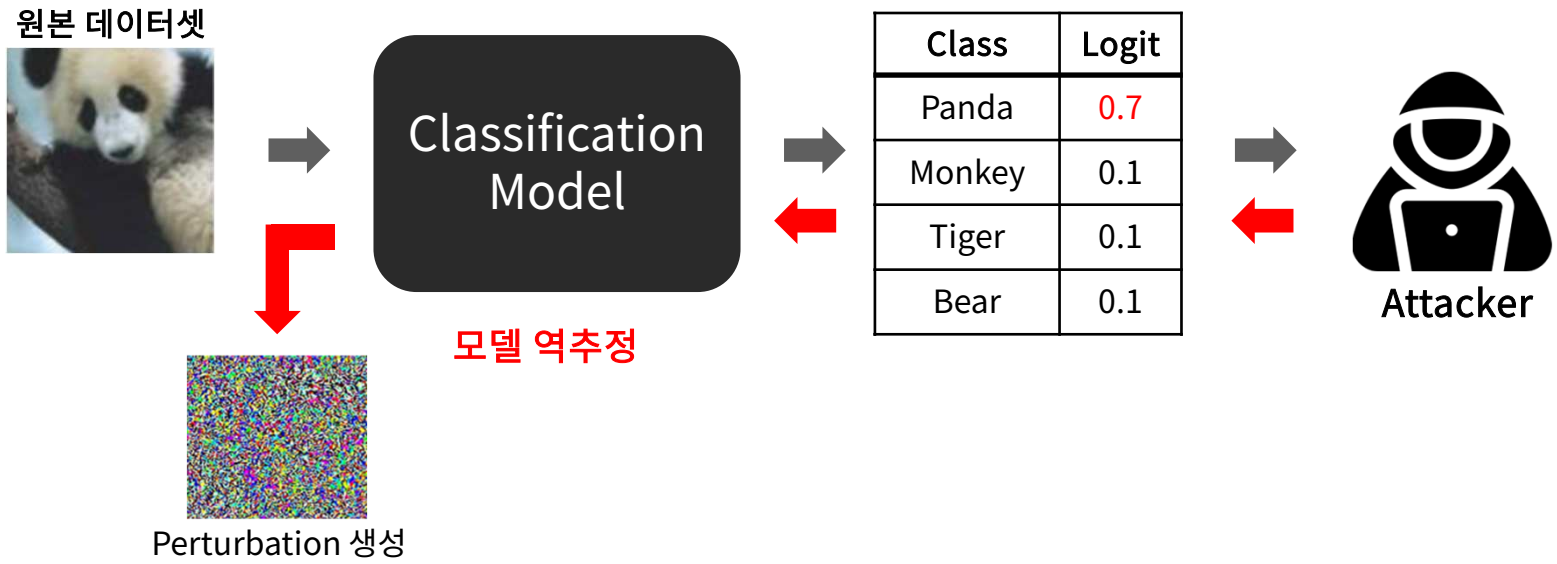
<Fig2. AI 오분류를 위한 디지털 공격 예시>

적대적 공격 (Adversarial Attack) ➔ Security for AI 분야에서 가장 위협적인 공격으로, AI의 Gradient 관련 취약점을 악용

적대적 공격에 대응하기 위해 많은 방어기법들이 연구되고 있지만
모델 파라미터 증가, 모델 훈련시간 증가, 모델 성능 하락 등 여러 한계가 존재함 ➔ 안전한 AI Model 상용화를 위한 연구가 필요함

02 Adversarial Attack

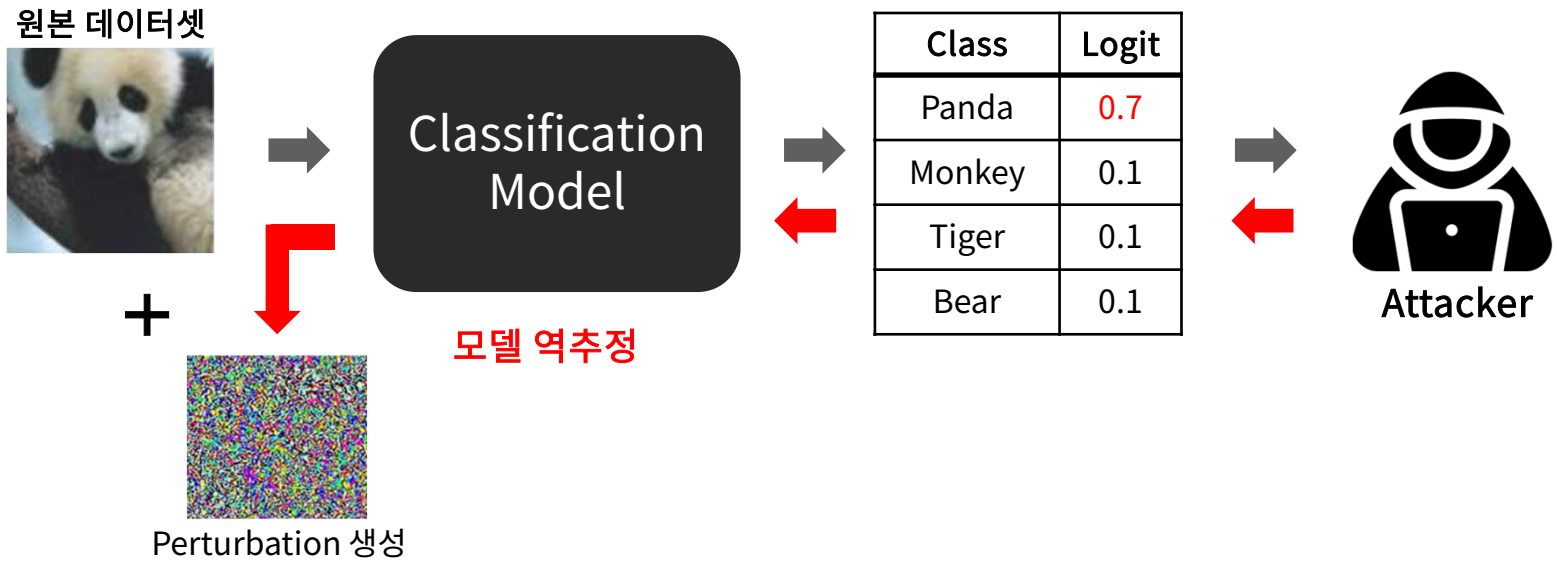
적대적 공격이란 Deep-learning 모델에 적대적 교란(Adversarial Perturbation)을 적용하여 모델의 오분류를 유발하고
신뢰도 감소를 야기하는 공격기법임



<Fig3. 적대적 공격 동작 원리>

02 Adversarial Attack

적대적 공격이란 Deep-learning 모델에 적대적 교란(Adversarial Perturbation)을 적용하여 모델의 오분류를 유발하고
신뢰도 감소를 야기하는 공격기법임

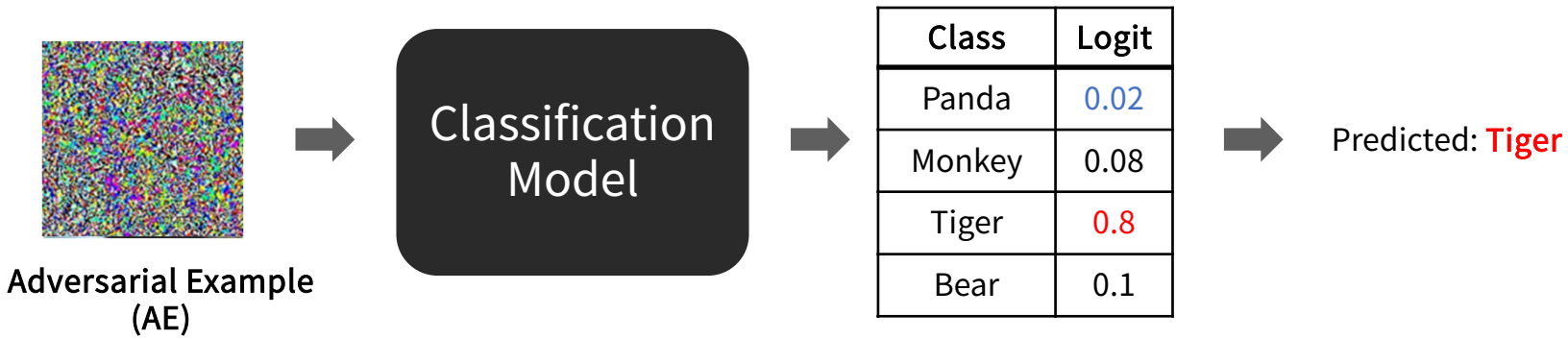


Class	Logit
Panda	0.7
Monkey	0.1
Tiger	0.1
Bear	0.1

<Fig3. 적대적 공격 동작 원리>

02 Adversarial Attack

적대적 공격이란 Deep-learning 모델에 적대적 교란(Adversarial Perturbation)을 적용하여 모델의 오분류를 유발하고
신뢰도 감소를 야기하는 공격기법임



<Fig3. 적대적 공격 동작 원리>

본 연구에서는 적대적 공격에서 base-line으로 사용되는 Fast Gradient Sign Method(FGSM)를 활용해 방어성능을 검증함

02 Defense Methods

적대적 공격에 따른 방어 기법은 다음과 같음

방어기법	기여	한계
적대적 훈련 [4]	현존 방어기법 중 **유일하게 근본적인 기법**, **방어성능이 가장 우수**	**훈련시간 소요 ↑**, **정확도 하락** 문제 발생
Defensive Distillation [7]	Gradient의 변형 ➔ 방어적 효과가 존재함을 제시 **지식증류의 방어적 활용 가능성** 제시	**Gradient 추정 시 방어성능 무력화**
Feature Denoising [3]	Perturbation의 제거 ➔ **AE에 직접적인 대처**	**원본데이터 손상**, Denoising모듈에 대한 AE 생성 시 방어성능 무력화

<Table1. 적대적 공격 방어기법 특징>

02 Defense Methods

적대적 공격에 대응하기 위해 **방어기법을 융합한 연구도 활발히 진행 중**이나, 여전히 한계가 존재하며 추가 연구가 필요함

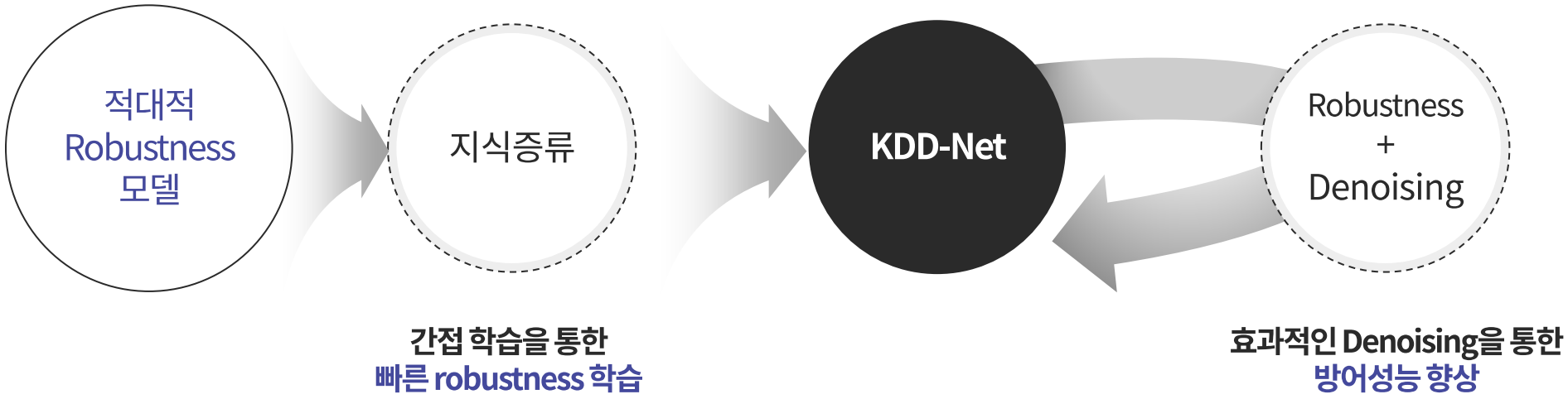
방어기법	기여	한계
적대적 훈련 + 지식증류 [8]	적대적 robustness의 **간접학습** ➔ **학습시간 ↓**	**직접적인 적대적 훈련**보다 **robustness ↓**
적대적 훈련 + Denoising [9]	정확도 하락 X, **효과적인 denoising ➔ Robustness ↑**	적대적 훈련 필수 ➔ **학습시간 ↑**

<Table2. 융합 방어기법 특징>

본 연구에서는 **각 융합기법**의 장점은 유지하며, 한계를 개선하기 위한 모델을 구축하고자 함

03 Proposed Method

본 연구는 지식증류 + Denoising을 융합한 방어 모델(KDD-Net: Knowledge Distillation and Denoising Network)을 제안함



03 Proposed Method

제안 모델의 contributions는 다음과 같음

1. 방어성능의 향상 및 정확도 하락 최소화

기존 방어기법보다
강인한 방어성능을 보이면서,
Accuracy의 하락 최소화

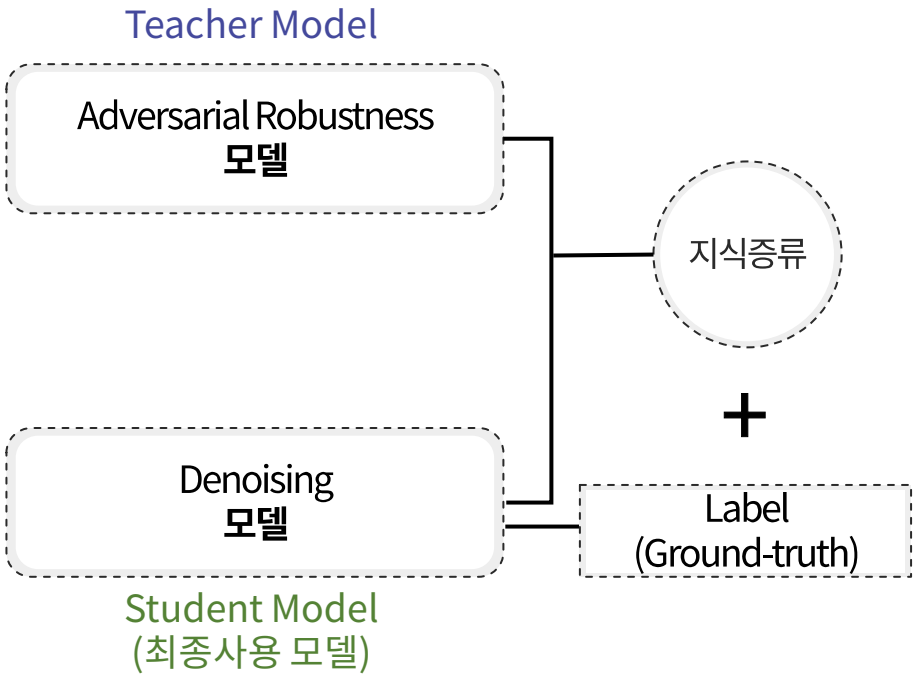
2. 파라미터 및 학습시간을 고려한 효율적 학습

파라미터 및 학습시간의 증가는 최소화시키며,
방어성능을 향상

security for AI 측면의 상용화 연구에 기여

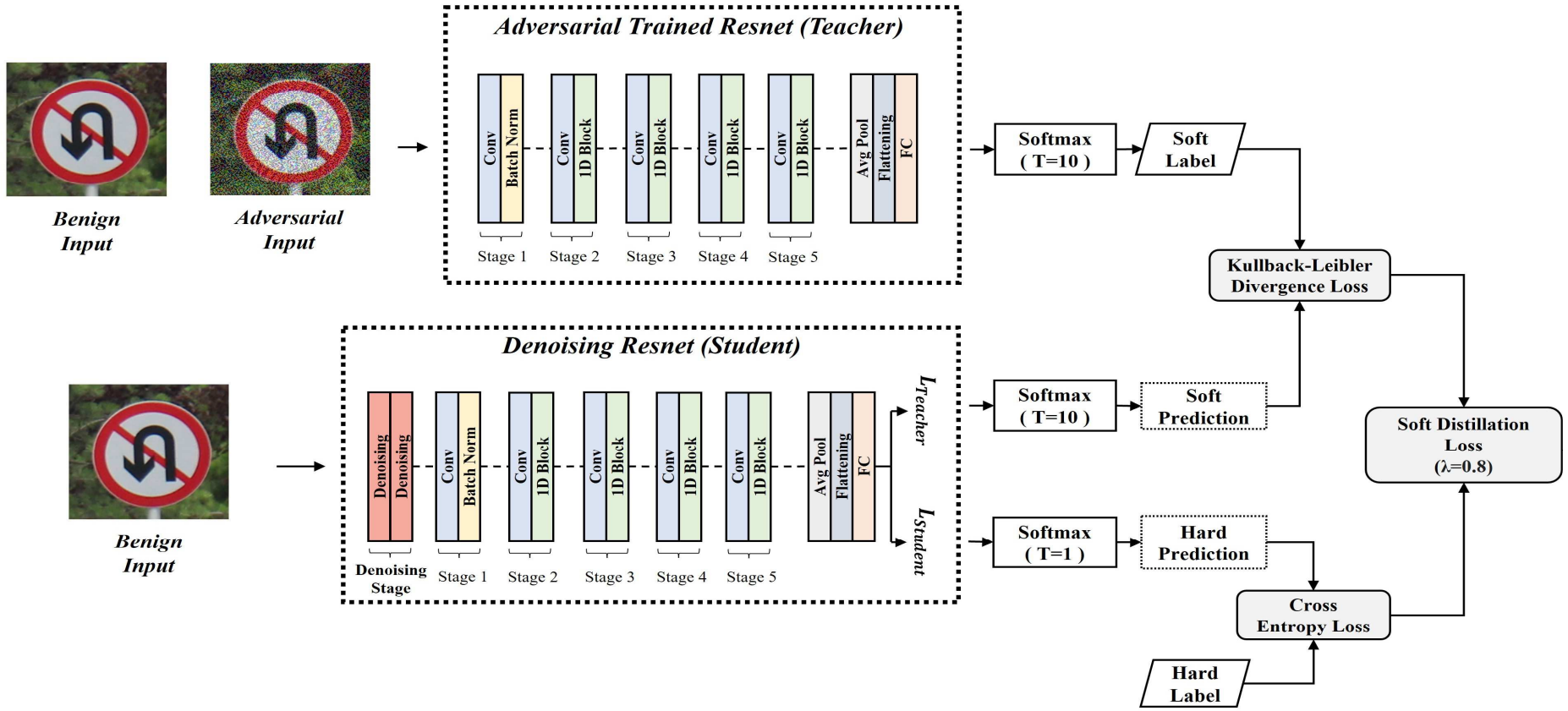
03 Proposed Method

제안모델 (KDD-Net)의 학습 구조는 다음과 같음



<Fig4. KDD-Net Training process>

03 Proposed Method



<Fig5. 제안모델(KDD-Net) 구조>

03 Proposed Method

제안모델 (KDD-Net)의 Algorithm은 다음과 같음

Algorithm 1: Defensive Model Training Process based on Distillation

Require: Teacher Model T, Denoising Model D, Temperature τ, λ

$\tau \leftarrow 10$

$\lambda \leftarrow 0.8$

for (x,y) **in** data **do**

  $Output_T \leftarrow \frac{T(x)}{\tau}$

  $Output_D \leftarrow D(x)$

  $Output_{Distilled\ D} \leftarrow \frac{T(x)}{\tau}$

  $loss_{soft} \leftarrow criterion_{soft}(O_{Distilled\ D}, Output_T)$

  $loss_{hard} \leftarrow criterion_{hard}(Output_D, y)$

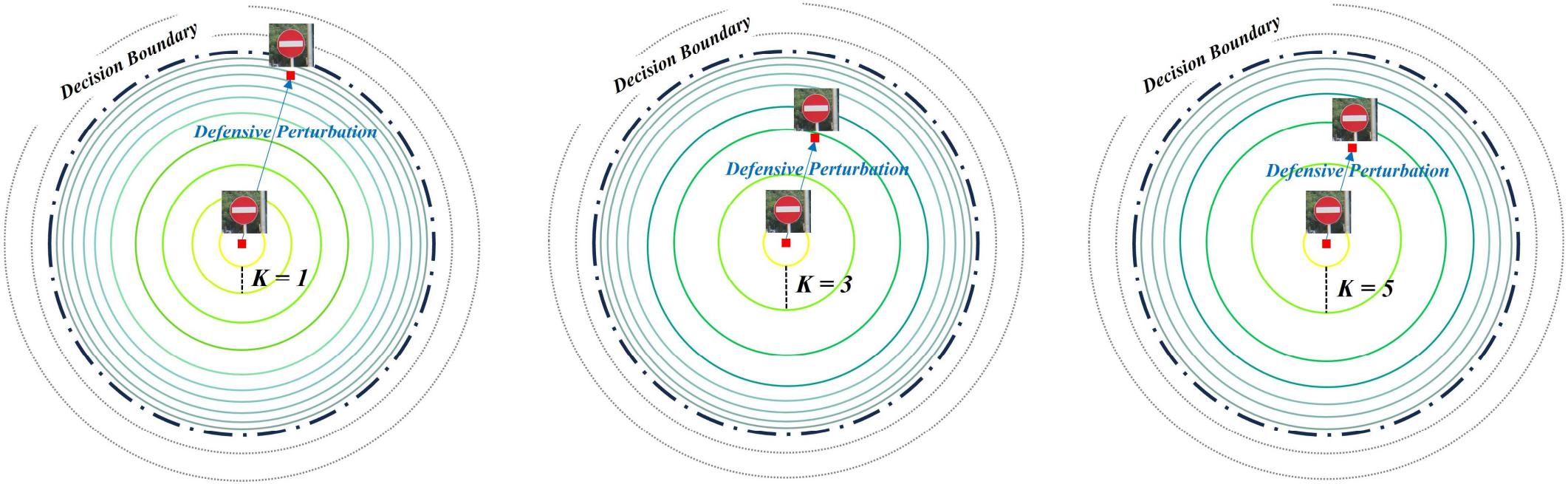
  $soft\ distillation\ loss \leftarrow (1\text{-}\lambda)\ loss_{soft} + \lambda\ loss_{hard}\ \tau^2$

  $soft\ distillation\ loss$ .gradient_update()

end

03 Proposed Method

제안모델 (KDD-Net)의 방어성능 향상 요인은 다음과 같음



(A) Denoising

(B) Adversarial Training + Denoising

(C) Knowledge Distillation + Denoising (Proposed Method)

04 Experimental Environment

실험환경:

Windows 10, NVIDIA RTX 3090, Pytorch

실험조건:

제안모델(KDD-Net)평가를 위해 **4개의 비교모델 구축,**
K-fold Validation (K=5)로 실험 진행

데이터셋:

Cifar-10 데이터셋

학습모델	Loss Function	Optimizer (Learning rate)	Epoch
비교 모델 (CNN)	Cross-Entropy Loss	Adam (1e-4)	300 (Early Stop)
비교모델 (지식증류)	Soft Distillation Loss (T=10, λ=0.8)	Adam (1e-4)	300 (Early Stop)
KDD-NET (Ours)	Soft Distillation Loss (T=10, λ=0.8)	Adam (1e-4)	300 (Early Stop)

<Table3. 모델에 따른 학습조건>

04 Experimental Evaluation

모델 실용화를 위해서는 방어성능 뿐만 아니라 모델 parameter, 학습시간 등이 중요한 영향을 미침
따라서, 모델 활용을 위한 종합적인 척도를 비교하여 분석함

1. 방어성공률 및 정확도

모델 성능 및 방어성능평가를 위한
방어성공률 및 정확도 평가

2. 모델 Parameter 및 학습시간

다양한 환경에서의 실용화를 위한
모델 크기 및 훈련시간 평가

Model	Defense Method	Attack	Precision	Recall	F1-Score	Accuracy	공격성공률 (%)	방어성공률 (%)
ResNet18 [2]	-	False	0.919 (±0.011)	0.918 (±0.010)	0.917 (±0.010)	0.918 (±0.010)	66.5	33.5
		True	0.259 (±0.015)	0.253 (±0.023)	0.254 (±0.019)	0.253 (±0.023)		
	DN [3]	False	0.822 (±0.007)	0.819 (±0.007)	0.819 (±0.007)	0.819 (±0.007)	68.3	31.7
		True	0.154 (±0.013)	0.136 (±0.017)	0.142 (±0.016)	0.136 (±0.017)		
	ADT [4]	False	0.774 (±0.003)	0.767 (±0.004)	0.766 (±0.005)	0.767 (±0.004)	28.2	71.8
		True	0.495 (±0.006)	0.484 (±0.004)	0.484 (±0.008)	0.484 (±0.004)		
	DN + ADT [6]	False	0.773 (±0.003)	0.770 (±0.004)	0.769 (±0.004)	0.770 (±0.004)	28.0	72.0
		True	0.493 (±0.010)	0.491 (±0.006)	0.488 (±0.007)	0.491 (±0.006)		
	KD [5]	False	0.848 (±0.004)	0.848 (±0.004)	0.848 (±0.004)	0.848 (±0.004)	28.4	71.6
		True	0.573 (±0.009)	0.564 (±0.009)	0.567 (±0.009)	0.564 (±0.009)		
KDD-Net (Proposed)	**KD + DN**	False	0.848 (±0.002)	0.847 (±0.002)	0.847 (±0.002)	0.847 (±0.002)	**27.3**	**72.7**
		True	0.586 (±0.004)	0.575 (±0.004)	0.579 (±0.004)	0.575 (±0.004)		

<Table 5. 방어모델 성능평가>

04 Experimental Evaluation

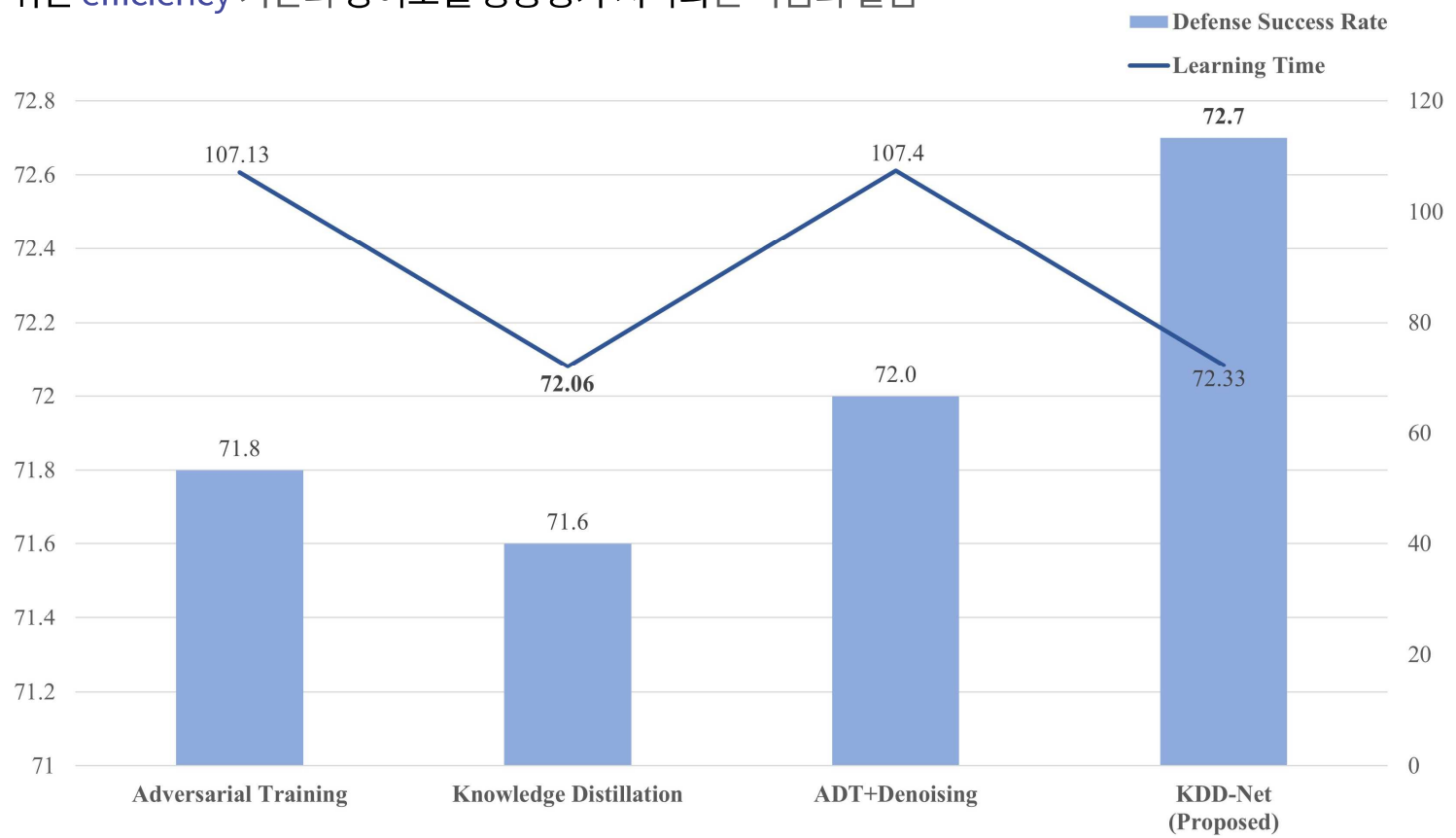
모델 상용화를 위한 efficiency 기반의 방어모델 성능평가는 다음 표와 같음

Model	Defense Method	Model Parameter	Learning Time	방어성공률 (%)
ResNet18 [2]	-	11.17M	1:06:22	33.5
	DN [3]	11.24M	1:11:46	31.7
	ADT [4]	11.17M	1:47:13	71.8
	DN + ADT [6]	11.24M	1:47:40	72.0
	KD [5]	11.17M	1:12:06	71.6
KDD-Net (Proposed)	**KD + DN**	**11.24M**	**1:12:33**	**72.7**

<Table 6. 방어모델 Efficiency 평가>

04 Experimental Evaluation

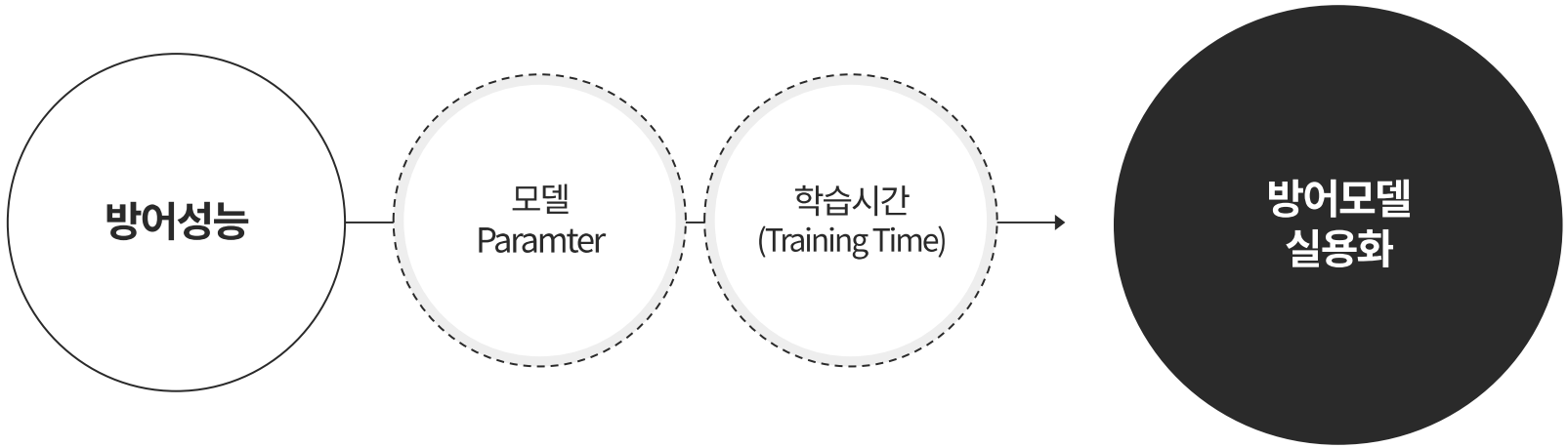
모델 상용화를 위한 efficiency 기반의 방어모델 성능평가 시각화는 다음과 같음



<Fig5. 방어모델 간 Efficiency 시각화 Graph>

05 Discussion & Conclusion

방어 모델 구축 시 가장 중요한 척도는 방어성능이지만, IoT 및 임베디드 기기의 적용을 위한 컴퓨팅 자원은 한정되어 있음.
따라서, 실제 사용자의 접근성을 향상시키고 방어모델의 실용화를 위해서는 방어성능 뿐만 아니라 여러 요소들이 고려되어야 함



05 Discussion & Conclusion

현재까지 적대적 공격을 100% 방어하는 방어기법은 없으며, 보안적 관점에서 완벽한 AI 방어모델은 존재하지 않음

하지만, AI는 이미 실용화되어 사용되고 있기 때문에 방어성능 뿐만 아닌 여러 요소도 고려된 방어모델 연구가 필요함

이에 본 연구에서는 모델크기 및 학습시간을 최소화 시킴과 동시에 방어성능이 향상된 모델(KDD-Net)을 제안함

하지만 보안적으로 더욱 강건한 AI모델 연구가 진행될 필요가 있음

따라서, 향후연구로 암호화 기법의 AI 적용 연구(E.g., Fully Homomorphic Encryption)를 진행할 예정임

Reference

[1] Povolny, Steve, and Shivangee Trivedi. "Model hacking ADAS to pave safer roads for autonomous vehicles." McAfee Blogs (2020).

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[3] X. Li and F. Li, “Adversarial examples detection in deep networks with convolutional filter statistics,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5764–5772.

[4] B. Wu, H. Pan, L. Shen, J. Gu, S. Zhao, Z. Li, D. Cai, X. He, and W. Liu, “Attacking adversarial attacks as a defense,” arXiv preprint arXiv:2106.04938, 2021.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” arXiv preprint arXiv:1412.6572, 2014.

[6] M. Goldblum, L. Fowl, S. Feizi, and T. Goldstein, “Adversarially robust distillation,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 04, 2020, pp. 3996–4003.

[7] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, “Distillation as a defense to adversarial perturbations against deep neural networks,” in 2016 IEEE symposium on security and privacy (SP). IEEE, 2016, pp. 582–597.

THANK YOU

hip9863@gachon.ac.kr