Smart Service Application

202245234 Inpyo-Hong

Index

Introduction

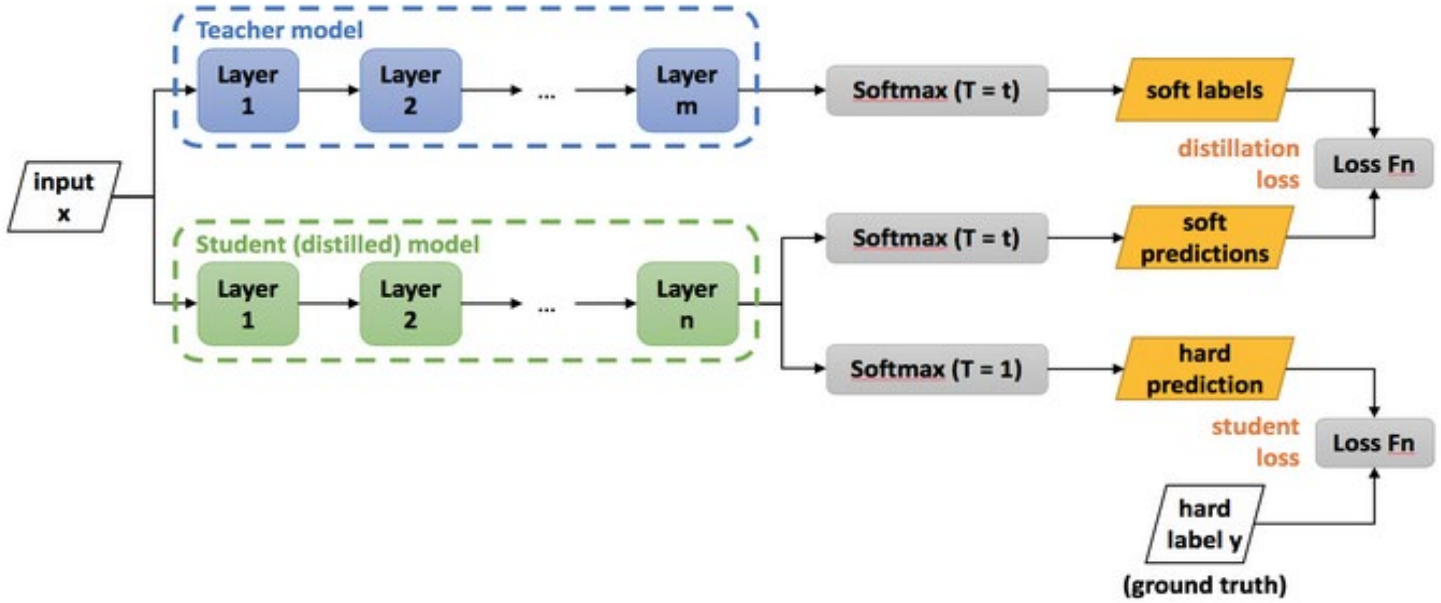
Current Computer Vision research focuses on model performance

The Emergence of new structure model, But not verified against adversarial attacks

We analyzed the vulnerability of **DeiT(ViT based on knowledge distillation)**

We verified by dividing it into **white box, semi-white box and black box** environments

Related Work



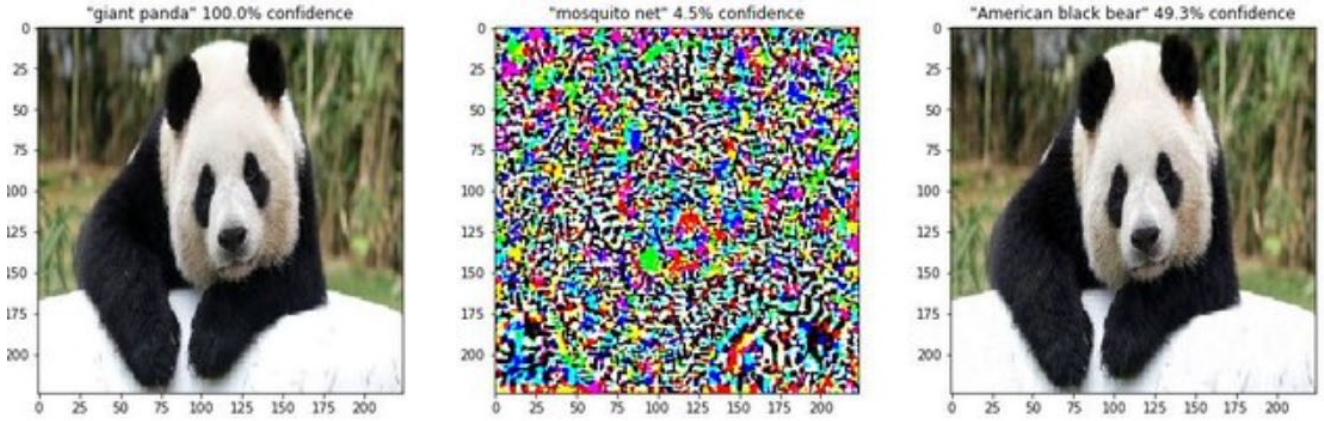
Knowledge Distillation Structure

Hard Distillation Loss : $$\mathcal{L}_{\text{global}}^{\text{hardDistill}} = \frac{1}{2}\mathcal{L}_{\text{CE}}(\psi(Z_s), y) + \frac{1}{2}\mathcal{L}_{\text{CE}}(\psi(Z_s), y_{\text{t}}).$$

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

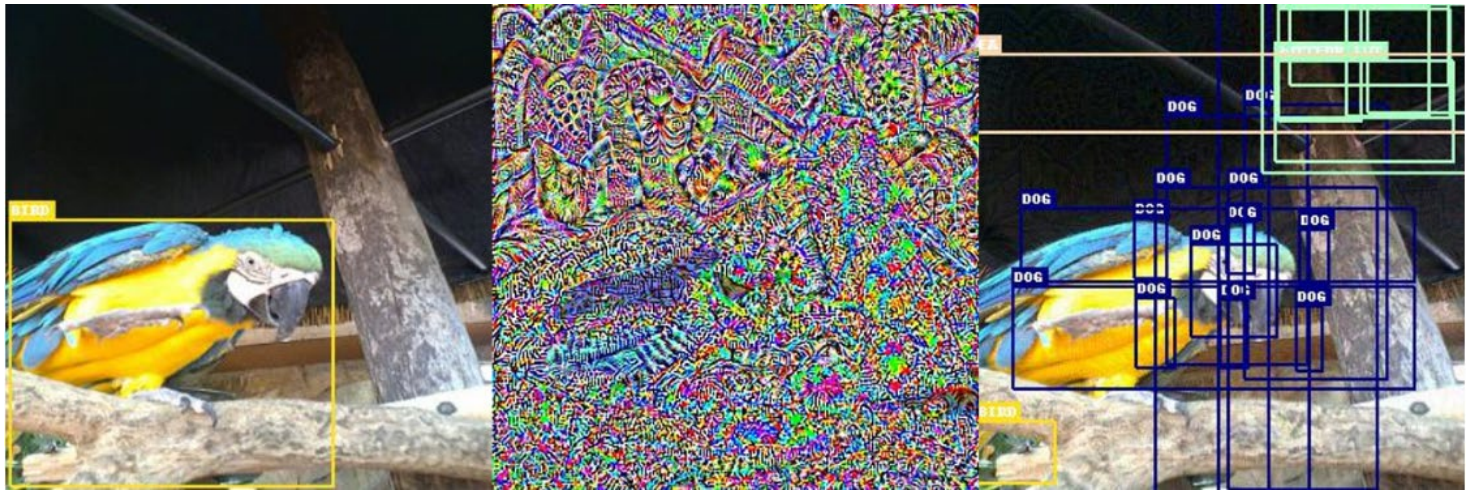
Related Work

- **FGSM (Fast Gradient Sign Method)**
  - 적대적 예제(Adversarial Example)를 생성하는 가장 기본적인 알고리즘 기법
  - 타겟모델의 손실 함수에 대한 기울기를 계산하고 noise를 추가하여 학습을 방해
  - Noise를 통한 경사상승법 (Gradient absent)으로 모델 학습 시 기울기 감소를 방해하는 방식



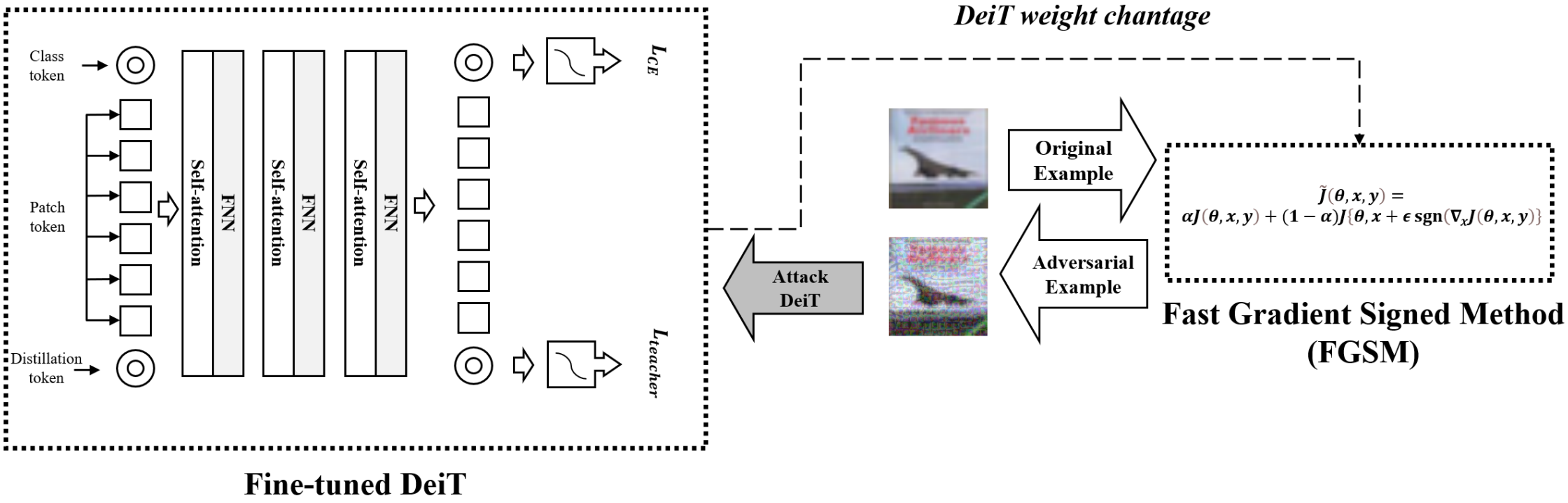
Related Work

- **PGD (Projected Gradient Descent)**
  - **FGSM 공격 기법을 여러 번 반복해 더욱 강력한 공격성능을 도출**
  - **$L_\infty - \mathrm{norm}$ 기반의 공격기법으로 FGSM을 기반으로 한 응용 공격**
  - **FGSM 공격의 반복을 step size로 조정하여 효율적인 공격 가능**



Method & Adversarial Attack

- **DeiT Vulnerability Verification (White Box Attack) – Previous Research**



Method & Adversarial Attack

① DeiT 자체 취약점 검증 - PGD 공격기법 추가 검증 (White Box Attack)

② DeiT의 일부 모델 weight를 이용한 DeiT 취약점 검증 (Semi-White Box Attack)

③ 일반 Deep-learning 모델 weight를 이용한 DeiT 취약점 검증 (Black Box Attack)

Method & Adversarial Attack

- **Adversarial Attack 환경**

  ① **White Box Environment**
  공격자가 타깃모델의 모든 정보를 알고 있다고 가정 (모델 구조, Hyper-parameter 등)

  ② **Gray Box Environment**
  공격자가 타깃모델의 정보를 어느정도 알고 있다고 가정 (대략적인 모델 구조)

  ③ **Black Box Environment**
  공격자가 타깃모델의 정보를 알고 있지 않다고 가정

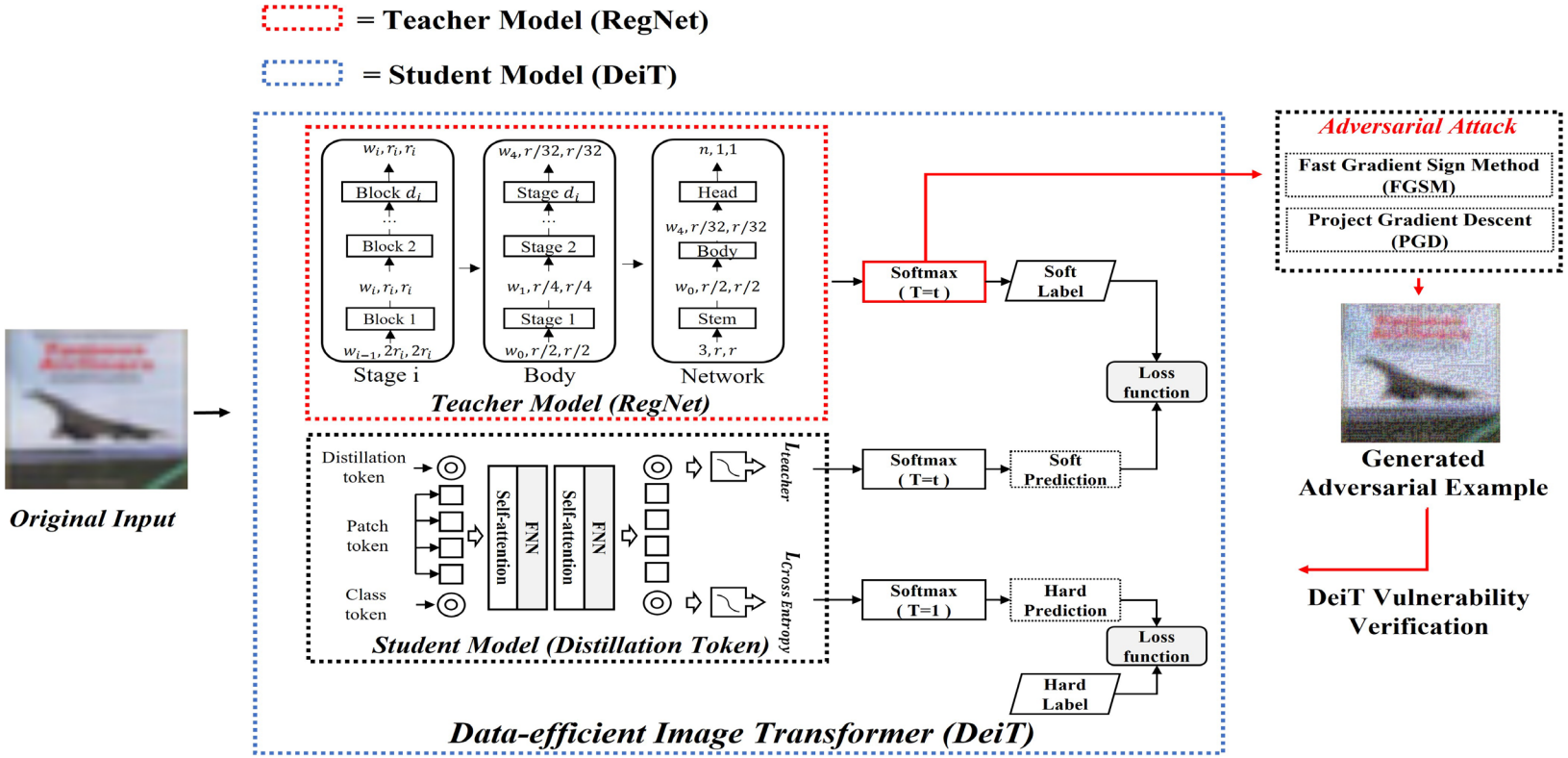
- **기존 적대적 공격 환경 ➔ 한번의 학습과정만 거치는 모델에서만 적용된 환경**

  ∴ **두번의 학습과정을 거치는 모델 ➔ 학습에 사용되는 일부 모델만 완벽히 파악하는 경우 환경정의 X**

  ➔ *일부모델의 정보만 파악한 경우 Semi-White Box Environment로 정의*

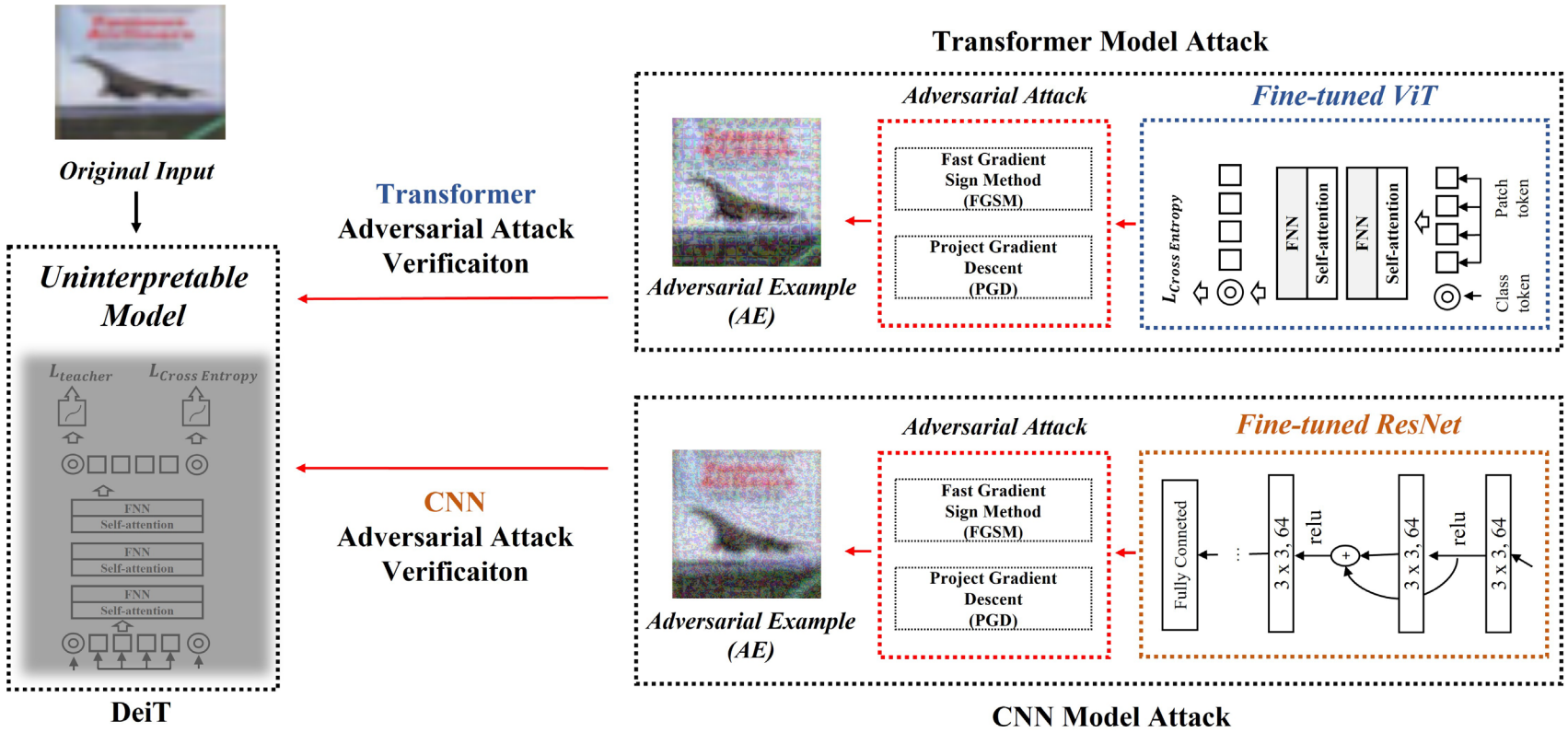
Method & Adversarial Attack

- **Teacher Model Vulnerability Verification Overview of DeiT (Semi-White Box Attack)**



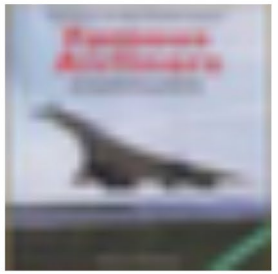
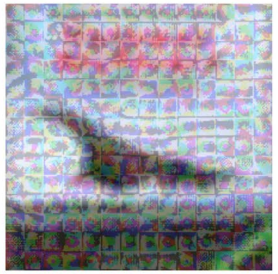
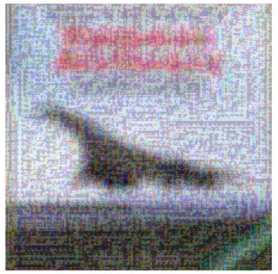
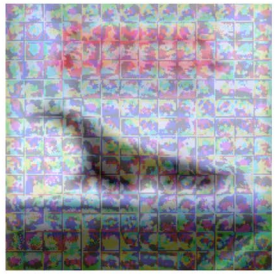
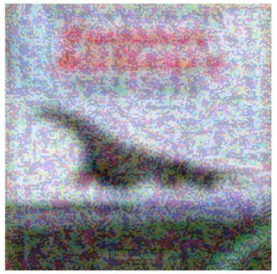
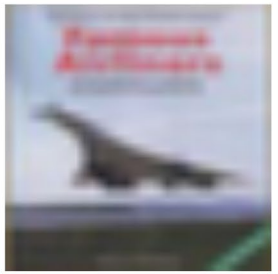
Method & Adversarial Attack

- **Black Box Attacks Verification Overview of DeiT**



Method & Adversarial Attack

- **Generated Adversarial Examples**

Attack	Un-Attacked	WhiteBox		BlackBox	
Target	-	DeiT	Teacher Model	Student Model	Teacher Model
FGSM					
	(a) Original	(b) DeiT	(c) RegNet	(d) ViT	(e) ResNet
PGD					
	(f) Original	(g) DeiT	(h) RegNet	(i) ViT	(j) ResNet

Experiments & Discussion

- **White Box Attacks of DeiT**

AEs	Target	Attack Method	Un-Attacked Accuracy	Attacked Accuracy	Accuracy Decrease Rate
DeiT	DeiT	FGSMIn-Pyo et al (2023)	0.9399	0.1050	**0.8349**
		PGD		0.0000	**0.9399**

- **FGSM, PGD공격 모두 DeiT에 치명적인 공격성능을 보임**

Experiments & Discussion

- **Semi-White Box Attacks of DeiT**

AEs	Target	Attack Method	Un-Attacked Accuracy	Attacked Accuracy	Accuracy Decrease Rate
RegNetY[1]	RegNetY	FGSM	0.8666	0.0920	0.7746
		PGD		0.0000	0.8666
	DeiT	FGSM	0.9399	**0.4750**	**0.4649**
		PGD		**0.2840**	**0.6559**

- **DeiT의 teacher model (RegNetY) 가중치 추출을 통한 DeiT 공격**

  ➔ **기존 white box 보다 치명적인 공격성능 X**
  하지만, teacher model만을 통해 충분히 DeiT에 치명적인 공격성능 도출

Experiments & Discussion

- **Black Box Attacks of DeiT (Different Learning Conditions)**

AEs	Target	Learning Conditions of AEs	Learning Conditions of Target	Attack Method	Un-Attacked Accuracy	Attacked Accuracy	Accuracy Decrease Rate
ViT	ViT	RMSprop (Step LR, 3e-5)	RMSprop (Step LR, 3e-5)	FGSM	0.9764	0.2650	0.7114
				PGD		0.0000	0.9764
ResNet-50	ResNet-50			FGSM	0.9249	0.1760	0.7489
				PGD		0.0000	0.9249
ViT	DeiT		Adam (Step LR, 5e-5)	FGSM	0.9399	**0.2350**	**0.7049**
				PGD		**0.2810**	**0.6589**
ResNet-50				FGSM		**0.4040**	**0.5359**
				PGD		**0.6100**	**0.3299**

- **Image Classification에서 대표적으로 쓰이는 2가지 모델의 weight를 사용한 공격 (ViT, ResNet)**

- **공격을 위한 모델 : DeiT와의 학습조건을 다르게 하여 DeiT의 정보를 모른다고 가정함**
- **ViT(Transformer 기반 모델) 공격 : 2가지 공격기법에서 모두 65%가 넘은 정확도 하락 유도**
- **ResNet(CNN 기반 모델) 공격 : FGSM공격에서 53%, PGD공격에서 32%로 DeiT에 치명적임을 확인**

Conclusion

Computer Vision➔ Transformer 기반 모델이 강세

∴ transformer모델을 활용한 다양한 모델이 출시되고 있으나, 해당모델에 대한 보안 취약점은 간과

CNN모델과 Transformer모델을 모두 사용하는 지식증류 모델(DeiT)의 보안 취약점을 분석

실험을 통해 일반적인 white box에서 취약점이 존재할 뿐 아니라,

학습에 사용되는 2개의 모델 중 1개의 모델만 공격자가 취득하더라도 치명적인 공격이 가능함을 제안

또한, DeiT는 CNN의 locality와 Transformer의 globality를 모두 반영하기 때문에,
공격자 입장에서 더 수월한 black box 공격이 가능함을 제안

추후, DeiT의 추가적인 취약점을 해결하여 adversarial robustness를 강화한 연구가 진행되기를 기대함