Adversarial Transferability and Beyond

Philipp Benz
https://phibenz.github.io

Chaoning Zhang
https://chaoningzhang.github.io

Author Introduction

Philipp Benz
https://phibenz.github.io
phibenz@gmail.com

Chaoning Zhang
https://chaoningzhang.github.io
chaoningzhang1990@gmail.com

We are Ph.D. students from the **Robotics and Computer Vision (RCV)** lab at **KAIST** in **South Korea**

Adversarial machine learning (AML) is our main research field.

We are **always** looking for **research collaborators!**

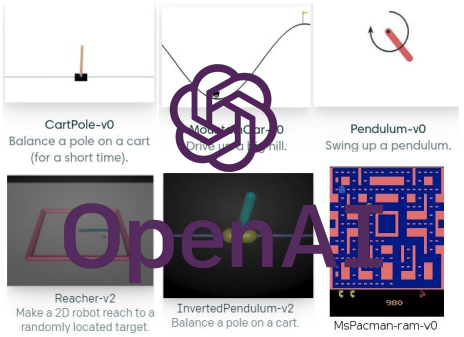
Author Introduction

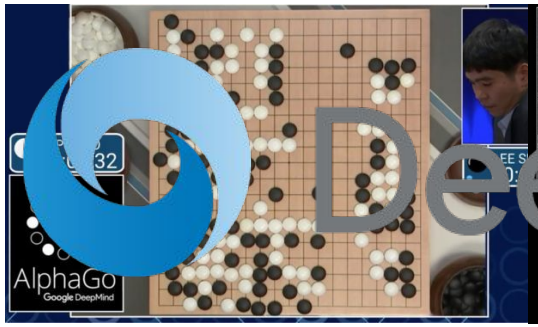
Selected recent works on AML (2020-2021):

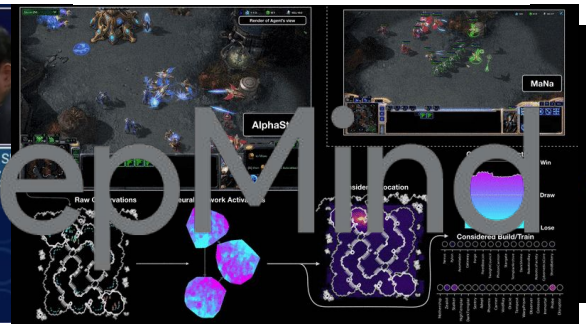
1. Universal Adversarial Training with Class-Wise Perturbations; **ICME 2021**
2. A Survey on Universal Adversarial Attack, **IJCAI 2021**
3. Universal Adversarial Perturbations Through the Lens of Deep Steganography: Towards a Fourier Perspective; **AAAI 2021**
4. Revisiting Batch Normalization for Improving Corruption Robustness; **WACV 2021**
5. UDH: Universal Deep Hiding for Steganography, Watermarking, and Light Field Messaging; **NeurIPS 2020**
6. Understanding Adversarial Examples from the Mutual Influence of Images and Perturbations; **CVPR 2020**
7. Double Targeted Universal Adversarial Perturbations; **ACCV, 2020**
8. CD-UAP: Class Discriminative Universal Adversarial Perturbations; **AAAI 2020**
9. Robustness Comparison of Vision Transformer and MLP-Mixer to CNNs; **CVPR-W 2021 (Outstanding paper award)**
10. The Triangular Trade-off Between Accuracy, Robustness, and Fairness; **CVPR-W 2021**
11. Backpropagating Smoothly Improves Transferability of Adversarial Examples; **CVPR-W 2021**
12. Is FGSM Optimal or Necessary for $L\infty$ Adversarial Attack? **CVPR-W 2021**
13. Stochastic Depth Boosts Transferability of Non-Targeted and Targeted Adversarial Attacks; **ICLR-W 2021**
14. On Strength and Transferability of Adversarial Examples: Stronger Attack Transfers Better; **ICLR-W 2021**
15. Towards Data-free Universal Adversarial Perturbations with Artificial Jigsaw Images; **ICLR-W 2021**
16. Batch Normalization Increases Adversarial Vulnerability and Decreases Adversarial Transferability: A feature perspective; **ICLR-W 2021**
17. Towards Simple Yet Effective Transferable Targeted Adversarial Attacks; **ICLR-W 2021**
18. Robustness May Be at Odds with Fairness: An Empirical Study on Class-wise Accuracy; **NeurIPS-W 2021**
19. Data from Model; **CVPR-W 2020**
20. Universal Adversarial Perturbations are Not Bugs, They are Features; **CVPR-W 2020**

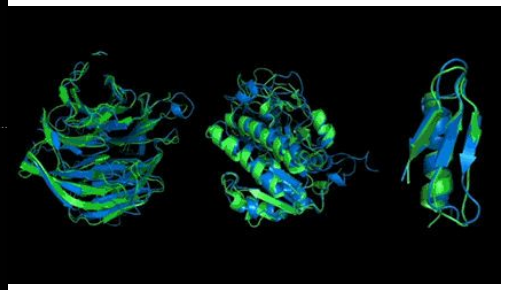
Introduction

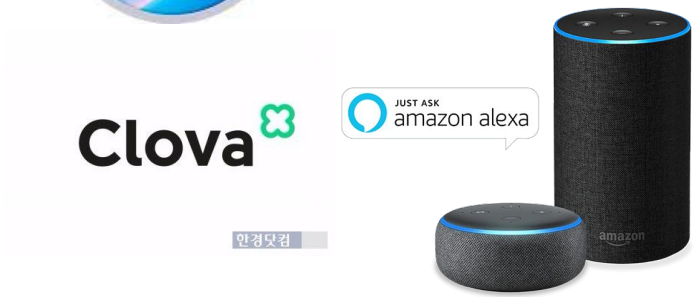
Deep learning is Awesome





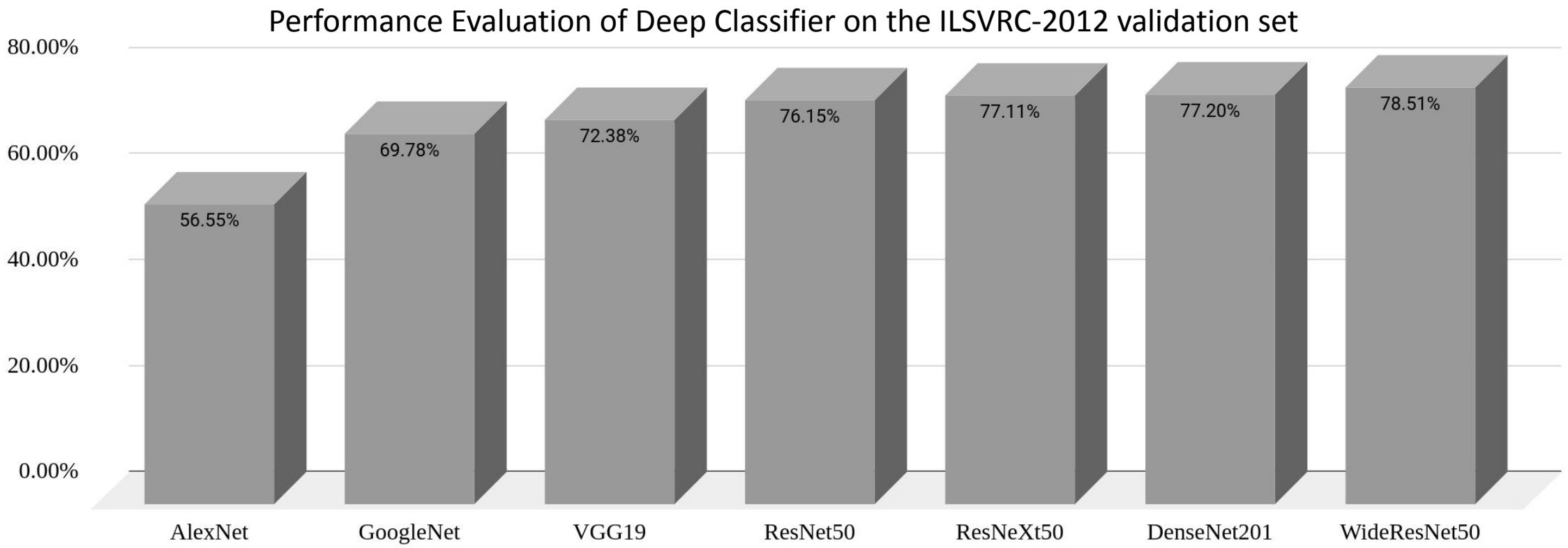






Deep Classifiers

Classification is one of the most fundamental tasks in machine learning.
After the advent of Deep Learning, Deep Classifiers dominate the field of image classification.



One fundamental concern about Deep Classifiers is their robustness.

Intriguing Adversarial Examples

Deep Neural Networks are sensitive to small perturbations in the image, which are specially crafted to deteriorate performance and to be mostly imperceptible for human observers.

One of the earliest and simplest adversarial attack methods is the **Fast Gradient Sign Method (FGSM)** [1]

Results on Googlenet

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

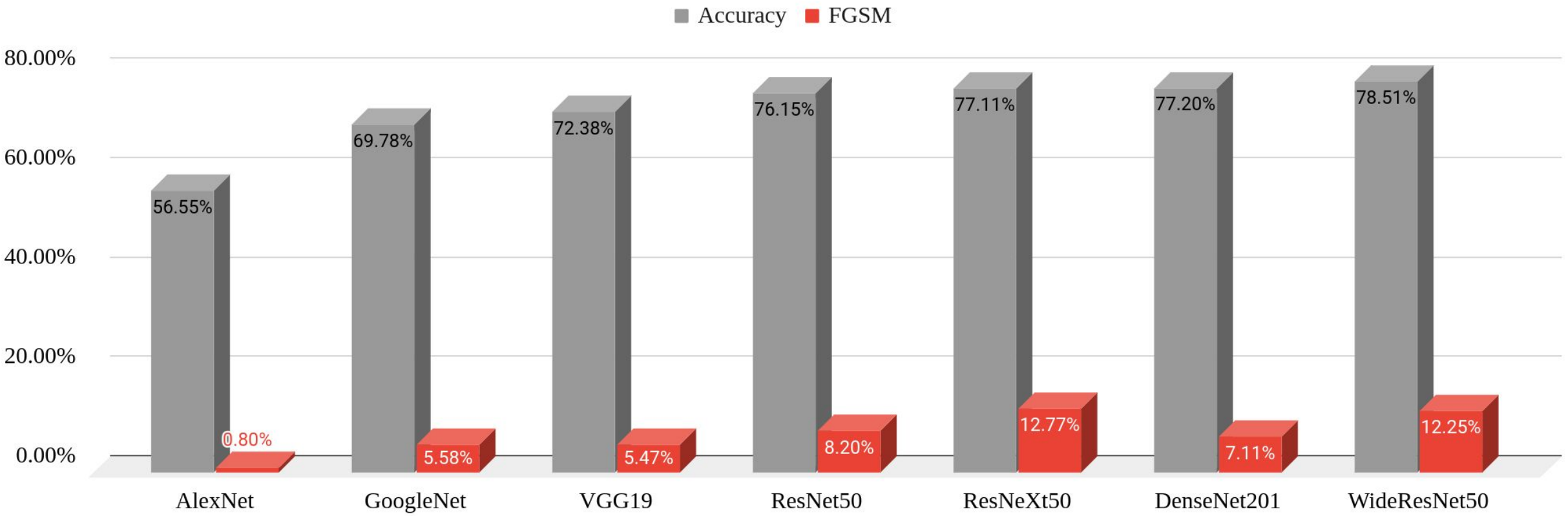
$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

[1] Explaining and Harnessing Adversarial Examples; Goodfellow, Shlens, Szegedy; ICLR 2015
[2] Intriguing properties of neural networks; Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, Fergus; arXiv 2013

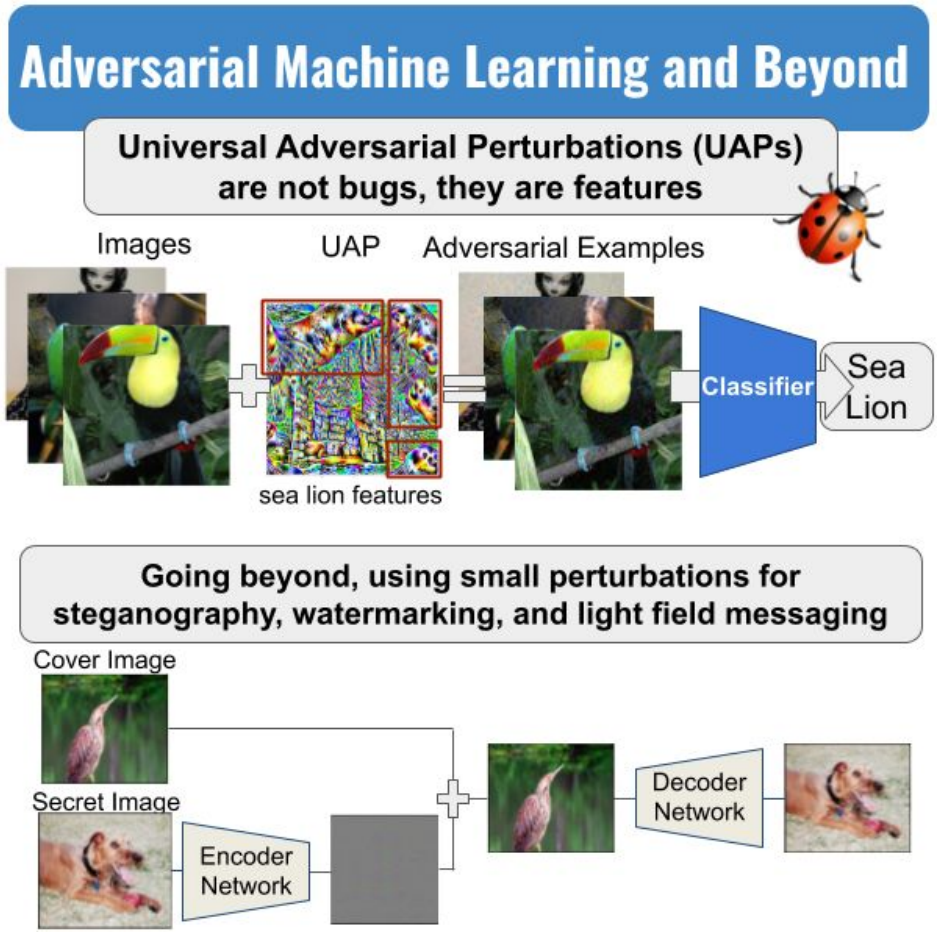
Performance Degradation with the simple FGSM attack



Allowable perturbation magnitude ε = 8/255, for images in range [0,1]

Adversarial Machine Learning and Beyond (Recap)

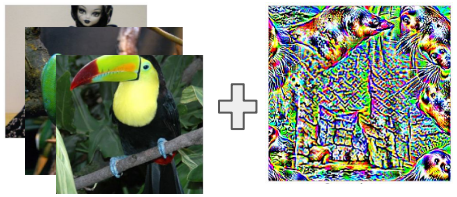
https://www.youtube.com/watch?v=ylEE1HtGNJc



Results presented in the proposal

An Alternative Perspective on UAPs

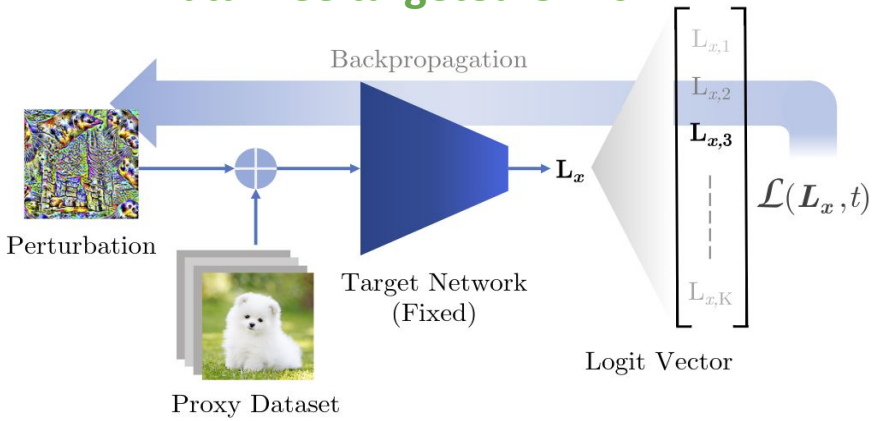
Previous Belief



New Perspective

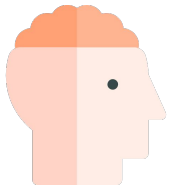
UAPs are features and images behave like noise to them

Data-free targeted UAPs



Simple, Fast and Effective

A Frequency Understanding on Targeted UAPs



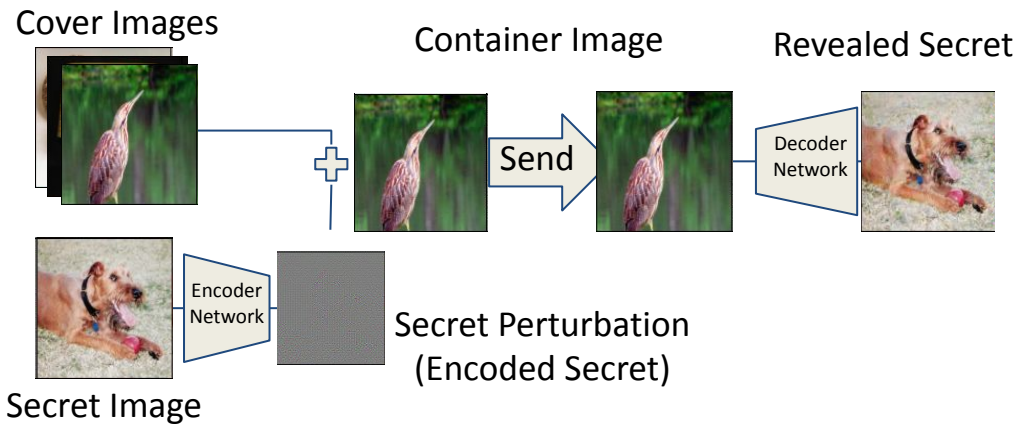
DNN

Sensitive to
low-frequency content

Sensitive to
high-frequency content

In contrary to human, DNN is more sensitive to high-frequency content

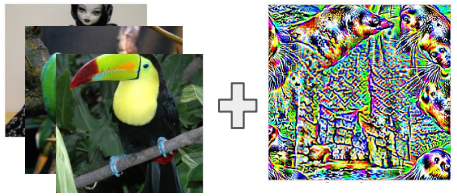
Towards Universal Deep Hiding



Steganography, Watermarking, and Light field messaging

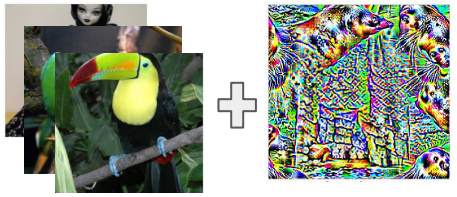
An Alternative Perspective on UAPs

Previous Belief



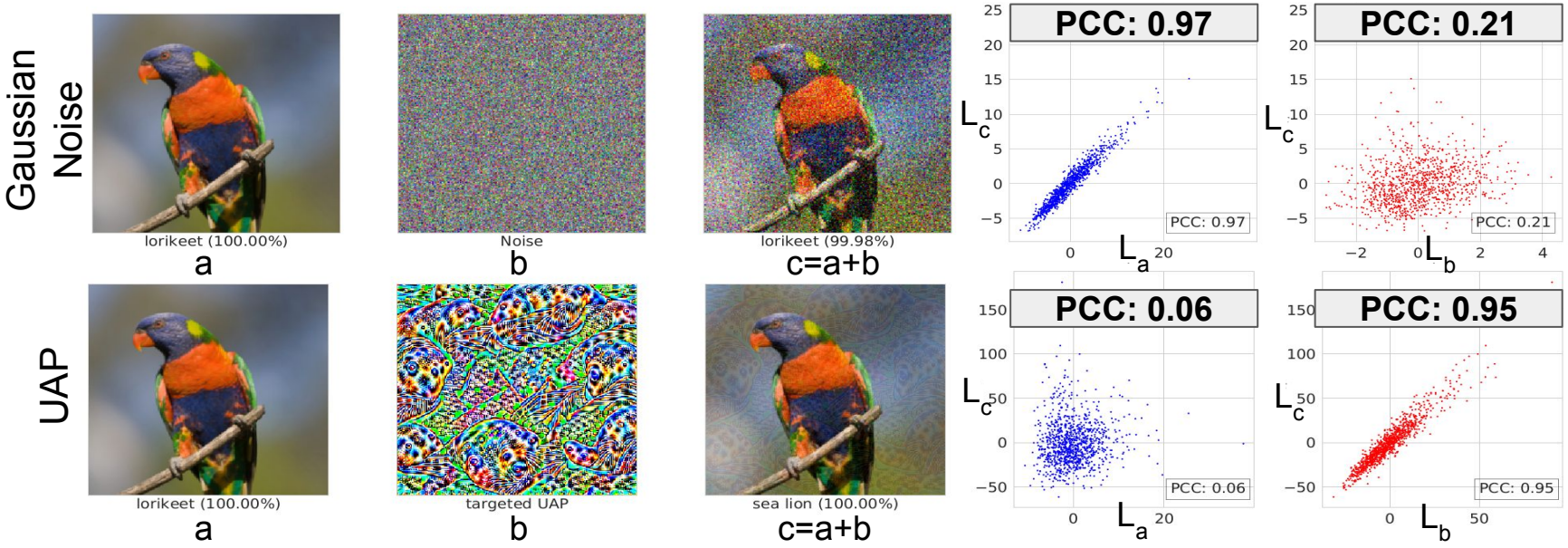
Features + Noise

Our Perspective



Noise + Features

Treat the DNN logits as a vector for feature representation and use them to analyze the mutual influence of two independent inputs based on the Pearson correlation coefficient (PCC)



$$\mathrm{PCC}_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

"Universal perturbations contain dominant features, and images behave like noise to them"

PCC-Analysis result for one sample image `lorikeet'. Three scenarios of input combinations are considered: 1: image + noise; 2: image + targeted UAP; 3: image + targeted image-dependant AE. The columns show input a, input b, input c=a+b, logit vector analysis of L_c over L_a and vector analysis of L_c over L_b

Data-free Targeted UAPs

If images behave like noise to the features in UAPs, we can leverage a proxy dataset to craft UAPs

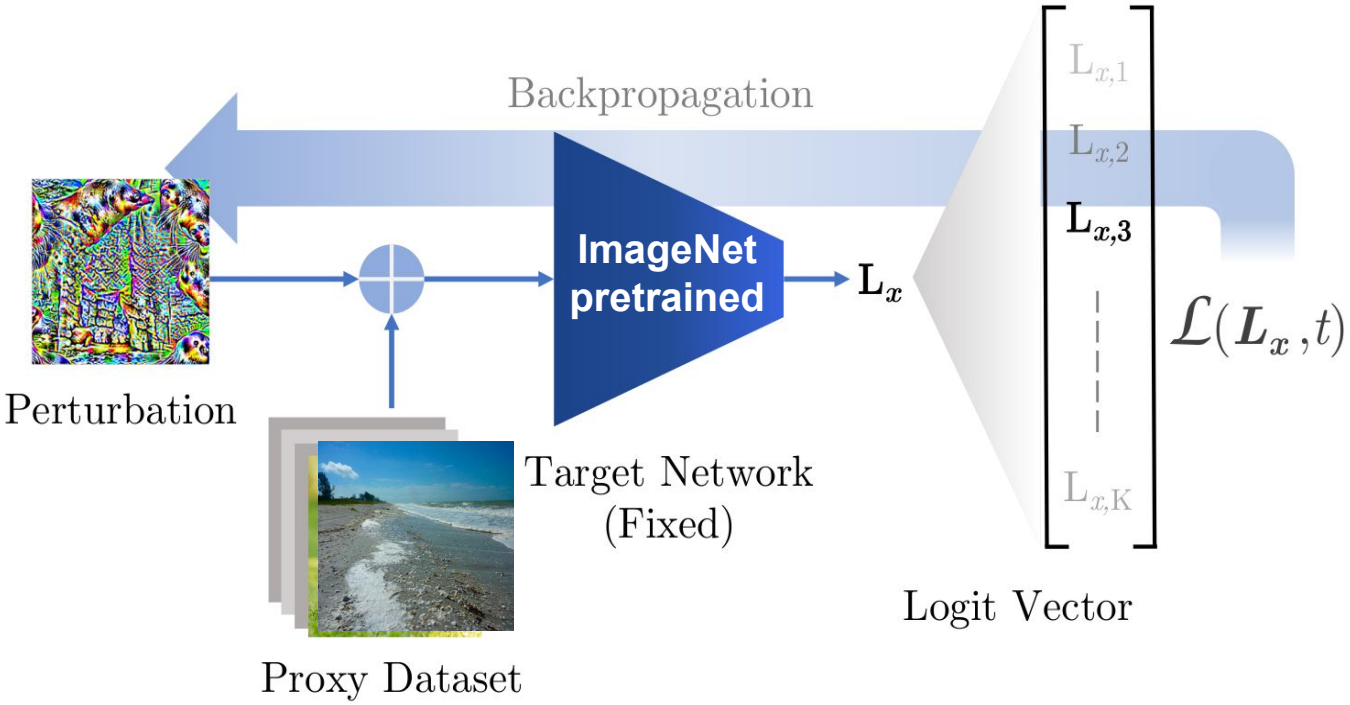


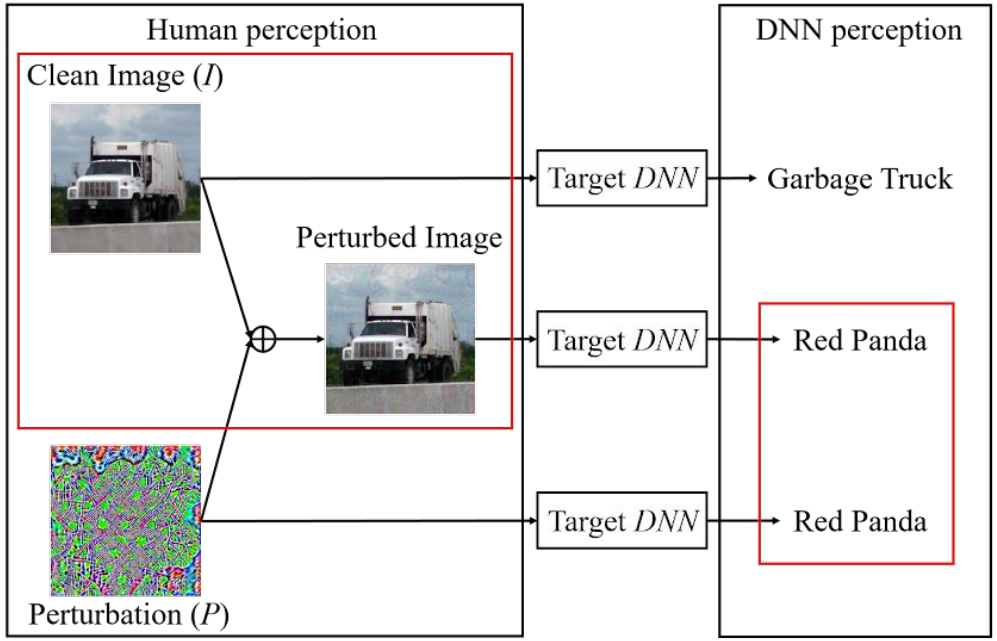
Table 1: Comparison to other methods. The results are divided in universal attacks with access to the original ImageNet training data (upper) and data-free methods (lower). The metric is reported in the non-targeted fooling ratio (%)

Method	AlexNet[1]	GoogleNet	VGG16	VGG19	ResNet152
UAP [14]	93.3	78.9	78.3	77.8	84.0
GAP [19]	-	82.7	83.7	80.1	-
Ours(ImageNet [11])	**96.17**	**88.94**	**94.30**	**94.98**	**90.08**
FFF [18]	80.92	56.44	47.10	43.62	-
AAA [21]	89.04	75.28	71.59	72.84	60.72
GD-UAP [17]	87.02	71.44	63.08	64.67	37.3
Ours (COCO [12])	89.9	**76.8**	**92.2**	**91.6**	**79.9**
Ours (VOC [5])	89.9	76.7	**92.2**	90.5	79.1
Ours (Places365 [29])	**90.0**	76.4	92.1	91.5	78.0

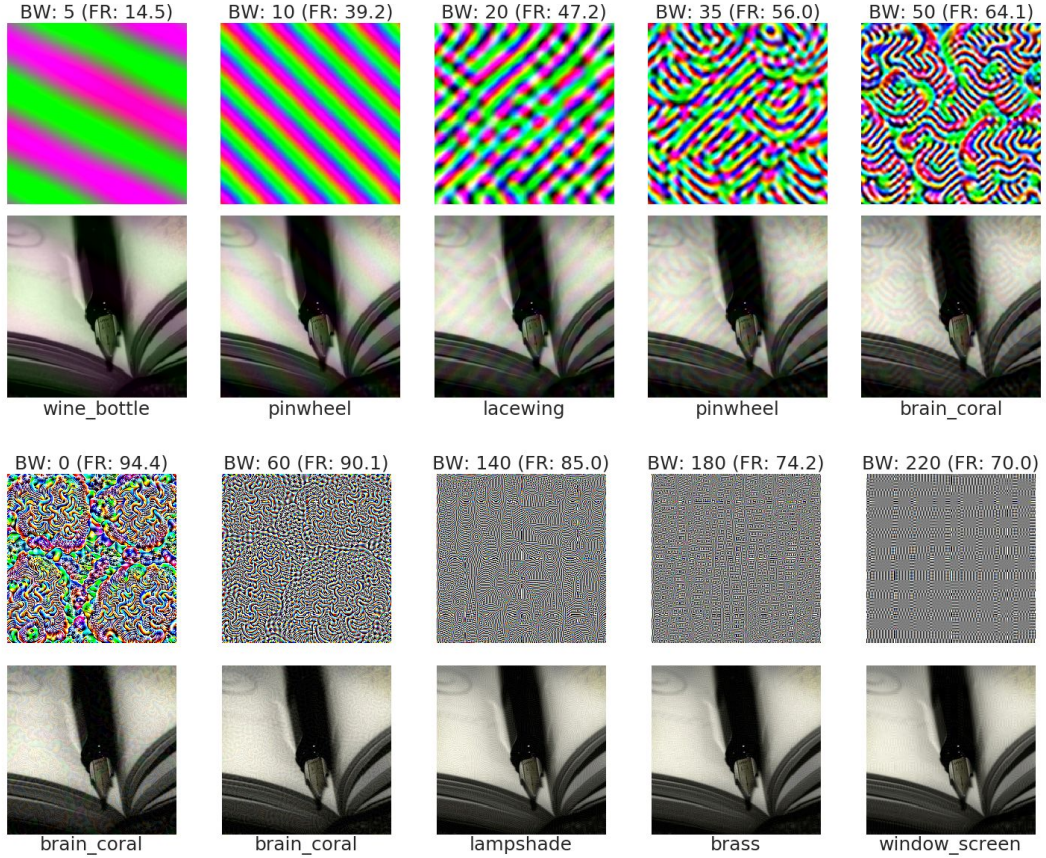
Fast

Our algorithm takes ~2 minutes vs. ~2 hours for the vanilla UAP algorithm

A Frequency Understanding on Targeted UAPs



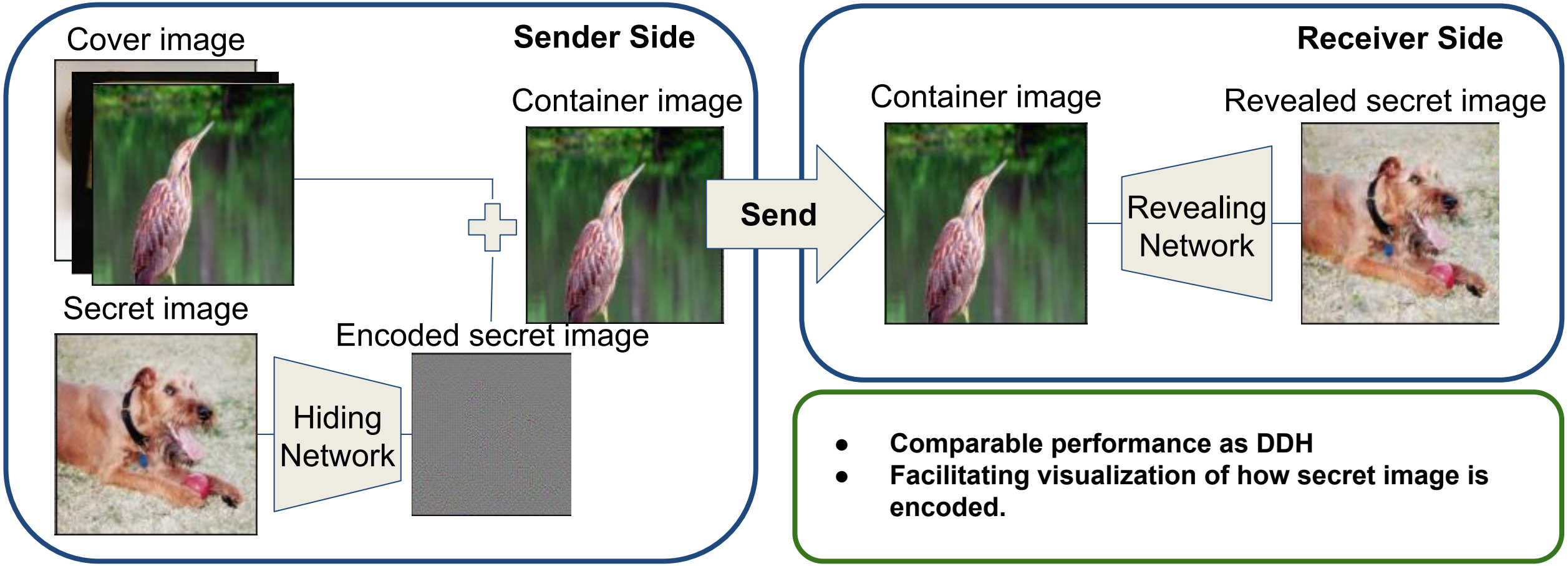
(Targeted) Universal Adversarial Attack



Why does such small universal perturbation dominate images?
DNN is highly sensitive to **high-frequency** patterns

Universal Deep Hiding (UDH)

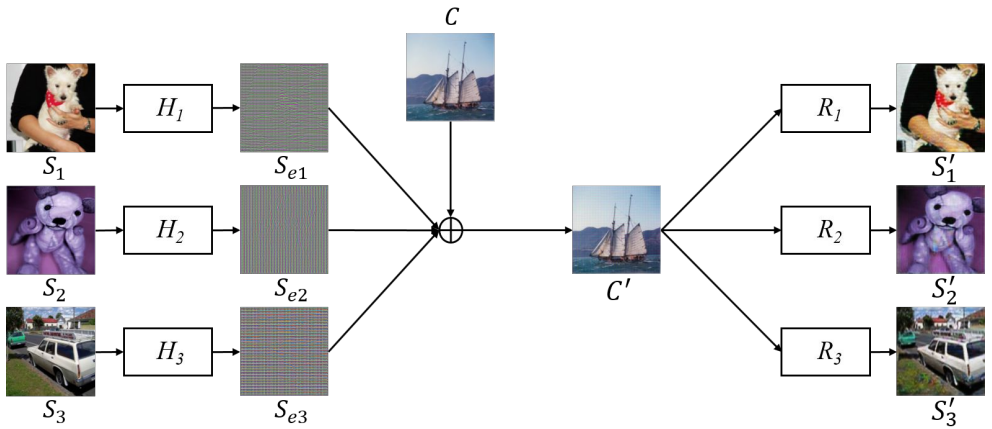
A novel Deep Hiding meta-architecture to hide a secret image in a **cover-agnostic** manner.



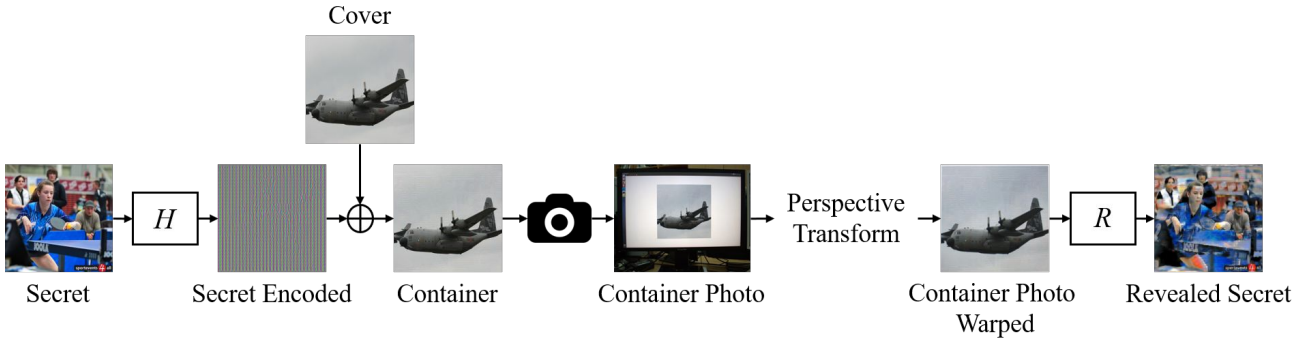
- **Comparable performance as DDH**
- **Facilitating visualization of how secret image is encoded.**

A secret image is fed to the hiding network to yield an encoded secret image, which can be added to a random cover image to form a container image. The revealing network then retrieves the secret image from the container.

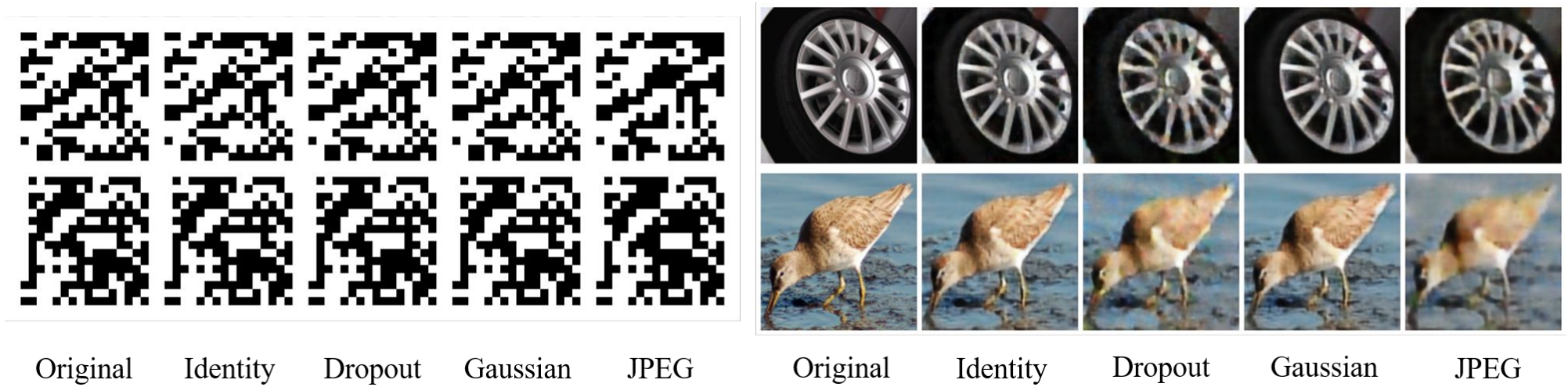
Universal Deep Hiding (UDH)



Steganography



Light field messaging



Watermarking

Adversarial Attacks

Adversarial attack: Definition and Goal

Finding a **Small** perturbation that **misclassifies** a sample

$$C(x+\delta) \neq C(x)$$

Misclassification

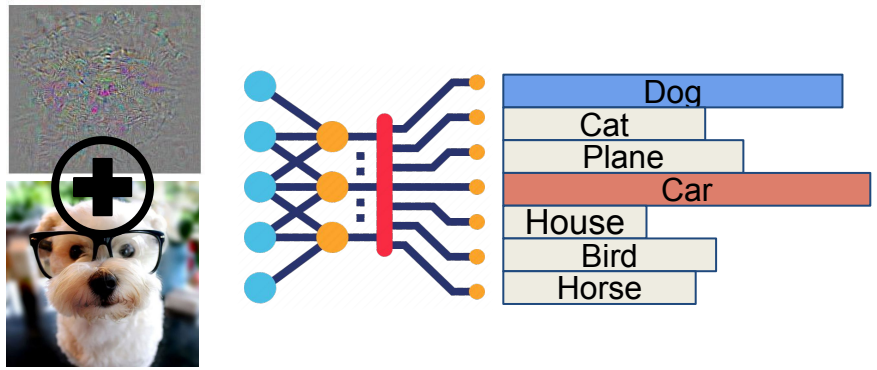
subject to $D(x, x+\delta) \leq \epsilon$

The perturbation is smaller than magnitude ϵ
D is some distance metric *L1*, *L2*, *Linf*

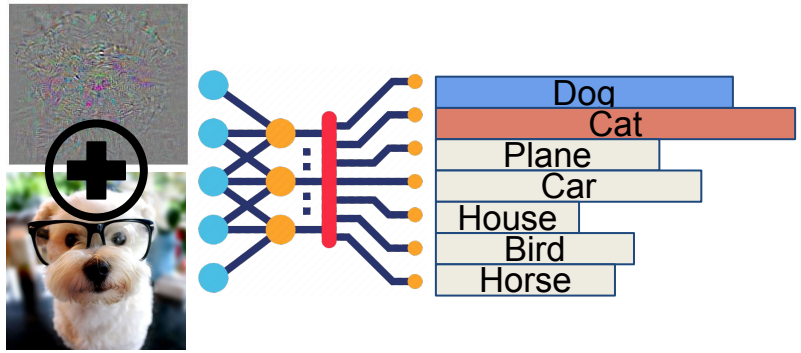
$$x+\delta \in [0,1]$$

Obey image range

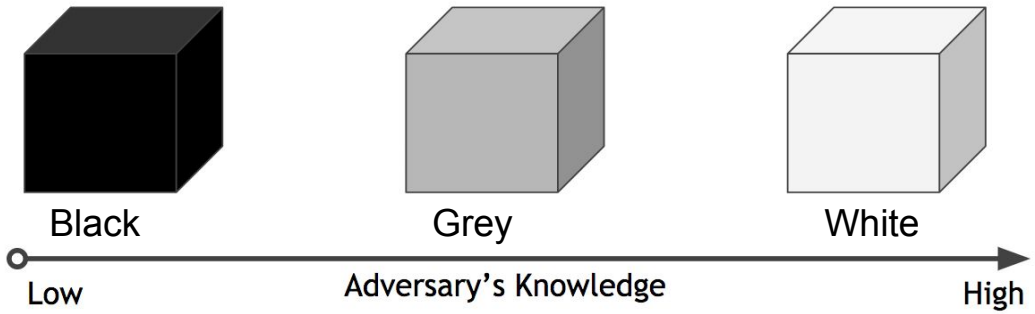
Untargeted scenario: Misclassify to **any other** class



Targeted scenario: Misclassify to a **predefined target** class (Cat for instance)



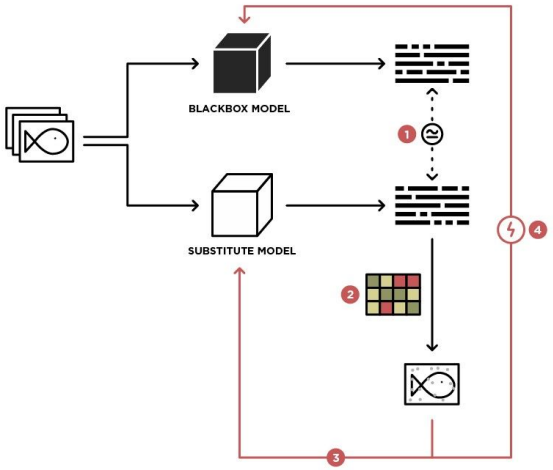
White-box Attacks

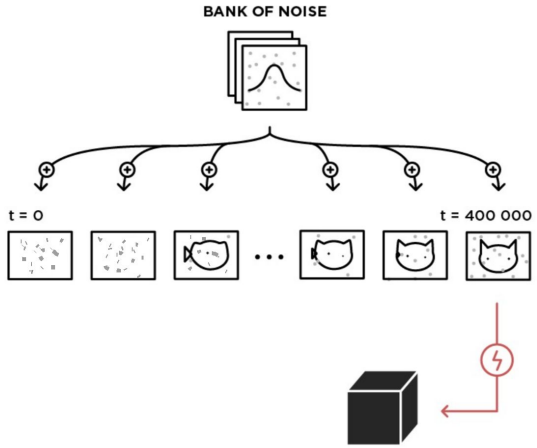


- In white-box attack, an adversary has total knowledge about the model used for classification (e.g., type of neural network along with number of layers).
- The attacker has information about the algorithm used in training (e.g., gradient-descent optimization) and can access the training data distribution. Also knows the parameters of the fully trained model architecture.
- The adversary utilizes available information to identify the feature space where the model may be vulnerable, i.e, for which the model has a high error rate. Then the adversary can make full use of the network information to carefully craft adversarial examples.

Picture Source: Team Panda, Class 1: Intro to Adversarial Machine Learning, <https://secml.github.io/class1/>
Reference: Chakraborty, Anirban et al. "Adversarial Attacks and Defences: A Survey." *ArXiv* abs/1810.00069 (2018): n. pag.

Black-Box Attacks





- Black-Box attack, assumes no knowledge about the model and uses information about the settings or past inputs to analyse the vulnerability of the model. For example, the adversary exploits a model by providing a series of carefully crafted inputs and observing outputs.

- Based on whether an attacker needs to query the victim model, there are query-free (transfer-based) and query-based attacks.

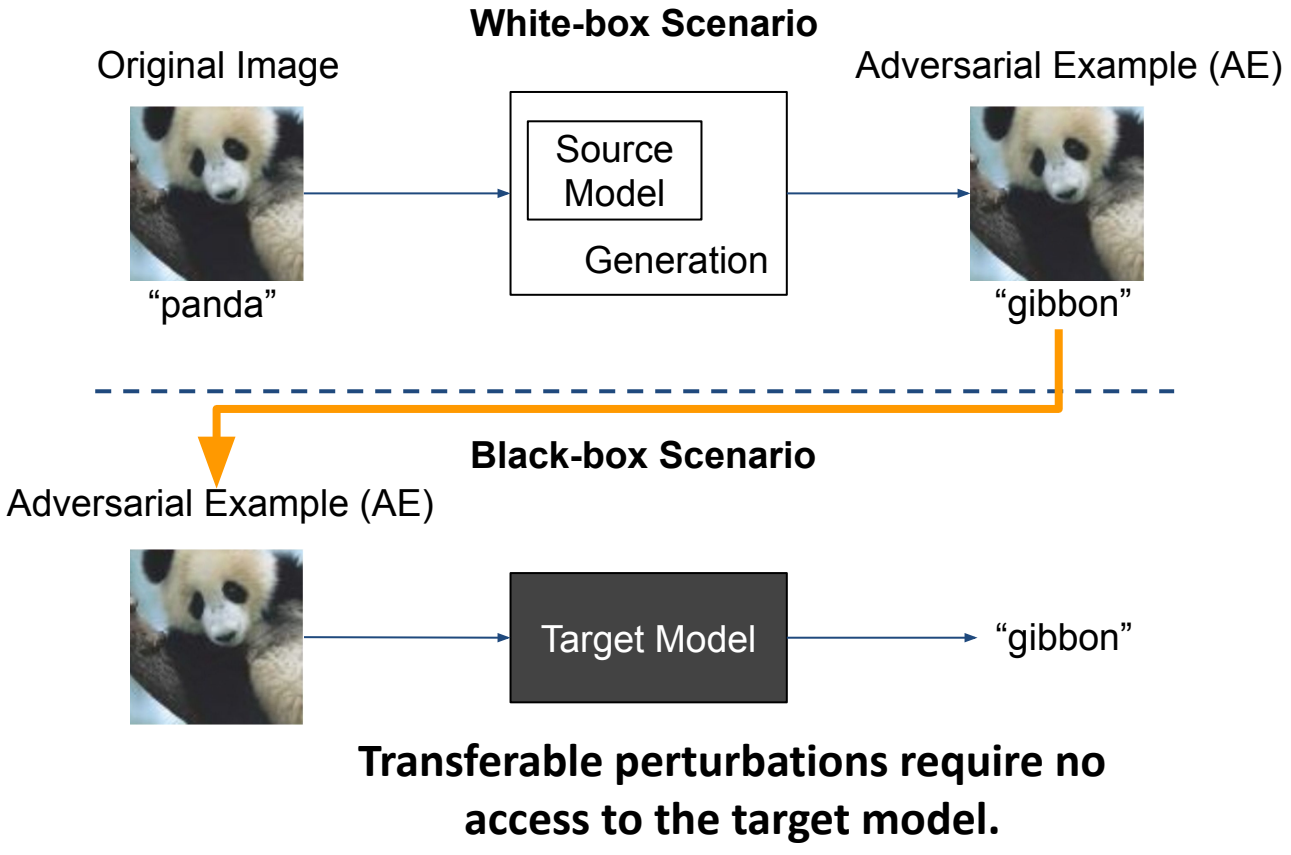
- *Transferability* is critical for Black-Box attacks where the victim model and the training data are not accessible. Attackers can train a substitute (source) model and then generate adversarial examples against substitute model. Then the victim model will be vulnerable to these adversarial examples due to transferability.

Reference: Yuan, Xiaoyong, et al. "Adversarial examples: Attacks and defenses for deep learning." *IEEE transactions on neural networks and learning systems* 30.9 (2019): 2805-2824.
Li, Huichen, et al. "QEBA: Query-Efficient Boundary-Based Blackbox Attack." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
Picture Source: Rey Reza Wiyatno, Tricking a Machine into Thinking You're Milla Jovovich, <https://medium.com/element-ai-research-lab/tricking-a-machine-into-thinking-youre-milla-jovovich-b19bf322d55c>

Transferability of Adversarial Perturbations

Transferable Adversarial Examples

Transferable Property (Model-agnostic) [1]
One perturbation generated on one network (source), transfers to another unseen one (target)



Transferable perturbations require no access to the target model.

[1] Adversarial machine learning at scale; Kurakin, Goodfellow, Bengio; ICLR 2017

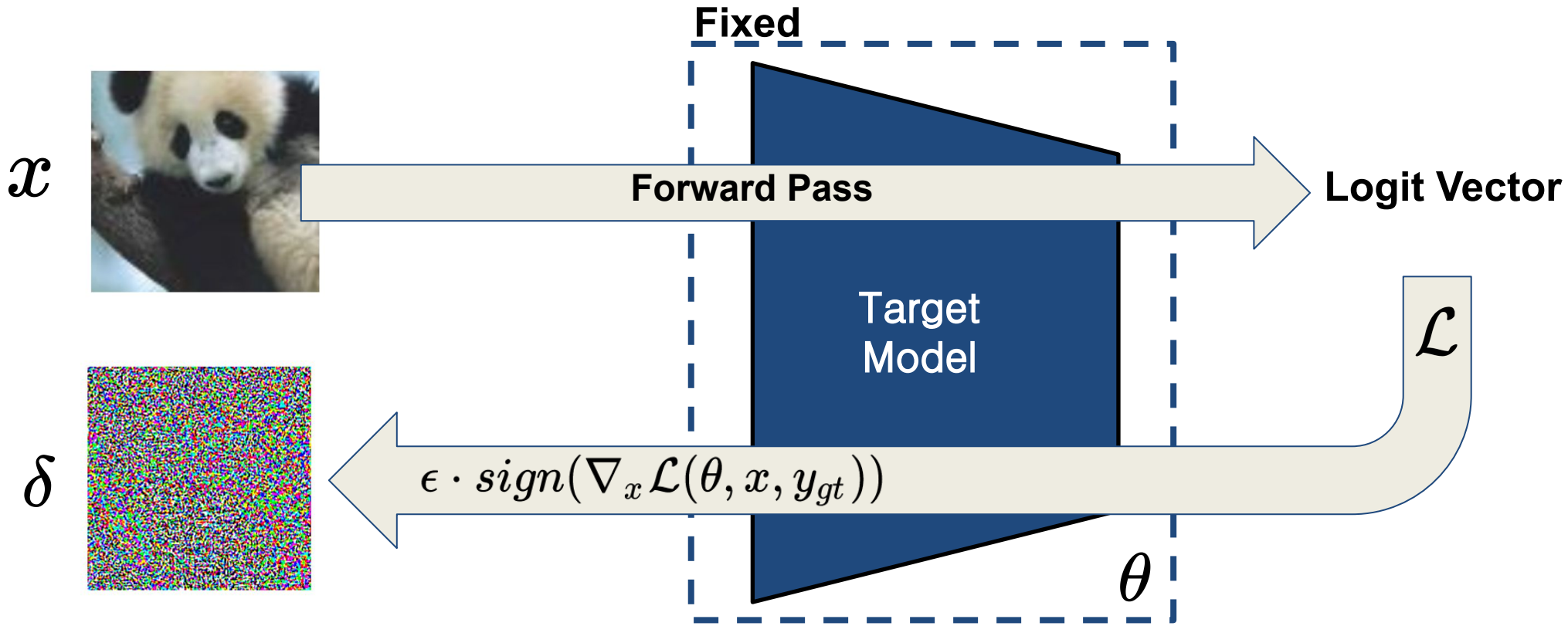
Background: FGSM attack

Fast Gradient Sign Method (FGSM) [1]
Stepping one step of step size ϵ into the gradient direction

$$X^{adv} = X + \epsilon sign(\nabla_X J(X, y_{true}))$$

Pro: Fast Con: Less effective $\mathcal{L} = CE$

ϵ: magnitude



[1] Explaining and Harnessing Adversarial Examples; ICLR 2015

Background: I-FGSM attack

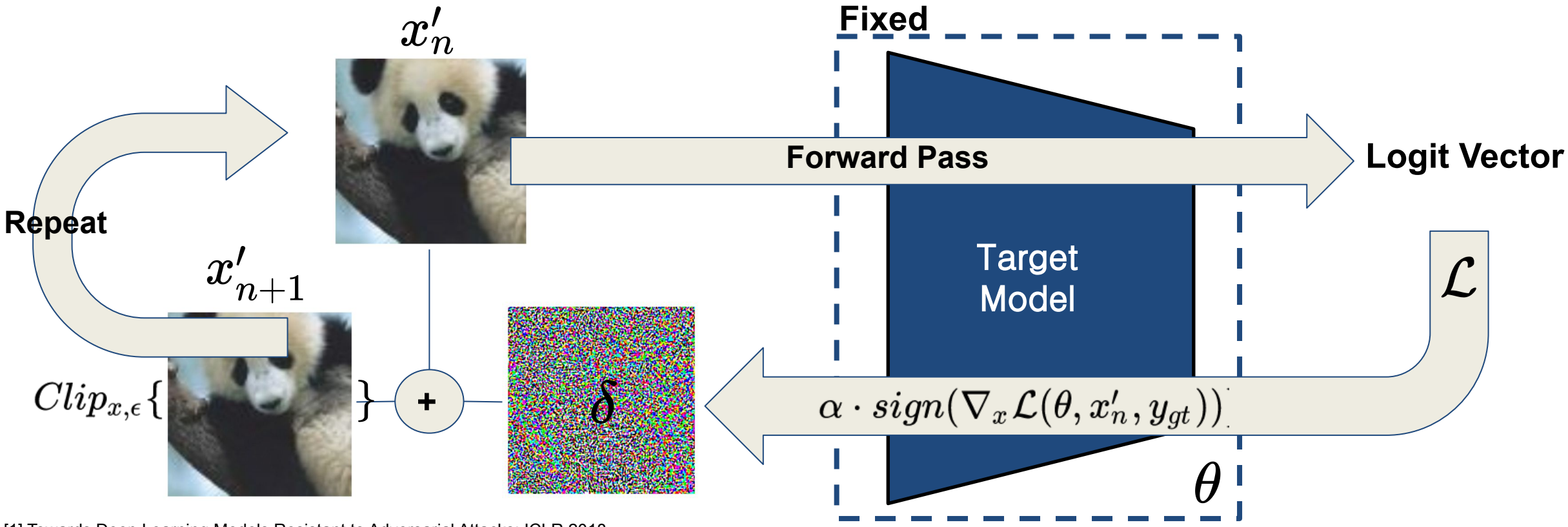
Iterative FGSM (I-FGSM)

Iteratively apply FGSM with step size α

$$X_0^{adv} = X, X_{t+1}^{adv} = X_t^{adv} + \alpha sign(\nabla_X J(X_t^{adv}, y))$$

Pro: Effective **Con: less transferable**

α: step size
α = ϵ/T, where T indicates total iterations



[1] Towards Deep Learning Models Resistant to Adversarial Attacks; ICLR 2018

Popular techniques to increase transferability: MI-FGSM

Momentum Iterative Fast Gradient Sign Method (MI-FGSM)

Use the momentum term to increase adversarial transferability

$$g_{t+1}^{adv} = \mu g_t^{adv} + \frac{\nabla_X L(X_t^{adv}, y)}{||\nabla_X L(X_t^{adv}, y)x||_1}$$

$$X_{t+1}^{adv} = X_t^{adv} + \alpha sign(\nabla_X J(X_t^{adv}, y))$$

μ: Decay Factor

	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-152	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	FGSM	72.3*	28.2	26.2	25.3	11.3	10.9	4.8
	I-FGSM	**100.0***	22.8	19.9	16.2	7.5	6.4	4.1
	MI-FGSM	**100.0***	**48.8**	**48.0**	**35.6**	**15.1**	**15.2**	**7.8**
Inc-v4	FGSM	32.7	61.0*	26.6	27.2	13.7	11.9	6.2
	I-FGSM	35.8	**99.9***	24.7	19.3	7.8	6.8	4.9
	MI-FGSM	**65.6**	**99.9***	**54.9**	**46.3**	**19.8**	**17.4**	**9.6**
IncRes-v2	FGSM	32.6	28.1	55.3*	25.8	13.1	12.1	7.5
	I-FGSM	37.8	20.8	**99.6***	22.8	8.9	7.8	5.8
	MI-FGSM	**69.8**	**62.1**	99.5*	**50.6**	**26.1**	**20.9**	**15.7**
Res-152	FGSM	35.0	28.2	27.5	72.9*	14.6	13.2	7.5
	I-FGSM	26.7	22.7	21.2	**98.6***	9.3	8.9	6.2
	MI-FGSM	**53.6**	**48.9**	**44.7**	98.5*	**22.1**	**21.7**	**12.9**

The success rates (%) of non-targeted adversarial attacks against seven models

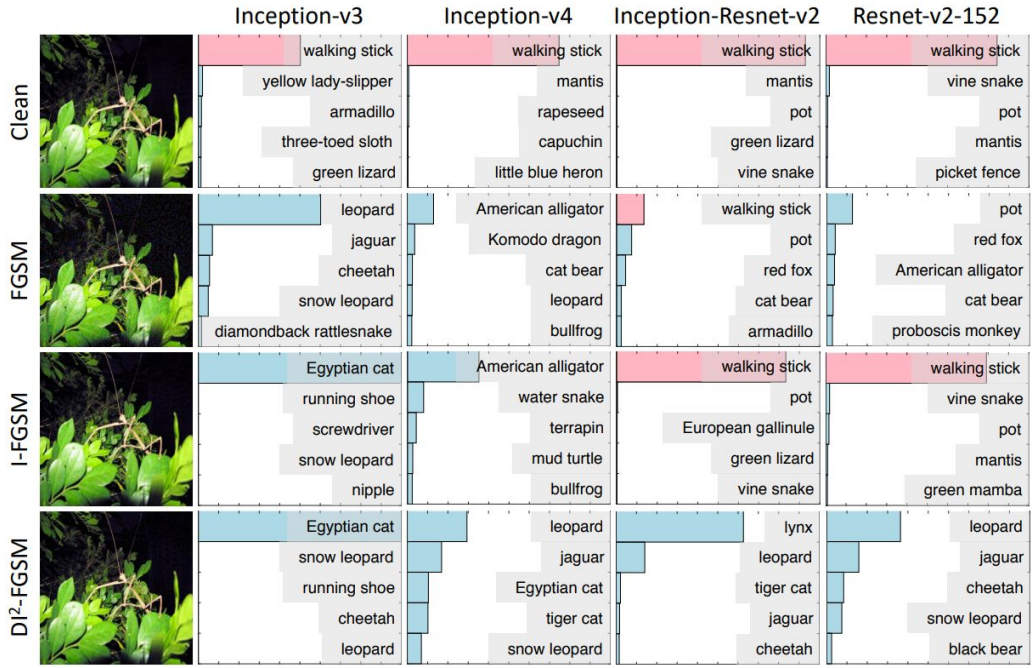
[1] Boosting Adversarial Attacks with Momentum; Dong, Liao, Pang, Su, Zhu, Hu, Li;CVPR 2018

Popular techniques to increase transferability: DI2-FGSM

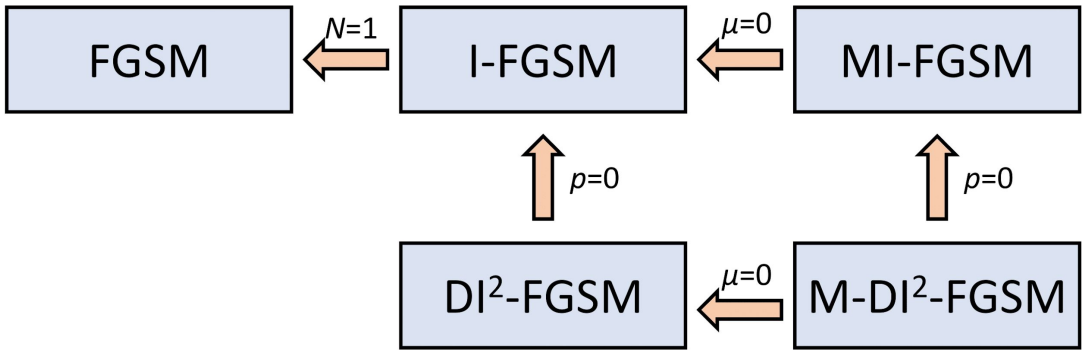
Diverse Inputs Iterative Fast Gradient Sign Method (DI^2-FGSM)

Applies random and differentiable transformations *Tr*, *e.g.* random resizing, random padding to the input images

$$X_{t+1}^{adv} = X_t^{adv} + \alpha sign(\nabla_X J(Tr(X_t^{adv}; p), y))$$



Comparison of success rates. The ground-truth "walking stick" is shown in pink. The adversarial examples are crafted on Inception-v3. DI^2-FGSM attacks the white-box model and all black-box models successfully.



Relationships between different attacks

μ: Decay Factor
N: # of Iterations
p: Probability of transformation

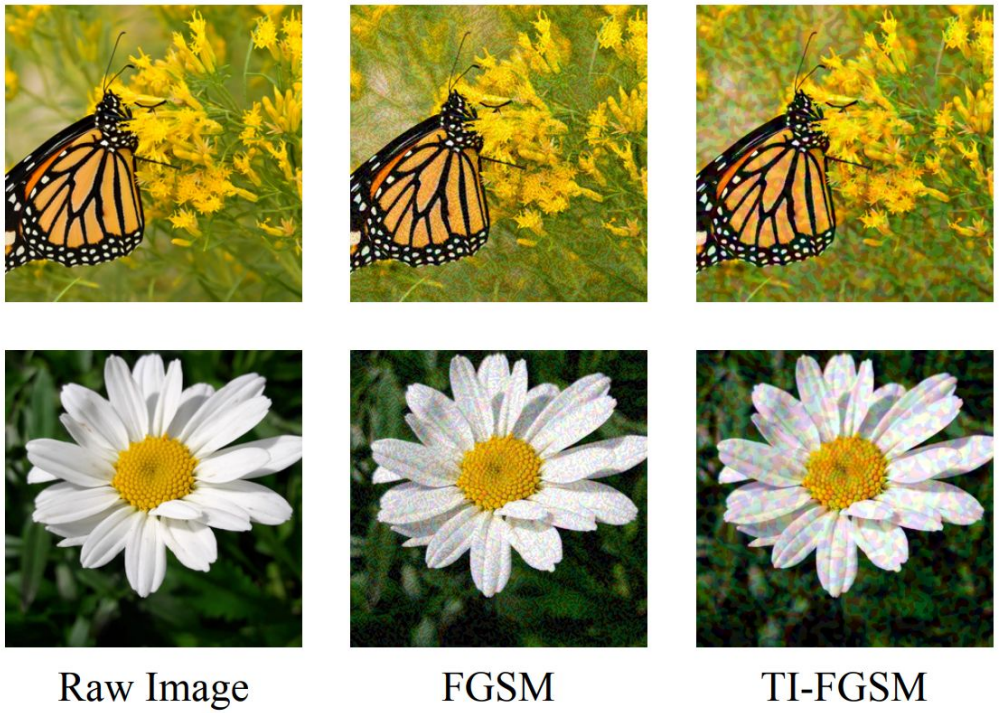
[1] Improving Transferability of Adversarial Examples with Input Diversity; Xie, Zhang, Zhou, Bai, Wang, Ren, Yuille; CVPR 2019

Popular techniques to increase transferability: TI-FGSM

Translation Invariant Fast Gradient Sign Method (TI-FGSM)

Optimizes a perturbation over an ensemble of translated images.
This method can be implemented by convolving the gradient at the untranslated image with a predefined kernel.

$$X_{t+1}^{adv} = X_t^{adv} + \alpha sign(W * \nabla_X J(X_t^{adv}, y))$$



Raw Image FGSM TI-FGSM

[1] Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks; Dong, Pang, Su, Zhu; CVPR 2019

On Strength and Transferability of Adversarial Examples: Stronger Attack Transfers Better

Chaoning Zhang*, Philipp Benz*, Adil Karjauv*, In So Kweon
Korea Advanced Institute of Science and Technology (KAIST)

RobustML workshop paper at ICLR 2021

Metric of Attack Strength (Transferability)

Conventional Attack Success Rate (ASR): Percentage of images being misclassified after attack

Drawback: Treating every misclassification equally.

GT class:
Swiss Greater mountain dog

No Attack

top 1

Weak attack

top 2

Stronger attack

top 5

Equally treated in the conventional ASR metric

Conventional ASR (ASR@1)

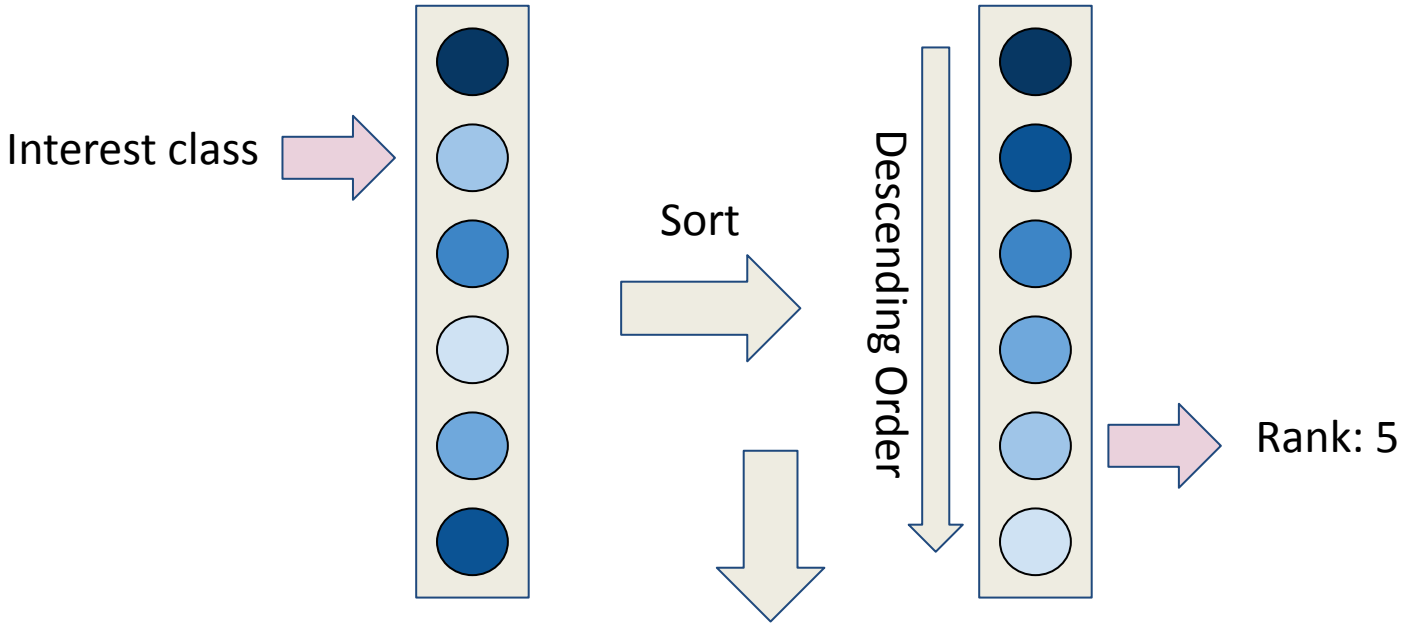
Proposed ASR@k

With the metric of ASR@k, a sample is successfully attacked if and only if the GT class is ranked higher than k.

Our Contributions

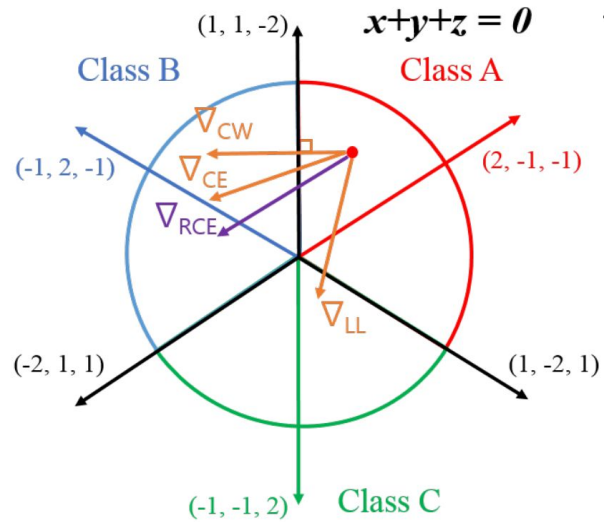
A unified simple metric to evaluate the ASR@k strength and/or transferability

Interest Class Rank (ICR)



A new loss for boosting semantic adversarial strength based on Geometry Perspective on loss design

$$RCE(X_t^{adv}, y_{gt}) = CE(X_t^{adv}, y_{gt}) - \frac{1}{K}\sum_{k=1}^{K} CE(X_t^{adv}, y_k)$$

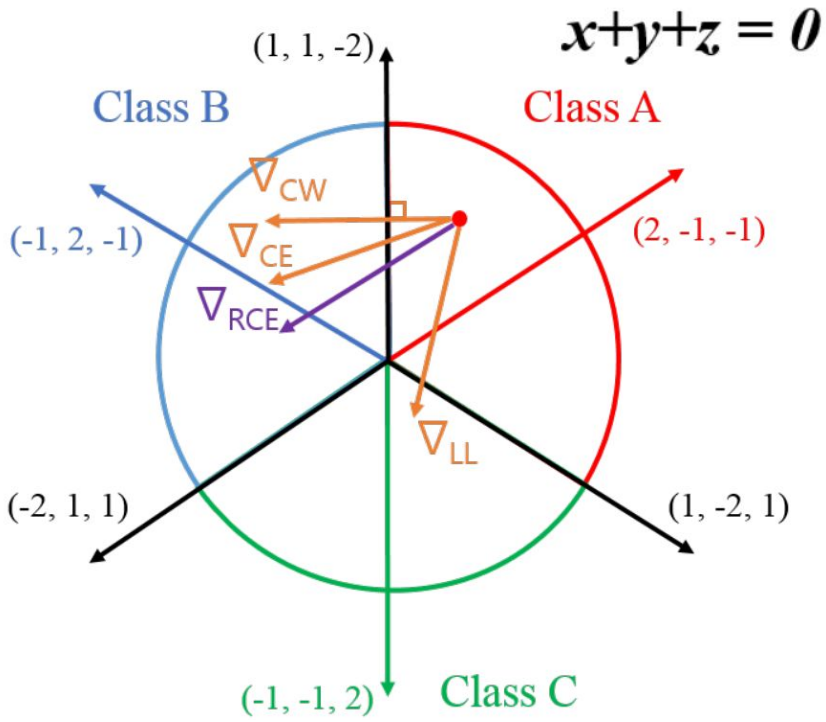


ICR attack strength is transferable? → **Transfer-based stronger attack**

Logit Vector Gradient Derivative

We calculate the derivative of various common losses with respect to logit vector **Z**. $\mathbf{Y}_{gt}$ is the ground-truth one-hot vector and **P** indicates the post-softmax probability vector $\mathbf{Y}_{LL}$ indicates one-hot label of the least likely class, and $j = \arg\max_{i \neq gt} Z(X^{adv})_i$ indicates the highest class except the gt class

Loss Type	Logit vector gradient	Sample value (x,y,z) = (1, 0.2, -1.2)
Cross Entropy (CE)	$\frac{\partial L_{CE}}{\partial \mathbf{Z}} = \mathbf{P} - \mathbf{Y}_{gt}$	(-0.36, 0.29, 0.07)
Carlini & Wagner (CW)	$\frac{\partial L_{CW}}{\partial \mathbf{Z}} = \mathbf{Y}_{j} - \mathbf{Y}_{gt}$	(-1.00, 1.00, 0.00)
CE (LL: Least Likely)	$\frac{\partial L_{CE(LL)}}{\partial \mathbf{Z}} = \mathbf{P} - \mathbf{Y}_{LL}$	(0.64, 0.29, -0.93)
Relative RCE (RCE)	$\frac{\partial L_{RCE}}{\partial \mathbf{Z}} = \frac{1}{K} - \mathbf{Y}_{gt}$	(-0.66, 0.33, 0.33)



Geometric illustration of the logit gradient with 3 classes

Given a logit vector (x, y, z), they can be represented in the 3-D space, assuming three logits are independent.

3-D space → Zero-sum: $x + y + z = 0$ → 2-D hyperplane

Loss Type	Gradient vector (x,y,z) = (1, 0.2, -1.2)
Cross Entropy (CE)	(-0.36, 0.29, 0.07)
Carlini & Wagner (CW)	(-1.00, 1.00, 0.00)
CE (LL: Least Likely)	(0.64, 0.29, -0.93)
Relative RCE (RCE)	(-0.66, 0.33, 0.33)

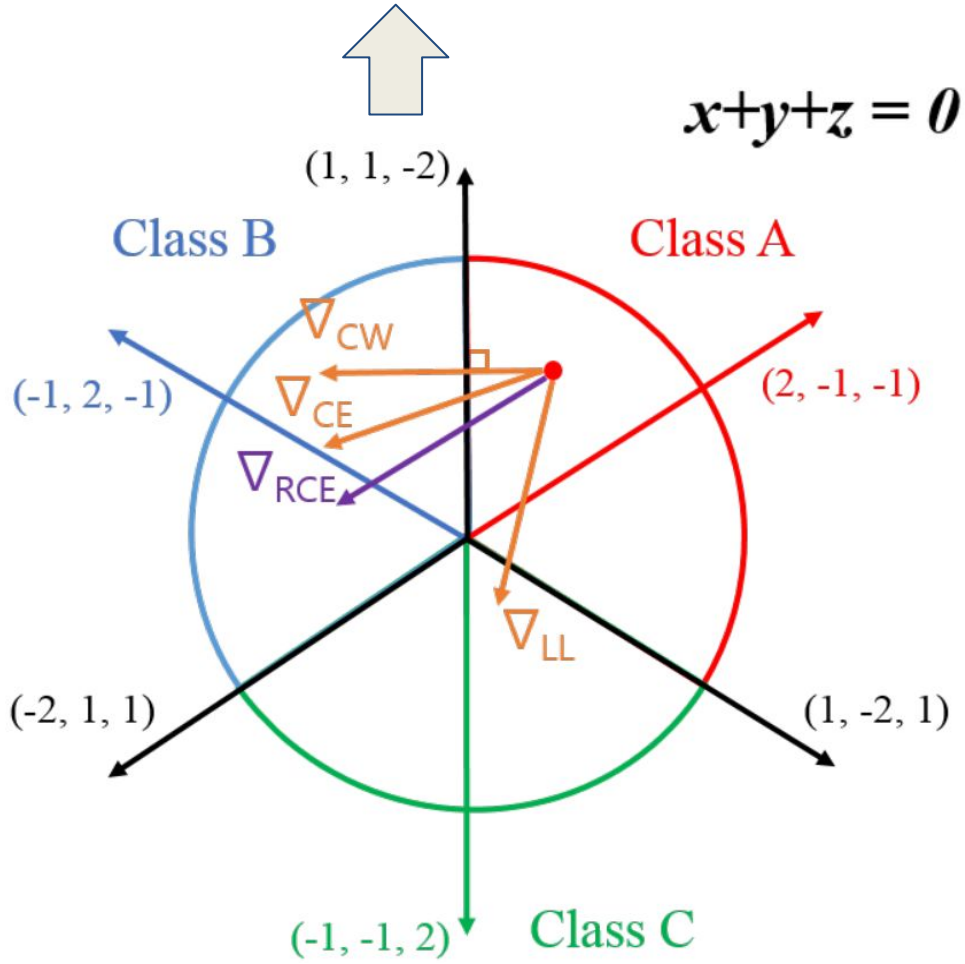
Interest class: A

Class A: (2, -1, -1)

Class B: (-1, 2, -1)

Class C: (-1, -1, 2)

RCE: find the opposite direction to shift **far from class A.**



Strongest White-box Attack with All Metrics

	non-targeted Acc.	ICR	OLNR	NLOR	NRT	CosSim
CE	100.00	752.90	712.35	159.52	279.53	0.25
CW	100.00	391.40	349.94	21.01	257.22	0.40
LL	99.20	491.02	490.46	888.96	306.12	0.08
FDA	100.00	619.90	608.84	517.28	311.49	0.06
RCE(Ours)	100.00	**1000.00**	**979.63**	570.94	**360.23**	**-0.21**
RCE(LL)	100.00	687.36	688.72	**996.32**	354.58	-0.17

FDA[1] is the existing **SOTA** approach.
RCE loss achieves **the strongest** attack among all losses for **all metrics** except for NLOR with CE(LL).

OLNR: Old Label New Ranking
NLOR: New Label Old Ranking
NRT: Normalized Rank Transformation
CosSim: Cosine Similarity

[1] FDA: Feature Disruptive Attack; Ganeshan, Vivek, Babu; CVPR 2019

ICR Results for Non-targeted and Targeted Attack

Non-targeted ICR:
higher is better

Targeted ICR:
lower is better

Table 4: Non-targeted ICR of I-FGSM (top), and MI-DI-TI-FGSM (bottom) attacks for source network ResNet50.

	RN50	DN121	VGG16bn	RN152	MNv2	IncV3
CW	390.00	14.80	18.59	24.15	22.68	5.49
CE	752.90	34.16	40.87	61.20	39.21	7.50
RCE (Ours)	1000.00	72.11	80.86	144.81	70.39	13.35
CW	427.49	77.82	77.13	81.67	84.88	39.03
CE	806.85	220.87	213.77	249.02	193.96	89.93
RCE (Ours)	999.94	482.58	430.97	517.85	366.30	141.90

Table 5: Targeted ICR of I-FGSM (Top), MI-DI-TI-FGSM (bottom) for source network ResNet50.

	RN50	DN121	VGG16bn	RN152	MNv2	IncV3
CE	2.52	320.73	355.33	264.20	345.40	607.46
Po-Trip	1.00	236.37	299.51	192.63	309.81	582.28
RCE (Ours)	1.02	161.13	208.61	108.22	244.40	559.95
CE	1.00	22.19	45.64	23.61	92.72	245.79
Po-Trip	1.00	13.84	40.33	18.46	76.37	215.26
RCE (Ours)	1.01	4.51	7.76	3.67	30.90	157.35

RCE loss **outperforms** other common losses in both non-targeted and targeted settings with the **ICR metric**

Towards Simple Yet Effective Transferable Targeted Adversarial Attacks

Philipp Benz*, Chaoning Zhang*, Adil Karjauv, In So Kweon
Korea Advanced Institute of Science and Technology (KAIST)

RobustML workshop paper at ICLR 2021

Techniques for Improving Transferability - Noise Augmentation

Two simple techniques to improve transferability

Input Noise Augmentation

Feature Noise Augmentation

Techniques for Improving Transferability - Push-Pull Loss

The targeted CE loss only maximizes the probability of the target class **without** explicitly encouraging the sample to decrease the probability of the ground-truth class.

Conjecture: The effectiveness of adversarial examples can be increased by **not only** increasing the logit of the target class **but also** decreasing the probability of the ground-truth class.

Push-Pull (PP) loss -- Combination of two CE losses:

- One for **pushing** the sample far from the ground-truth class (CE_{gt}), i.e. non-targeted CE loss
- One for pulling the sample close to the target class (CE_{tar}), i.e. targeted CE loss

$$PP(Z, y_{gt}, y_{tar}) = CE_{tar}(Z, y_{tar}) - \beta CE_{gt}(Z, y_{gt})$$

Z: Output logit
y_{gt}: ground-truth one-hot label
y_{tar}: target one-hot label
β: balancing weight for the two CE losses

Transfer-based attack in non-targeted setting

Performance evaluation in the non-targeted attack setting with a single substitute model, i.e. DenseNet121. The results are reported in the ASR (%) for various baselines: MI, TI, DI, MI-DI-TI. All experiments were performed with the non-targeted CE loss

Substitute	FGSM variant	Attack	RN50	VGG16	DN201	MNv2	IncV3	Avg.
DN121	MI-FGSM	CE	88.8	85.9	95.1	84.0	58.1	82.4
		CE + I_{Aug}	94.1	91.8	97.6	88.0	70.3	88.4
		CE + F_{Aug}	98.1	96.5	99.4	95.5	73.7	92.6
		CE + I_{Aug} + F_{Aug}	98.2	97.5	99.5	96.3	83.0	94.9
	TI-FGSM	CE	86.2	82.7	93.2	78.9	46.7	77.5
		CE + I_{Aug}	90.8	87.7	96.3	82.8	59.3	83.4
		CE + F_{Aug}	96.2	93.3	98.5	92.8	65.4	89.2
		CE + I_{Aug} + F_{Aug}	97.8	95.2	99.0	93.5	74.2	91.9
	DI-FGSM	CE	96.3	96.3	98.1	91.7	62.1	88.9
		CE + I_{Aug}	96.6	96.4	97.9	91.7	71.4	90.8
		CE + F_{Aug}	98.9	98.5	99.4	97.1	74.6	93.7
		CE + I_{Aug} + F_{Aug}	98.6	98.4	99.0	96.8	78.1	94.2
	MI-DI-TI-FGSM	CE	98.3	97.2	99.3	95.6	83.6	94.8
		CE + I_{Aug}	98.2	97.7	99.1	96.3	86.8	95.6
		CE + F_{Aug}	99.7	99.2	99.8	98.6	91.6	97.8
		CE + I_{Aug} + F_{Aug}	99.5	99.4	99.7	99.0	91.6	97.8

Noise and Feature Augmentation as well as their combination improve adversarial transferability

Transfer-based Targeted Attacks - Single Surrogate

Non-targeted ASR/targeted ASR for a targeted MI-DI-TI attack with a single substitute model (ResNet50) in the targeted attack scenario.

Attack	DN121	VGG16	RN152	MNv2	IncV3	Avg.
CE	84.2/40.2	88.6/28.0	82.6/43.1	84.7/10.4	52.9/4.6	78.6/25.3
Po-Trip	84.0/56.7	86.0/33.1	83.0/55.5	81.5/15.1	51.0/7.1	77.1/33.5
FDA$^{(5)}$+xent	90.9/57.9	88.8/43.5	89.7/51.6	86.4/22.9	-	-
PP	97.6/73.1	97.8/62.5	98.2/78.2	95.0/28.5	71.8/10.8	92.1/50.6
PP + I_{Aug}	98.8/78.3	98.5/68.5	99.3/82.3	96.7/38.4	79.6/21.4	94.6/57.8
PP + F_{Aug}	99.8/87.6	99.7/82.6	99.8/90.5	97.7/56.0	80.1/28.7	95.4/69.1
PP + I_{Aug} + F_{Aug}	99.9/87.2	99.8/81.0	100.0/90.8	99.0/67.2	87.4/42.6	97.2/73.8

The Push-Pull Loss itself and in combination with noise and feature augmentation improve adversarial transferability in the single surrogate scenario

Transfer-based Targeted Attacks - Ensemble Surrogate

Non-targeted ASR/Targeted ASR for a targeted MI-DI-TI attack with an ensemble of two substitute models.

Source	Attack	VGG16	RN152	DN201	MNv2	IncV3	IncV4	IncRes	Avg.
RN50 + DN121	CE	93.9/54.9	91.1/68.5	95.4/86.2	90.8/24.4	64.7/15.6	66.5/14.2	49.6/8.3	78.86/39.0
	Po-Trip	88.2/44.9	84.7/63.5	92.6/82.9	83.9/21.7	59.3/14.6	61.2/13.1	46.1/7.2	73.71/35.0
	PP	99.2/81.6	99.3/87.2	100.0/93.7	97.7/51.0	84.8/33.2	85.9/29.4	68.1/17.5	90.71/56.0
	PP + I_{Aug}	99.8/82.2	99.8/90.7	100.0/94.3	98.8/61.2	90.8/50.4	91.9/45.0	78.9/33.0	94.29/65.0
	PP + F_{Aug}	100.0/92.4	100.0/94.3	100.0/95.3	99.9/81.6	94.2/63.9	94.6/62.9	81.5/41.4	95.74/76.0
	PP + I_{Aug} + F_{Aug}	**100.0/90.9**	**100.0/94.3**	**100.0/94.6**	**99.9/85.3**	**97.8/74.2**	**97.3/73.0**	**88.6/55.0**	**97.66/81.0**

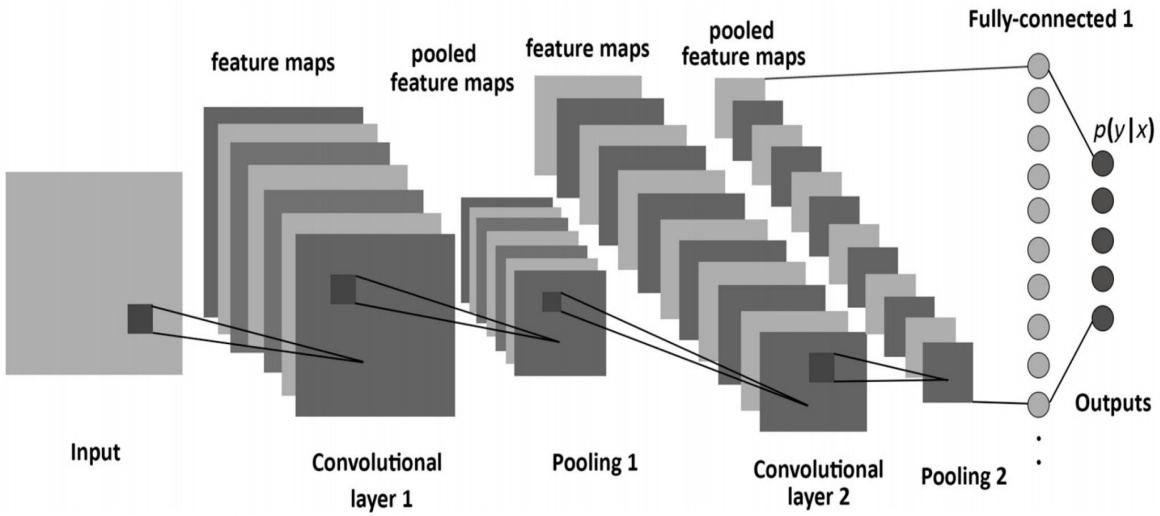
The Push-Pull Loss itself and in combination with noise and feature augmentation improve adversarial transferability in the ensemble surrogate scenario

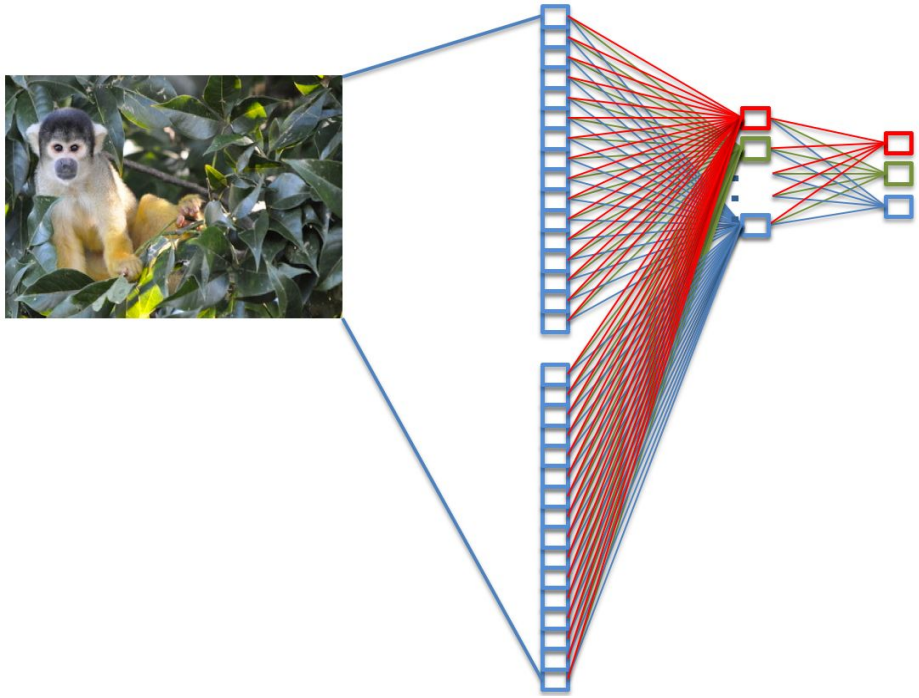
Robustness Comparison of Vision Transformer and MLP-Mixer to CNNs

Philipp Benz*, Chaoning Zhang*, Soomin Ham*, Adil Karjauv, In So Kweon
Korea Advanced Institute of Science and Technology (KAIST)

CVPR 2021 Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges (AML-CV)

Convolutional Neural Networks & Fully Connected Neural Networks





CNN

CNN is locally connected and sharing weights by convolving kernels.

- Consists of convolutional layers followed by pooling layer
- Shift-invariant

FC

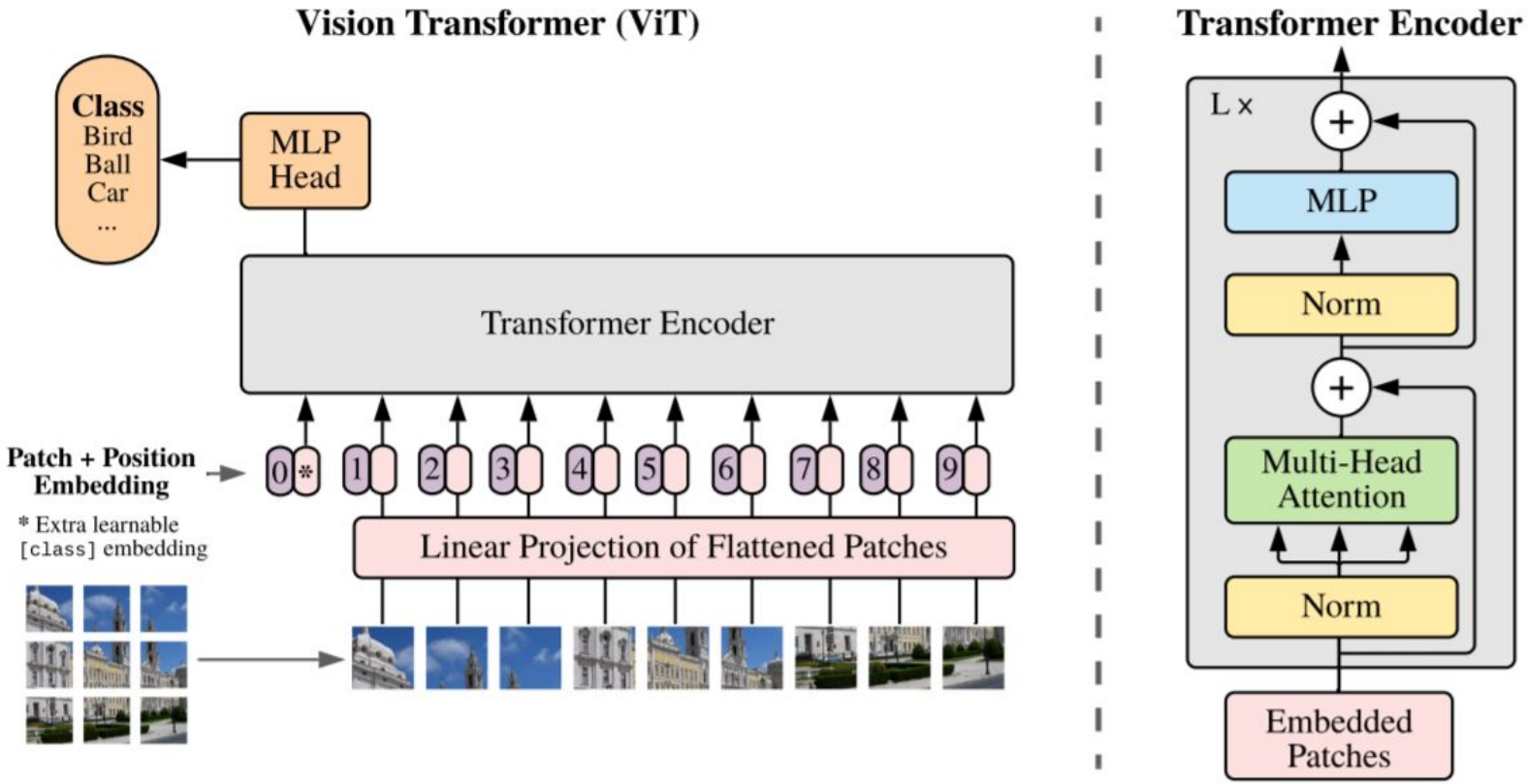
FC reacts differently to an input image and its shifted version.

- Ignores the information brought by pixel position and correlation with neighbors.
- Cannot handle translation (shift-variant)

Picture Source: Matthew Stewart, Simple-introduction-to-convolutional-neural-networks, <https://towardsdatascience.com> and Christian wolf, what-is-translation-equivariance-and-why-do-we-use-convolutions-to-get-it, <https://chriswolfvision.medium.com>

Vision Transformers [1]

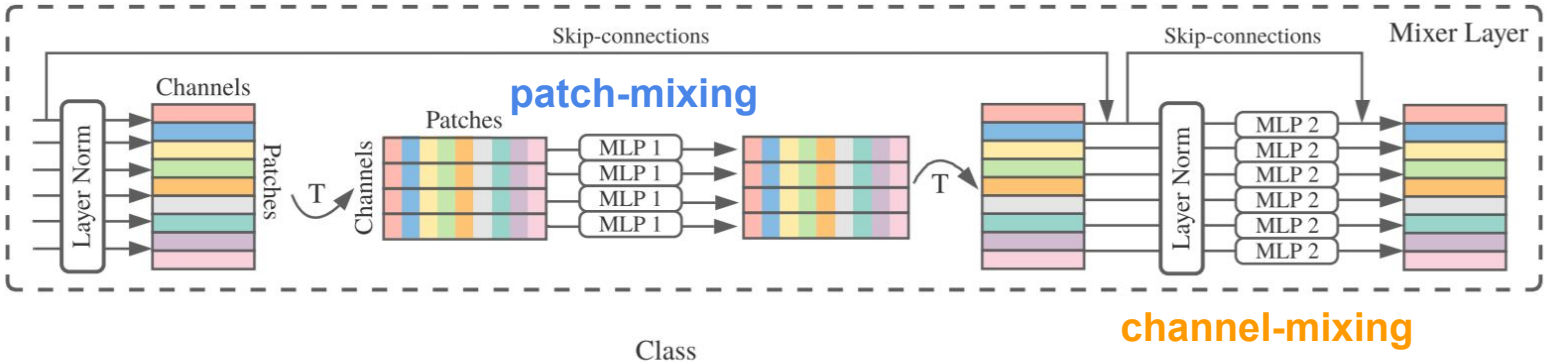
Vision Transformer splits the input image into patches and feeds the linear embedding sequence of these patches as inputs to the Transformer[2]. Image patches are processed in the same way as token(word) in NLP applications.

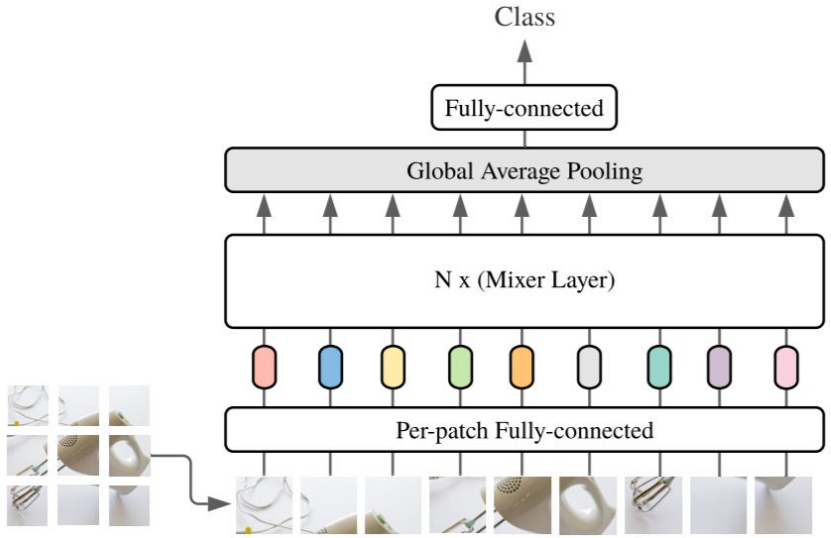


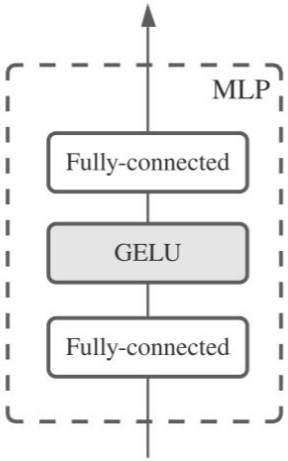
[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).
[2] Vaswani, Ashish, et al. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

MLP-Mixer [3]

Mixer architecture clearly separates the per-location (channel-mixing) operations and cross-location (patch-mixing) operations. Both operations are implemented with MLPs.







[3] Tolstikhin, Ilya, et al. "Mlp-mixer: An all-mlp architecture for vision." *arXiv preprint arXiv:2105.01601* (2021).

Overview

- Despite the success of CNNs, they remain vulnerable to adversarial examples whose small additive perturbations of the input cause the CNN to misclassify a sample.

- Due to the rather recent introduction of the ViT and Mixer architecture, the adversarial vulnerability of these novel architectures has not been well studied yet.

- **This work sets out to explore and analyze the adversarial vulnerability of ViT and Mixer architectures and compare the findings against the CNN models.**

Methodology

- **Models and Dataset**
  - ViT-B/16, ViT-L/16
  - Mixer-B/16, Mixer-L/16
  - ResNet18 (SWSL), ResNet50 (SWSL),
    ResNet18 (SSL), ResNet50 (SSL),
    ResNet18, ResNet50

- **Test on NeurIPS 2017 adversarial challenge dataset**
  - ImageNet-compatible dataset composed of 1,000 images in 430 classes.

Robustness Against White-Box Attacks

Attack Success Rate (ASR): Percentage of samples which were classified differently from the ground-truth class

Clean accuracy on NeurIPS & ImageNet dataset, the attack success rate (%) of PGD and FGSM under l_∞ distortion, and the l_2-norm of C&W and DeepFool

	Clean		PGD (ℓ_∞)					FGSM (ℓ_∞)					C&W (ℓ_2)	DeepFool (ℓ_2)
Model	ImageNet	NeurIPS	0.1	0.3	0.5	1	3	0.1	0.3	0.5	1	3		
ViT-B/16	81.4	90.7	**22.6**	63.6	86.5	97.5	99.9	**19.1**	38.7	52.8	66.3	79.7	**0.468**	0.425
ViT-L/16	82.9	89.3	22.8	**60.1**	80.9	95.8	100	19.5	**35.9**	**44.9**	**57.9**	**67.3**	0.459	**0.548**
Mixer-B/16	76.5	86.2	29.5	63.4	82.0	96.2	100	27.7	49.3	59.5	69.3	78.0	0.375	0.339
Mixer-L/16	71.8	80.0	41.1	67.3	**80.4**	**92.1**	**99.4**	36.7	51.8	56.9	61.6	67.4	0.297	0.377
ResNet-18 (SWSL)	73.3	90.4	47.9	93.7	98.7	99.5	99.6	38.0	76.3	89.9	96.2	97.6	0.295	0.132
ResNet-50 (SWSL)	81.2	96.3	39.4	90.2	97.0	98.4	**99.4**	26.3	60.9	73.0	83.8	87.5	0.380	0.149
ResNet-18 (SSL)	72.6	90.5	42.3	93.2	98.8	99.8	99.8	34.3	75.1	88.9	96.6	97.9	0.312	0.142
ResNet-50 (SSL)	79.2	95.3	39.5	91.8	97.6	99.5	99.9	26.3	60.5	75.2	85.8	89.5	0.372	0.149
ResNet-18	69.8	83.7	46.1	90.0	97.8	99.9	100	42.0	75.2	88.5	95.7	98.2	0.302	0.237
ResNet-50	76.1	93.0	35.8	86.3	97.9	99.5	100	27.5	63.1	77.6	89.4	93.9	0.371	0.287

ViT & Mixer have a lower attack success rate compared with the CNN architecture

Lower l2-norm for the C&W and DeepFool attacks when applied to the CNNs

→ViT & Mixer are more robust

Exception: Mixer model exhibits increased vulnerability to very small perturbations

Robustness Against Black-Box Attacks

❑ Query-based

Boundary Attack: A decision-based attack that starts from a large adversarial perturbation and then seeks to reduce the perturbation while staying adversarial. [1]

We test 100 random samples from NeurIPS dataset, and the l2-norm of adversarial perturbation is presented

	ViT-B	ViT-L	Mix-B	Mix-L	RN18 (SWSL)	RN50 (SWSL)	RN18 (SSL)	RN50 (SSL)	RN18	RN50
Boundary (ℓ_2)	3.980	7.408	1.968	1.951	1.403	1.846	1.434	1.780	1.468	1.740

ViT and Mixer models are more robust, indicated by the relatively higher l2-norm of the adversarial perturbation

[1] Brendel et al.; Reliable attacks against black-box machine learning models; ICLR 2018

Robustness Against Black-Box Attacks

❑ Transfer-based

We report the attack success rate (%) and a model with a lower ASR is considered to be more robust.

		Target model									
Source model	**Variant**	**ViT-B/16**	**ViT-L/16**	**Mixer-B/16**	**Mixer-L/16**	**ResNet-18 (SWSL)**	**ResNet-50 (SWSL)**	**ResNet-18 (SSL)**	**ResNet-50 (SSL)**	**ResNet-18**	**ResNet-50**
ViT-B/16	I-FGSM	100	84.7	48.8	50.5	32.0	20.5	34.3	23.4	40.9	31.7
ViT-L/16	I-FGSM	90.9	99.9	45.7	48.0	30.4	22.2	34.4	23.6	40.8	30.9
Mixer-B/16	I-FGSM	33.9	25.3	100	89.1	30.6	20.5	34.5	23.3	40.8	32.0
Mixer-L/16	I-FGSM	27.7	20.1	80.3	99.7	27.7	17.0	31.5	17.5	38.2	28.4
ResNet-18 (SWSL)	I-FGSM	16.2	13.6	24.8	29.5	99.6	57.1	80.2	58.0	73.5	63.4
ResNet-50 (SWSL)	I-FGSM	15.3	13.5	23.6	29.9	56.5	99.5	51.6	69.1	49.4	51.0
ResNet-18 (SSL)	I-FGSM	17.7	13.7	28.6	34.4	84.4	54.6	99.9	65.4	78.2	66.8
ResNet-50 (SSL)	I-FGSM	18.1	15.0	26.4	32.3	58.9	73.3	64.7	100	54.7	62.2
ResNet-18	I-FGSM	18.2	14.7	28.9	35.6	84.6	49.9	85.3	60.4	100	81.6
ResNet-50	I-FGSM	17.7	13.6	28.4	34.5	73.9	63.9	74.3	74.7	80.6	100

Adversarial examples from the same family (or similar structure) exhibit higher transferability, suggesting models from the same family learn similar features.

When a different model architecture is used as the source model, there is also a trend that CNNs are relatively more vulnerable (i.e., transfer poorly toward foreign architectures).

Toy Example

Binary classification: Horizontal and vertical black stripe on a black background

Images for the binary classification toy example

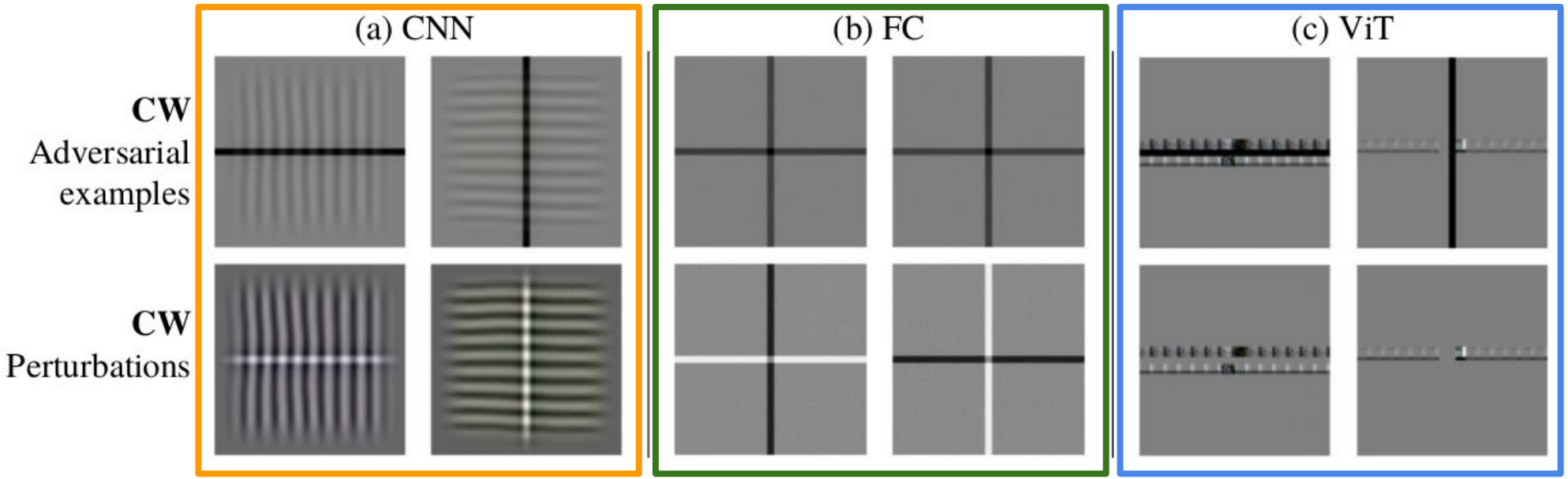
Train a Fully Connected network (FC), a Convolution Neural Network (CNN), and a Vision Transformer (ViT) on the images of similar (small) capacity

Evaluation with the C&W and DDN attack for the models trained on the toy example

	C&W (ℓ_2)	DDN (ℓ_2)	$\#params$
CNN	12.55	13.91	$4.59M$
FC	25.06	25.39	$4.82M$
ViT	27.82	59.99	$4.88M$

The CNN is also less robust than the FC and the ViT in this toy example setup

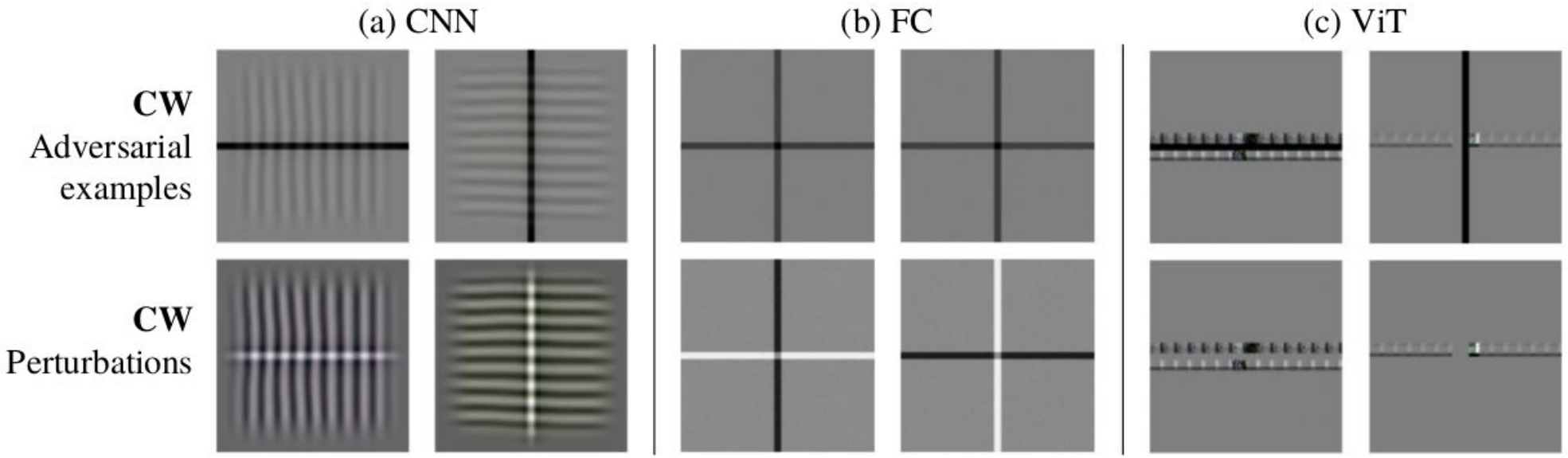
Explanation from the perspective of shift-invariance



Adversarial examples and perturbations generated against C&W attack using different architectures trained on toy example.

- ViT: square patches → likely due to the division of the input image into patches in the ViT architecture
- CNN: repeated stripes
- FC: only a single stripe in the center

Explanation from the perspective of shift-invariance



Adversarial examples and perturbations generated against C&W attack using different architectures trained on toy example.

Due to shift-invariance, CNN has perturbations all over the image (in the direction of the opposite class' stripe), and the other models have perturbations only in the middle part. → lower l2-norm refers to less robust

Summary & Take Aways

- **An empirical study on the adversarial robustness comparison of ViT and MLP-Mixer to the widely used CNN on image classification.**

  - White-box adversarial attack
  - Black-box adversarial attack (query-based and transfer-based)
  - Toy example on binary classification
  - Frequency Analysis

- **ViT is significantly more robust than CNN, and Mixer is generally located between ViT and CNN.**

- **The lower robustness of CNN can be partially attributed to the shift-invariant property of CNNs.**

Thank You