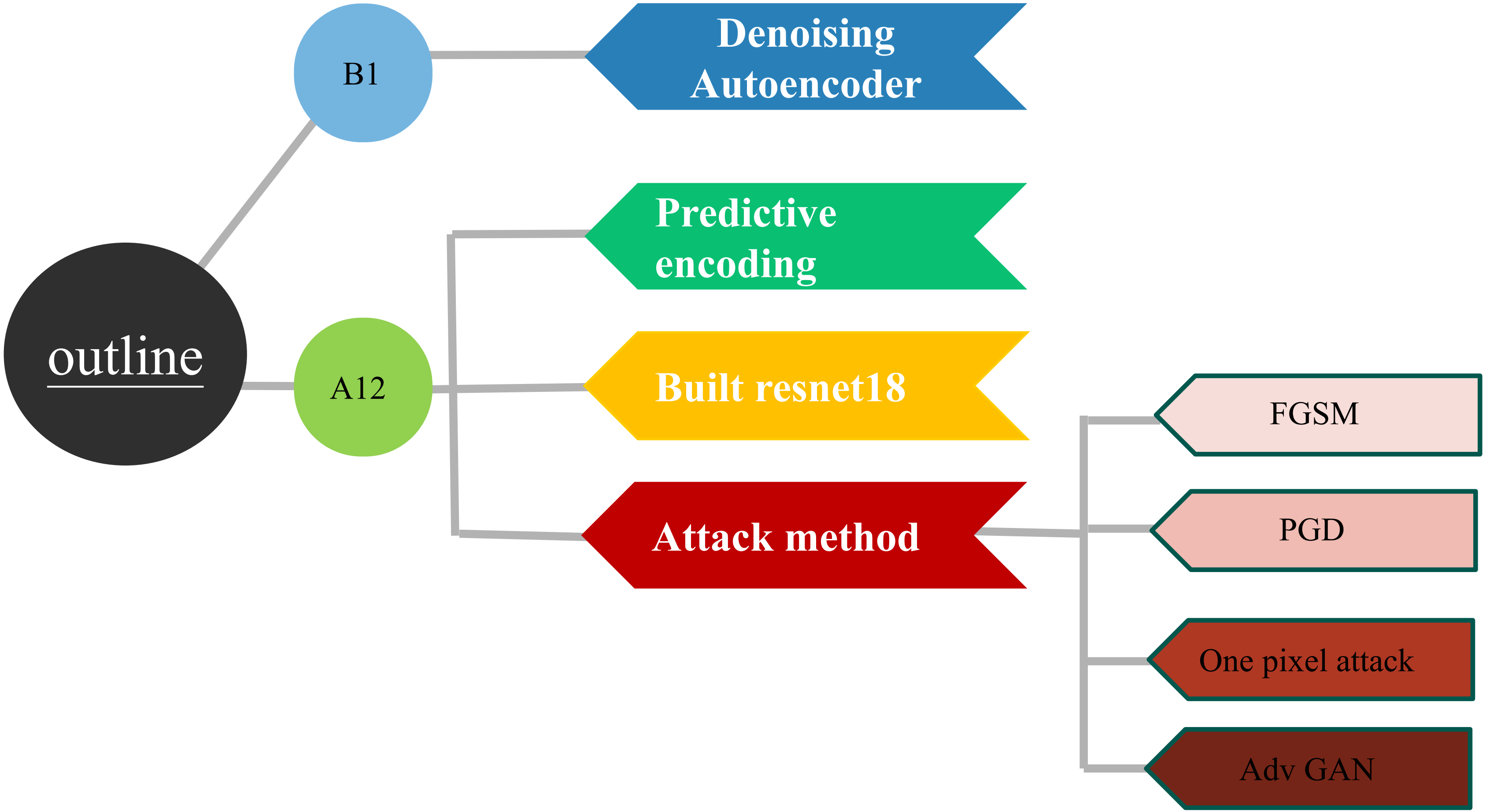
Result report

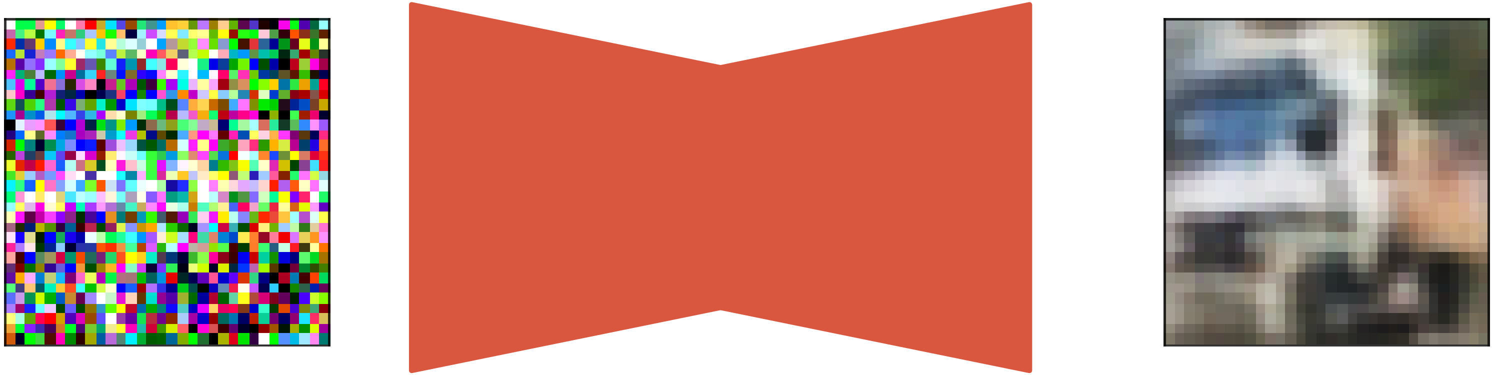
毛柏元 Mao po yuan
E-mail:zxc596666123@gmail.com


outline
B1
Denoising Autoencoder
A12
Predictive encoding
Built resnet18
Attack method
FGSM
PGD
One pixel attack
Adv GAN

B1 Reconstruct images from CIFAR 10

Denoising Autoencoder

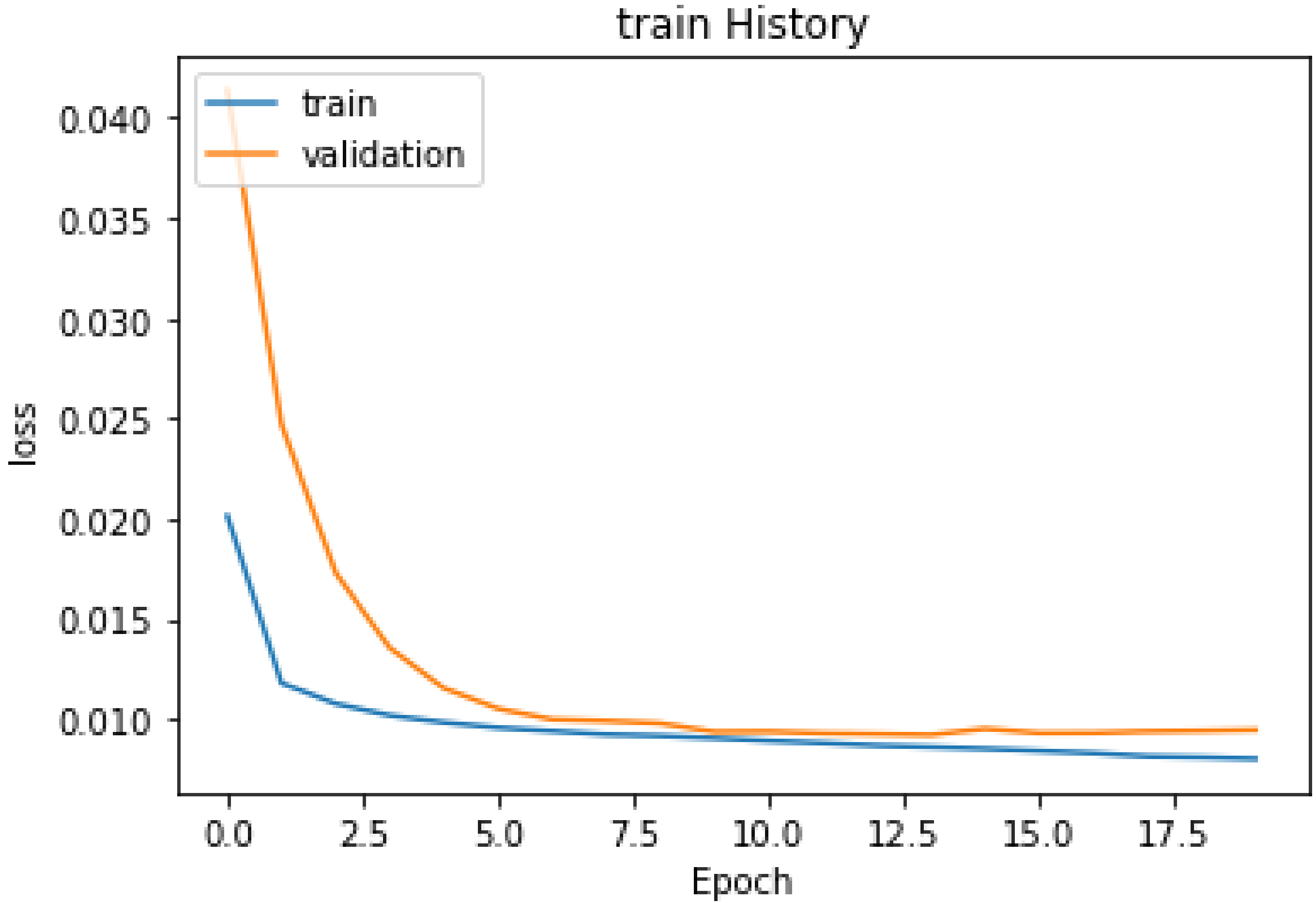
Reference https://codahead.com/blog/a-denoising-autoencoder-for-cifar-datasets



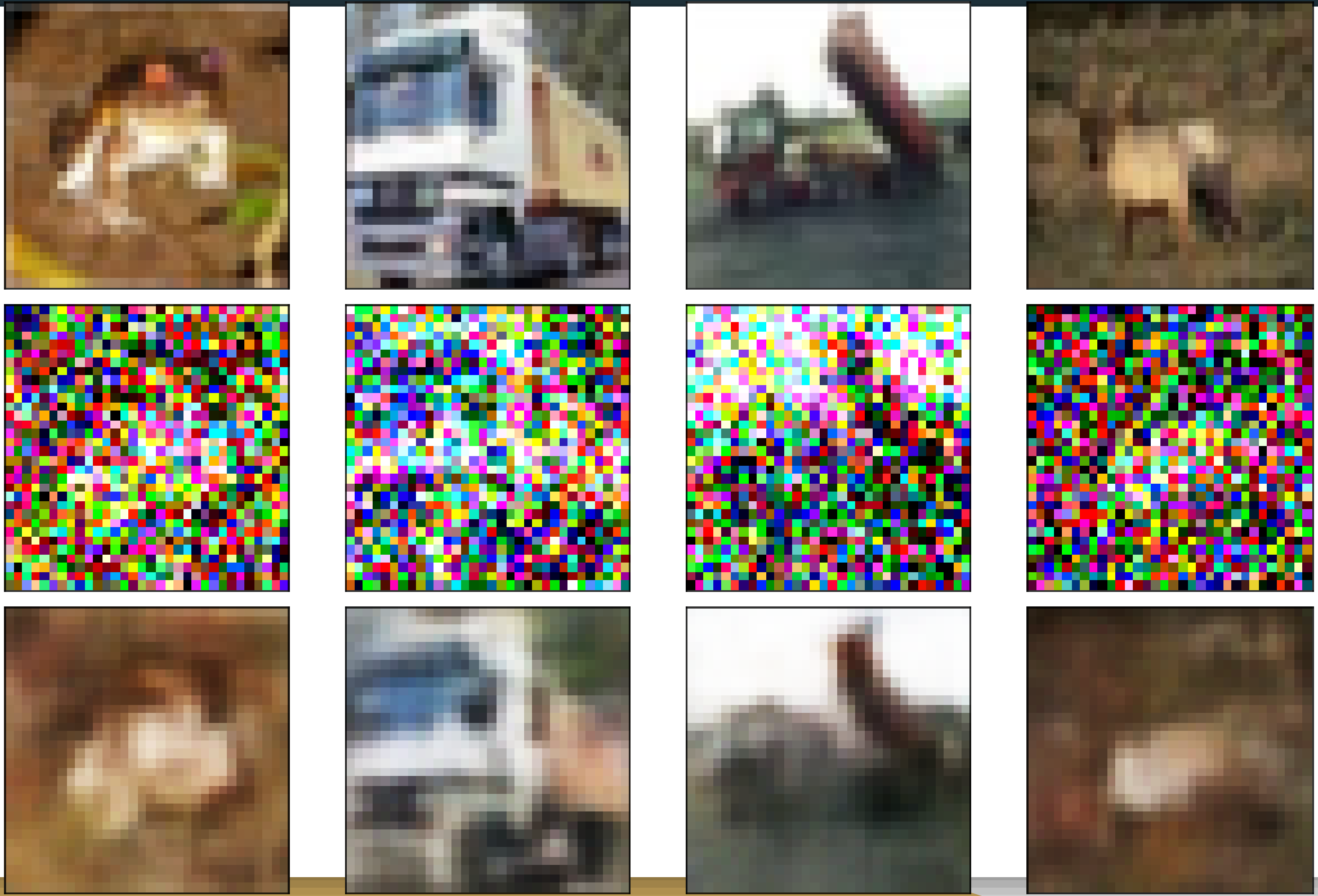
Noise picture
(clean picture with random noise)

decode picture

history

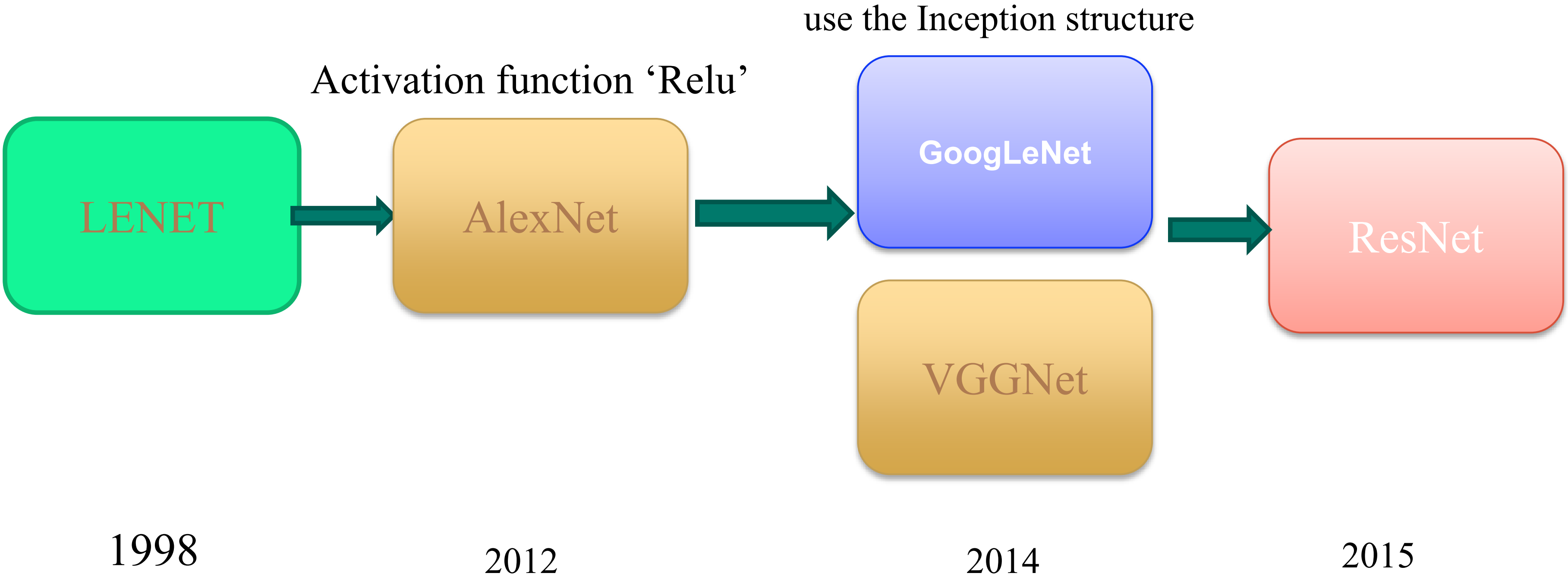

train History
train
validation
loss
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epoch

result



A11 reate a predictive encoding Model on CIFAR10 and attack it to verify its robustness

predictive encoding



use the Inception structure

Activation function ‘Relu’

LENET

AlexNet

GoogLeNet

VGGNet

ResNet

1998

2012

2014

2015

Reference :(AlexNet、VGG、Inception、ResNet)+Keras Coding

Fast Gradient Sign Method (FGSM)

white box attack

Input : image and Model
Output : Perturbation

FGSM

$$x^{*} \leftarrow x^{0} - \varepsilon \Delta x$$

$$\Delta x = \begin{bmatrix} sign(\partial L/\partial x_{1}) \\ sign(\partial L/\partial x_{2}) \\ \vdots \\ \vdots \end{bmatrix}$$

x^{*}: picture with perturbation

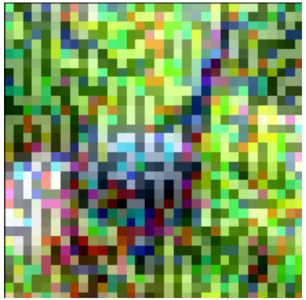
L:loss function

ε :limit of perturbation

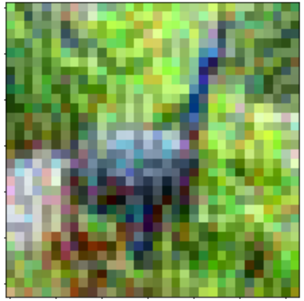
x^{0}: original picture

ε	**0.2**	**0.1**	**0.05**	**0.01**
Success rate	66.03%	68.44%	69.37%	65.83%
Success Robustness	0.404	0.203	0.1019	0.02

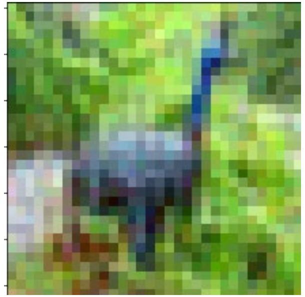
Bird 99.9%



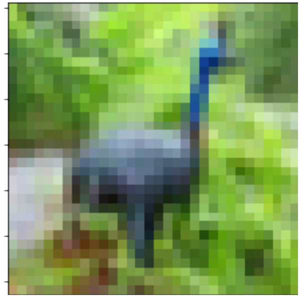
4.9e-5% bird
99.9% frog



2.0e-2% bird
65.5% frog



5.1e-2% bird
42.5% cat



12.3% bird
29.3% cat

PGD ATTACK

white box attack

Input : image and Model
Output : Perturbation

Reference : Towards Deep Learning Models Resistant to Adversarial Attacks
https://arxiv.org/pdf/1706.06083.pdf

PGD

$$x^{\mathrm{n}} \leftarrow x^{n-1} - a\Delta x^{n-1}$$

$$\Delta x^{n-1} = \begin{bmatrix} sign(\partial L/\partial x_1) \\ sign(\partial L/\partial x_2) \\ \vdots \\ \vdots \end{bmatrix}$$

$$\varepsilon \geq \Sigma a\Delta x^{n-1}$$

x^{n}: picture with all perturbation at epochs n

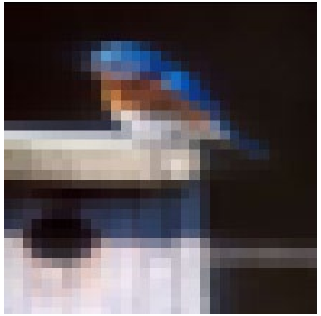
a : each epochs step perturbation

ε : limit of perturbation

ε	**0.1**	**0.1**
Epochs	10	20
a(each epochs step perturbation)	0.01	0.001
Success Robustness	0.0314	0.0149
Success rate	70.19	70.2%

Result

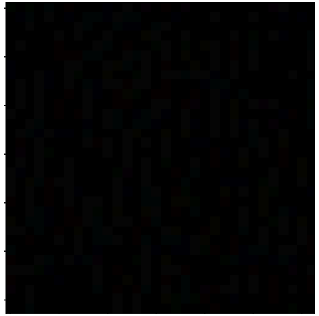
Original image



Original 99.9% bird

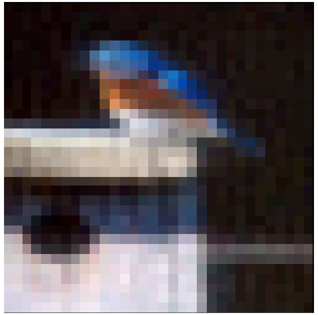
+

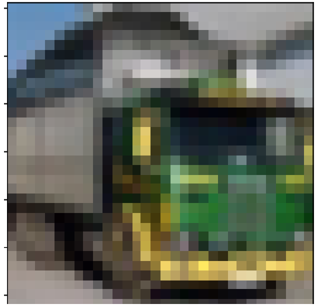
perturbation



=

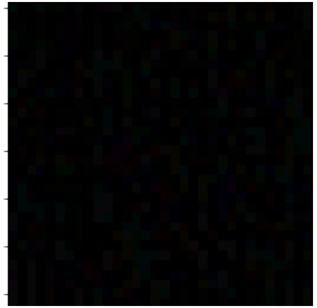
100% truck
3.2e-14% bird



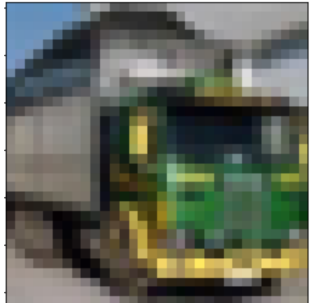


Original 99.9% truck

+



=



99.9% bird
1.14e-10% truck

ONE pixel attack

black box attack

Input : image
Output : Perturbation

Reference :One pixel attack
https://arxiv.org/abs/1710.08864

Difference from others attack

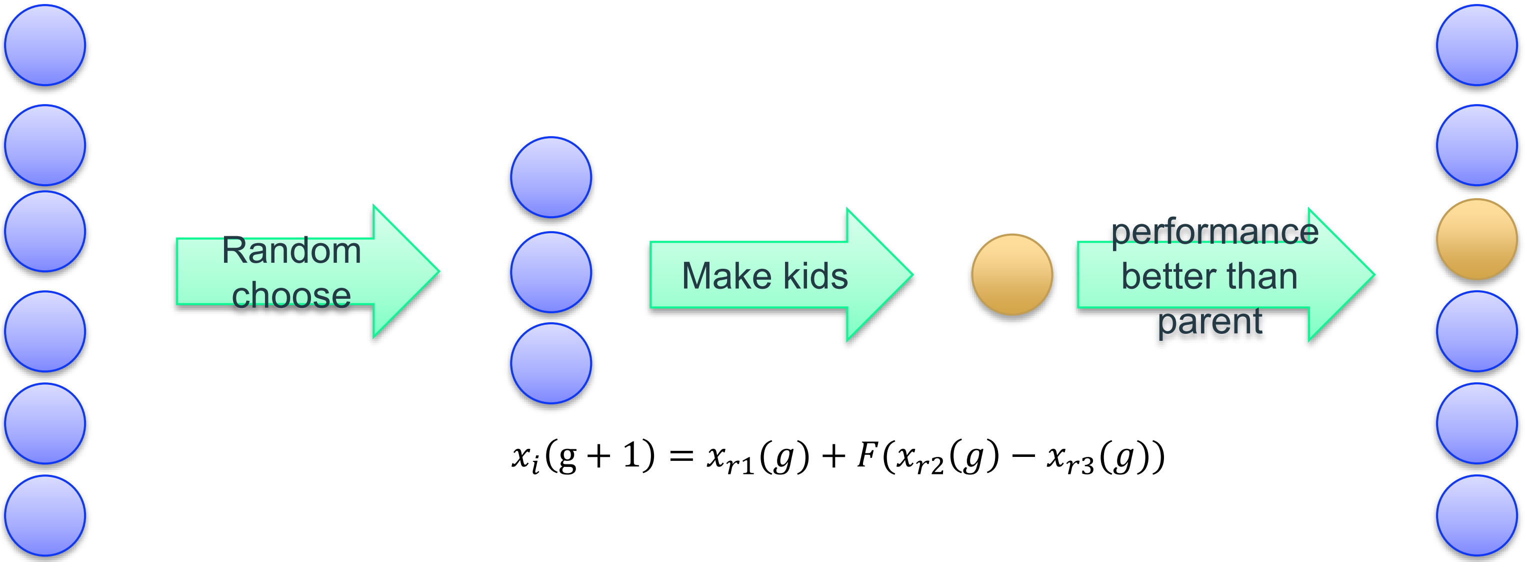
$$\begin{array}{ll} \underset{e(\mathbf{x})^*}{\text{maximize}} & f_{adv}(\mathbf{x}+e(\mathbf{x})) \\ \text{subject to} & \|e(\mathbf{x})\| \leq L \end{array}$$

Other attacks

$$\begin{array}{ll} \underset{e(\mathbf{x})^*}{\text{maximize}} & f_{adv}(\mathbf{x}+e(\mathbf{x})) \\ \text{subject to} & \|e(\mathbf{x})\|_0 \leq d, \end{array}$$

One pixel attack

Method: Differential Evolution

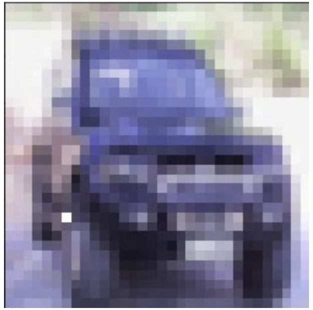


$$x_i(\mathrm{g}+1) = x_{r1}(g) + F(x_{r2}(g) - x_{r3}(g))$$

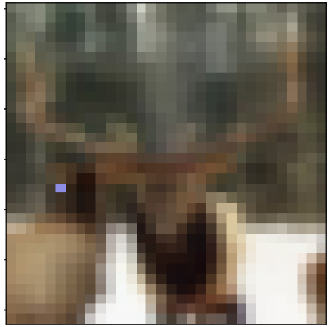
Parents(x , y , R,G,B)

New Parents(x , y , R,G,B)

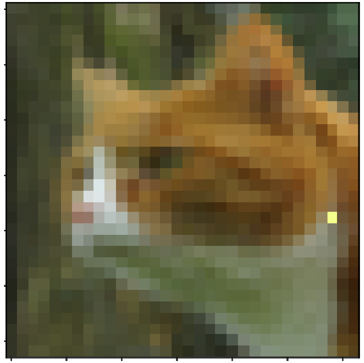
Result



Original 81.8% automobile
After 34% automobile
37.8% truck



Original 78.8% deer
After 6.7% deer
91.6% cat



Original 81.9% cat
After 8.5% cat
90.6% frog

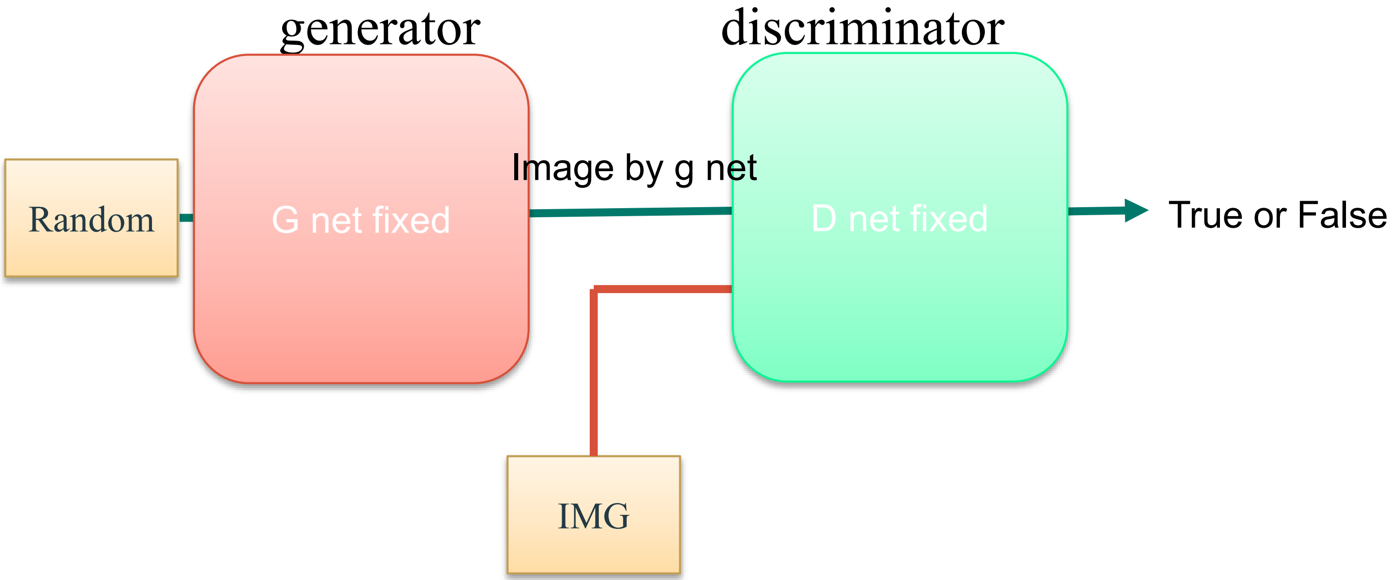
ADV GAN

white box attack

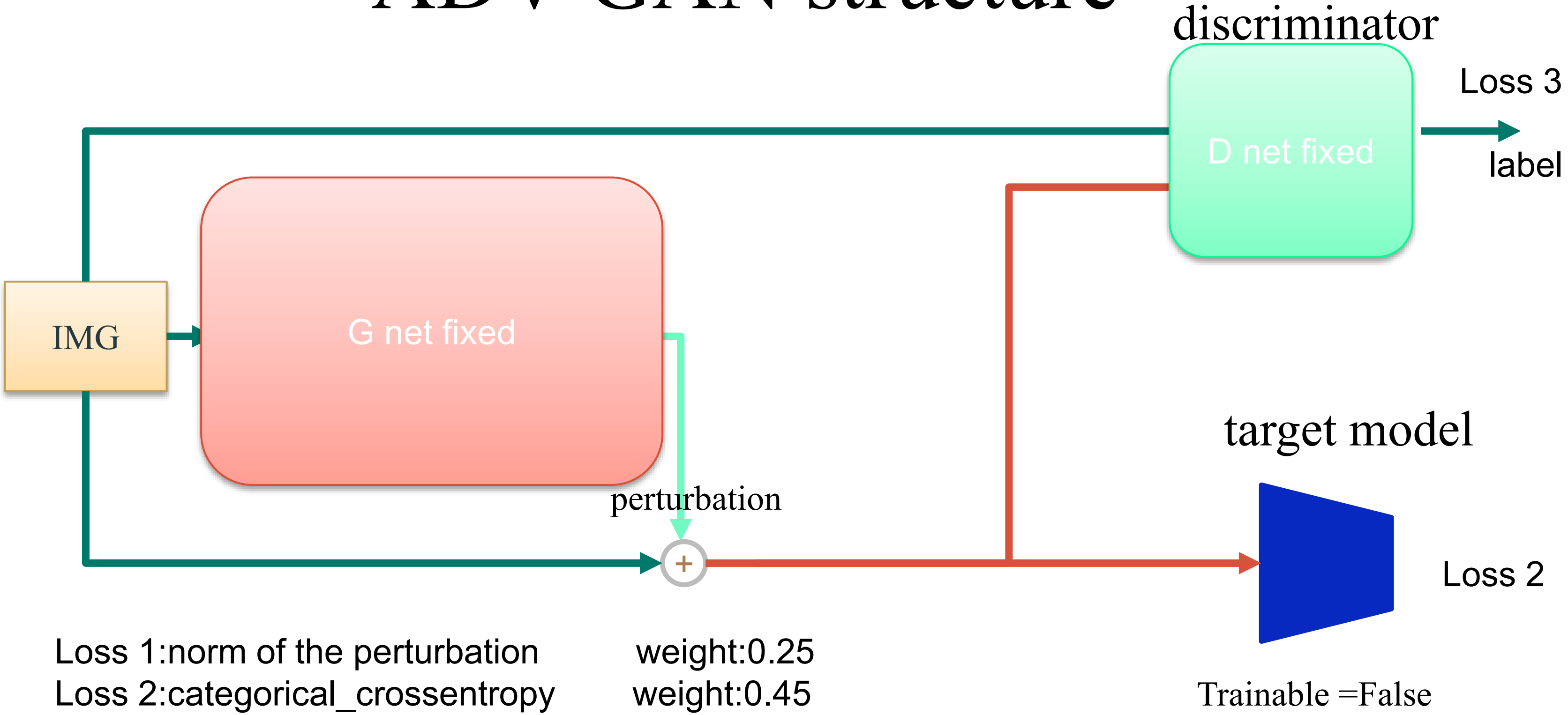
Input : image and Model
Output : Perturbation

Reference :ADVGAN https://arxiv.org/abs/1801.02610

GAN

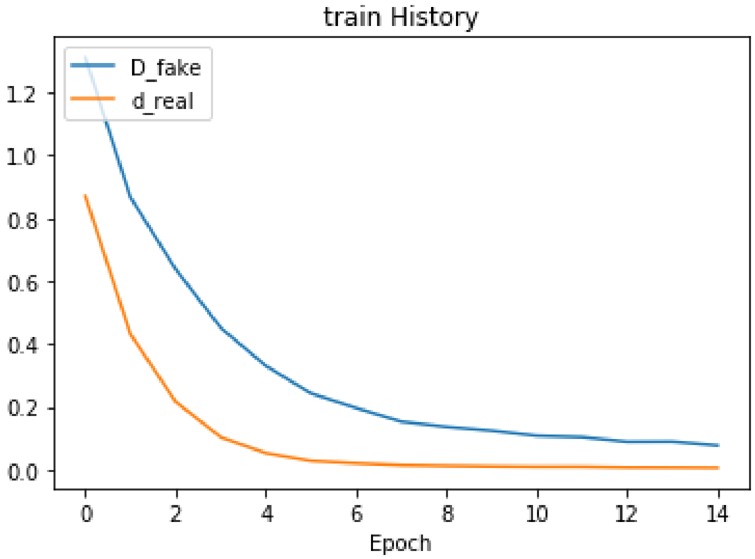

generator
discriminator
Random
G net fixed
Image by g net
D net fixed
True or False
IMG

ADV GAN structure

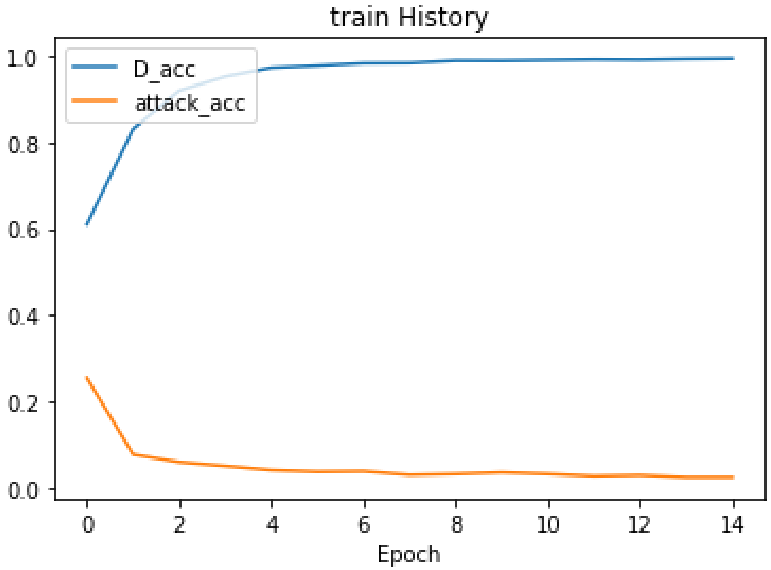


Loss 1:norm of the perturbation weight:0.25
Loss 2:categorical_crossentropy weight:0.45
Loss 3:categorical_crossentropy weight:1.2

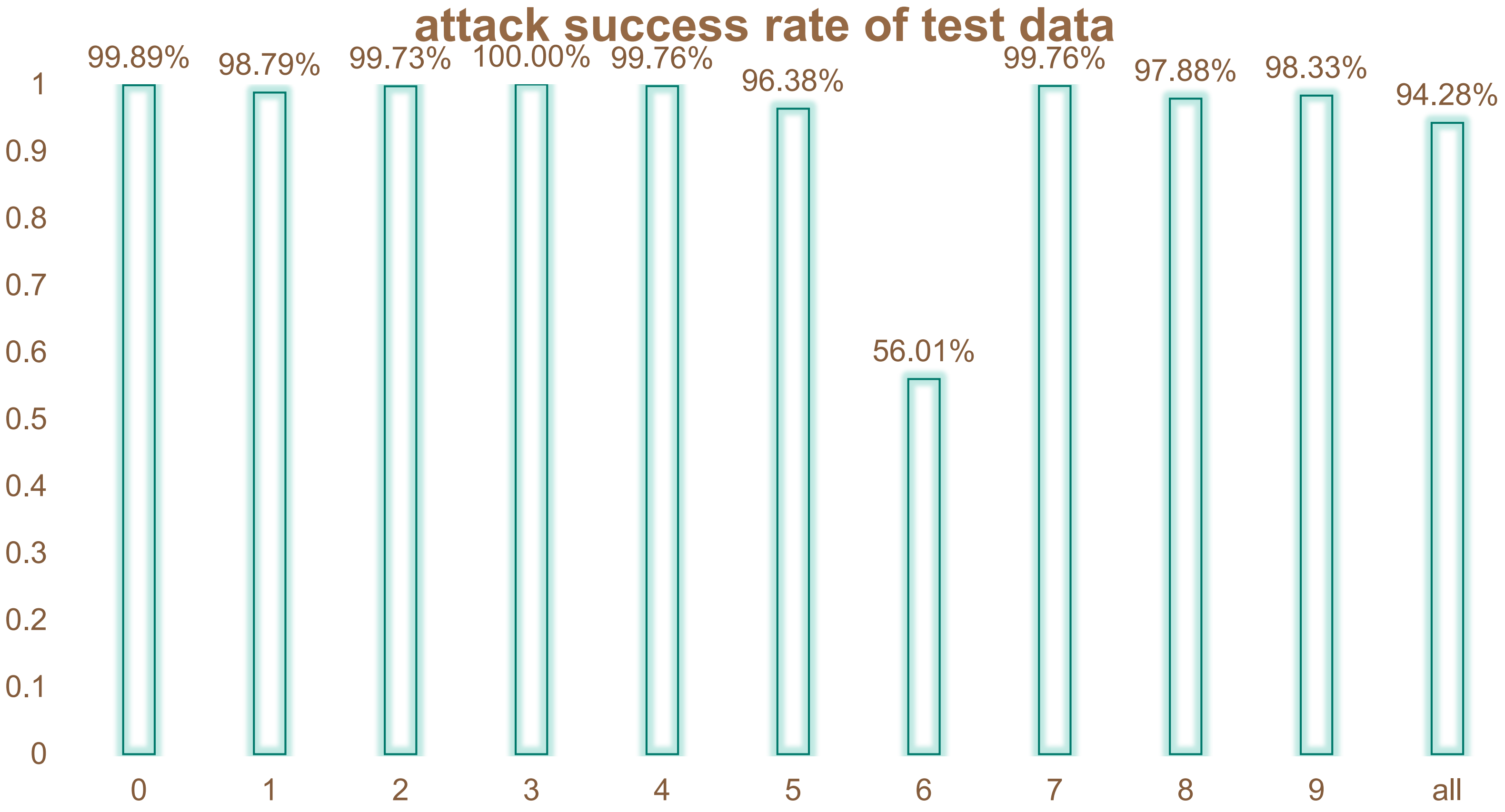
history



train History
D_fake
d_real
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
14
Epoch

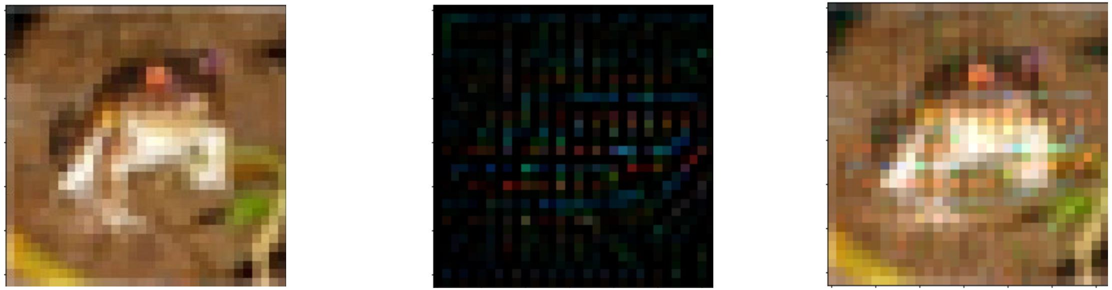


train History
D_acc
attack_acc
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
14
Epoch

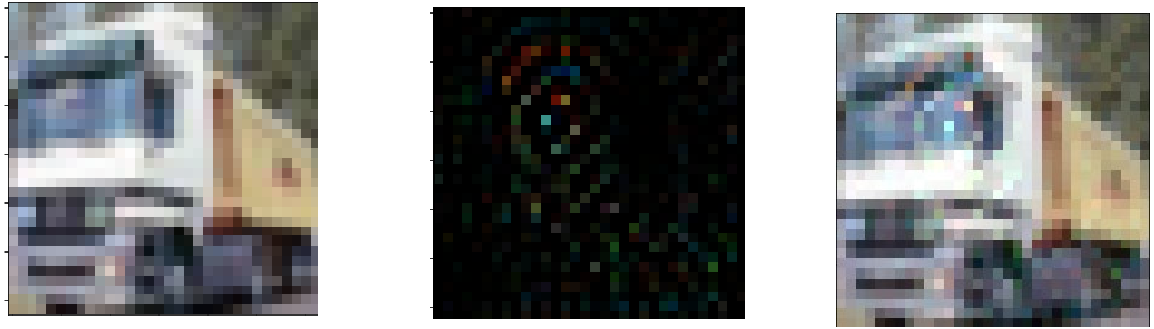


attack success rate of test data

Result



100% ship
2.68e-11% frog



99.99%dog
6.47e-7% truck

Comparison

Attack method	FGSM	PDG	one pixel attack	Adv gan
Black box or White box	White box	White box	Black box	White box
Success rate	69.37%	70.2%	53%	96.24%
Robustness	0.1019	0.0149		0.0967
advantage	Fast	1. Stable success rate 2. Get the less perturbation to mis-lead the model	Do not need the model's detail	1.high success rate 2.Easy to use

Thank you for listening