

PuVAE

Nao Rho, Fan Yang, Junran Yang

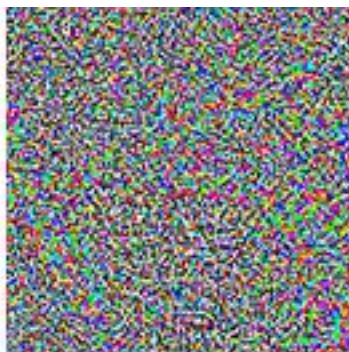
Introduction



"panda"

57.7% confidence

+ ϵ



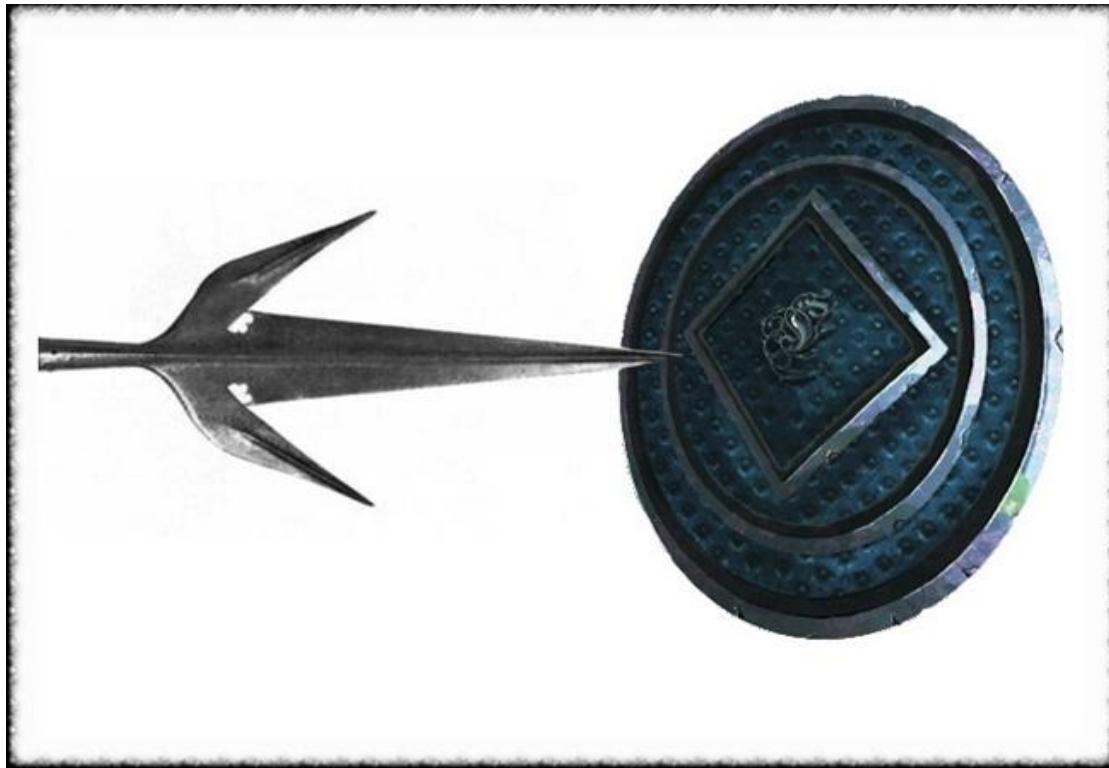
=



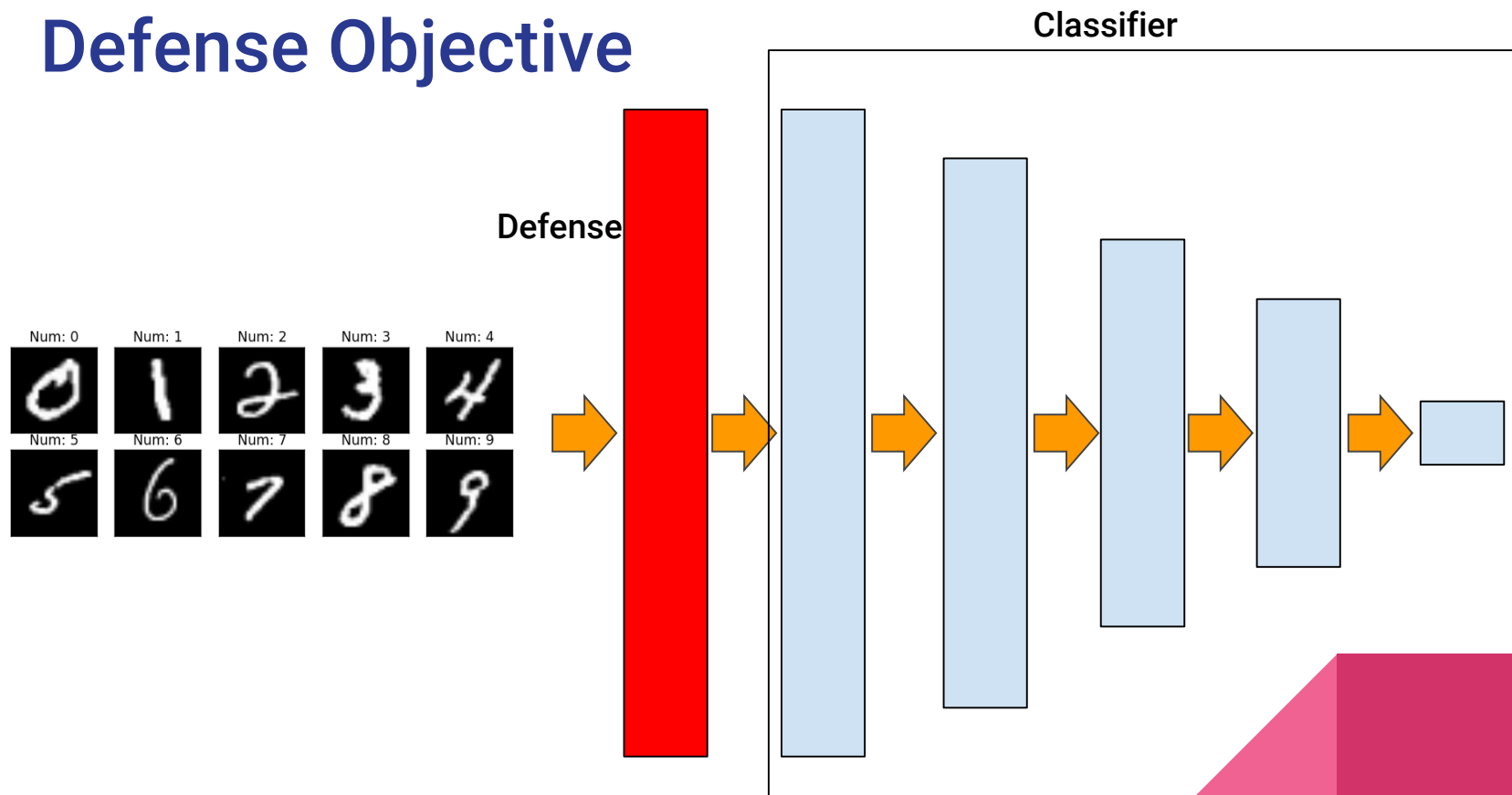
"gibbon"

99.3% confidence

Attack vs Defense



Defense Objective



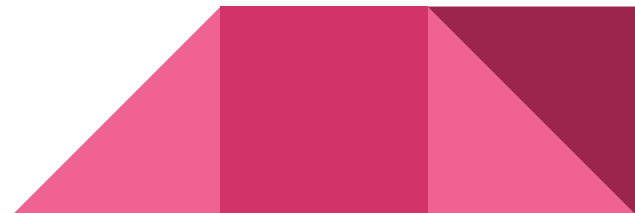
Attack and Defense

PUVAE

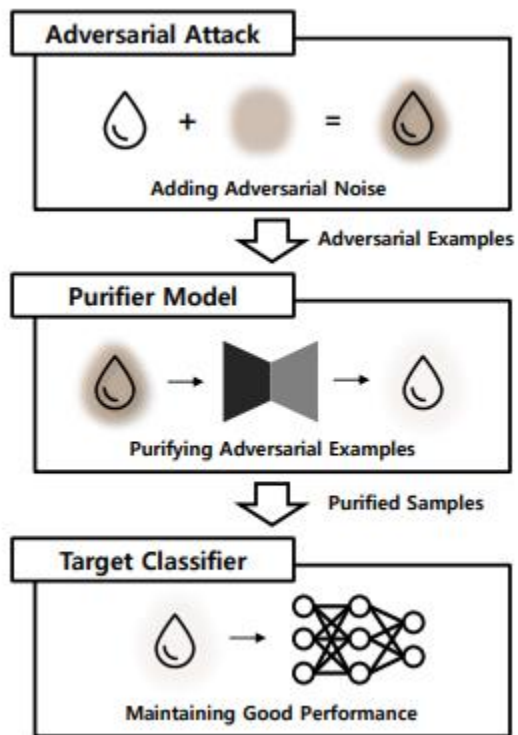
VS

advGAN

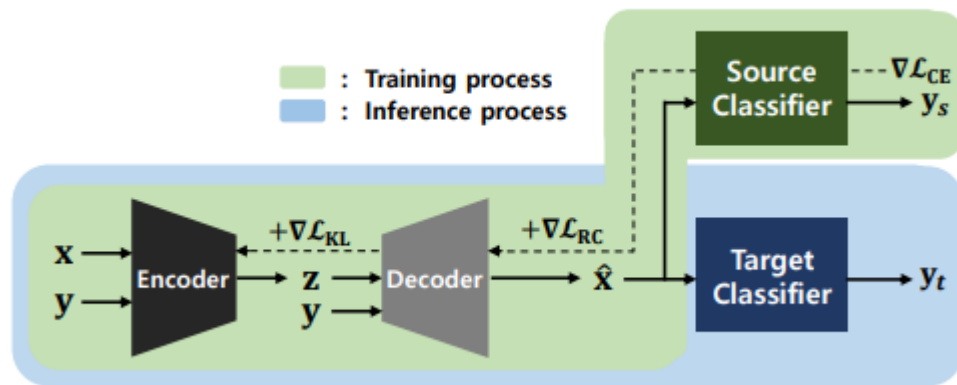
defenseGAN



PuVAE1



PuVAE2



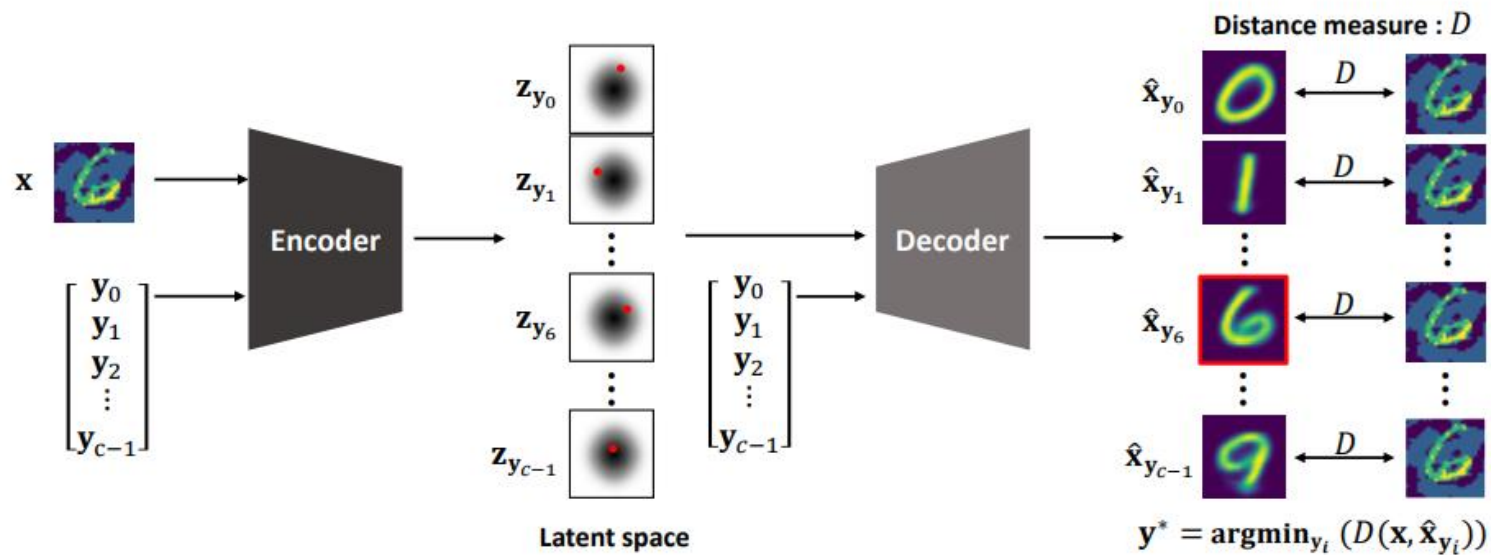
$$\mathcal{L}_{RC} = \mathbf{x}_{\text{data}} \log \hat{\mathbf{x}} + (1 - \mathbf{x}_{\text{data}}) \log(1 - \hat{\mathbf{x}})$$

$$\mathcal{L}_{KL} = \mu^2 + \sigma^2 - \log(\sigma^2 - 1)$$

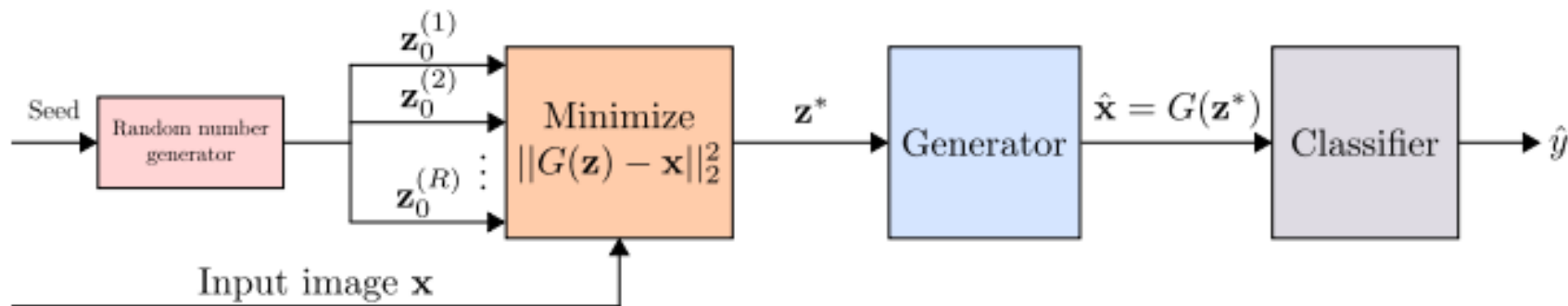
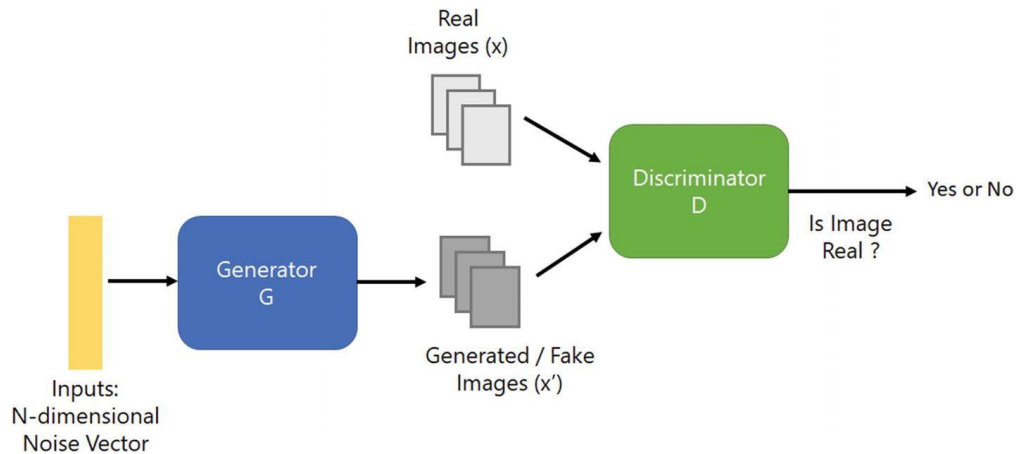
$$\mathcal{L}_{CE} = \mathbf{y}_{\text{data}} \log \mathbf{y}_s + (1 - \mathbf{y}_{\text{data}}) \log(1 - \mathbf{y}_s)$$

$$\nabla(\lambda_{RC} \mathcal{L}_{RC} + \lambda_{KL} \mathcal{L}_{KL} + \lambda_{CE} \mathcal{L}_{CE})$$

PuVAE3



defenseGAN1



defenseGAN2

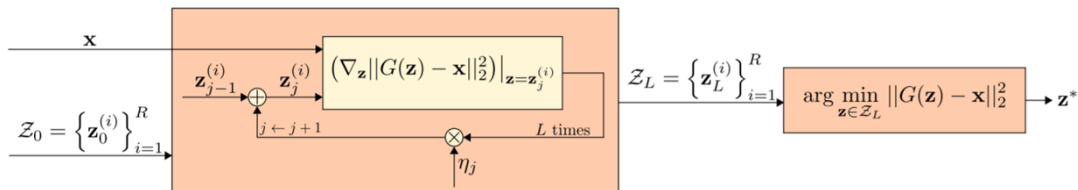
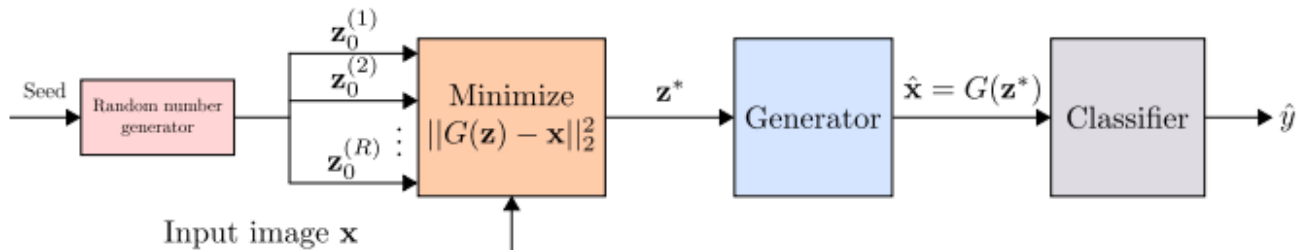
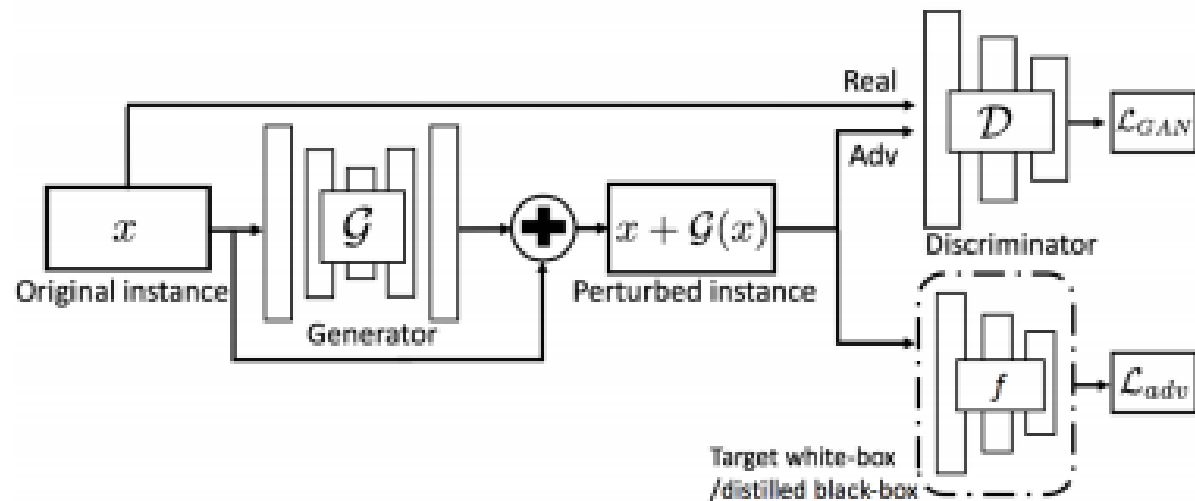


Figure 2: L steps of Gradient Descent are used to estimate the projection of the image onto the range of the generator.

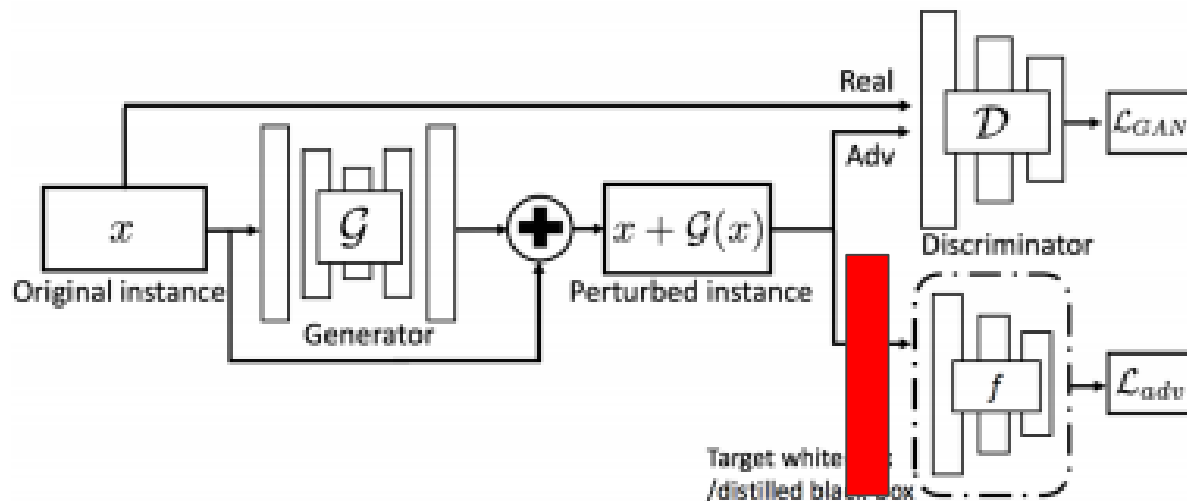


advGAN1



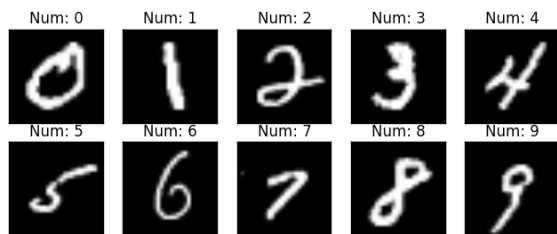
$$\mathcal{L} = \mathcal{L}_{adv}^f + \alpha \mathcal{L}_{GAN} + \beta \mathcal{L}_{hinge}$$

Experiment 2 - White Box Attack

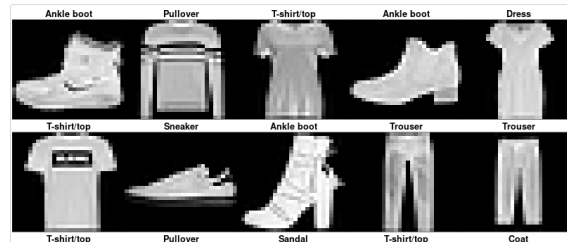


test data set

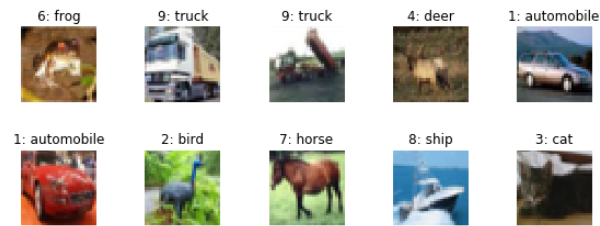
MNIST



Fashion MNIST



CIFAR10



advGAN2



Buckeye

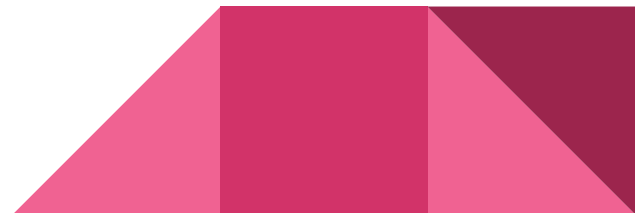
Toy poodle

Perturbation bound = 0.2 on a $[0, 1]$ scale

Baseline Result

MNIST	Fashion MNIST	CIFAR10
99.5	92.4	88.6

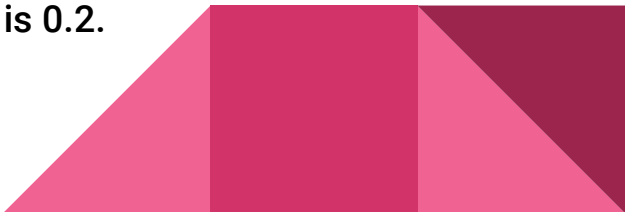
Table 1. Accuracy of target classifiers under no attack



Black Box Result

	No defense	PuVAE defense	defenseGAN
MNIST	56.6	17.2	34.0
FashionMNIST	65.2	22.5	54.8
CIFAR10	23.3	16.0	20.7

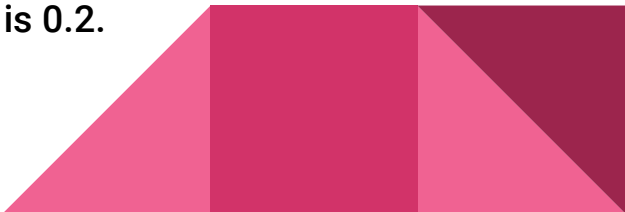
Table 2. Attack success rates on target classifiers under black box setting. All attacks were untargeted. Perturbation bound is 0.2.



White Box Result

	No defense	PuVAE defense	defenseGAN
MNIST	93.8	74.1	90.6
FashionMNIST	99.1	92.2	95.3
CIFAR10	96.7	91.1	87.5

Table 2. Attack success rates on target classifiers under white box setting. All attacks were untargeted. Perturbation bound is 0.2.



References

Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models, *ICLR***2018**

Pouya, Samangouei Maya, Kabkab Rama Chellappa

Generating Adversarial Examples with Adversarial Networks.

Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, Dawn Song

PuVAE: A Variational Autoencoder to Purify Adversarial Examples, *ArXiv***2019**

Uiwon Hwang, Jaewoo Park, Hyemi Jang, Sungroh Yoon, Nam Ik Cho



Any question?

