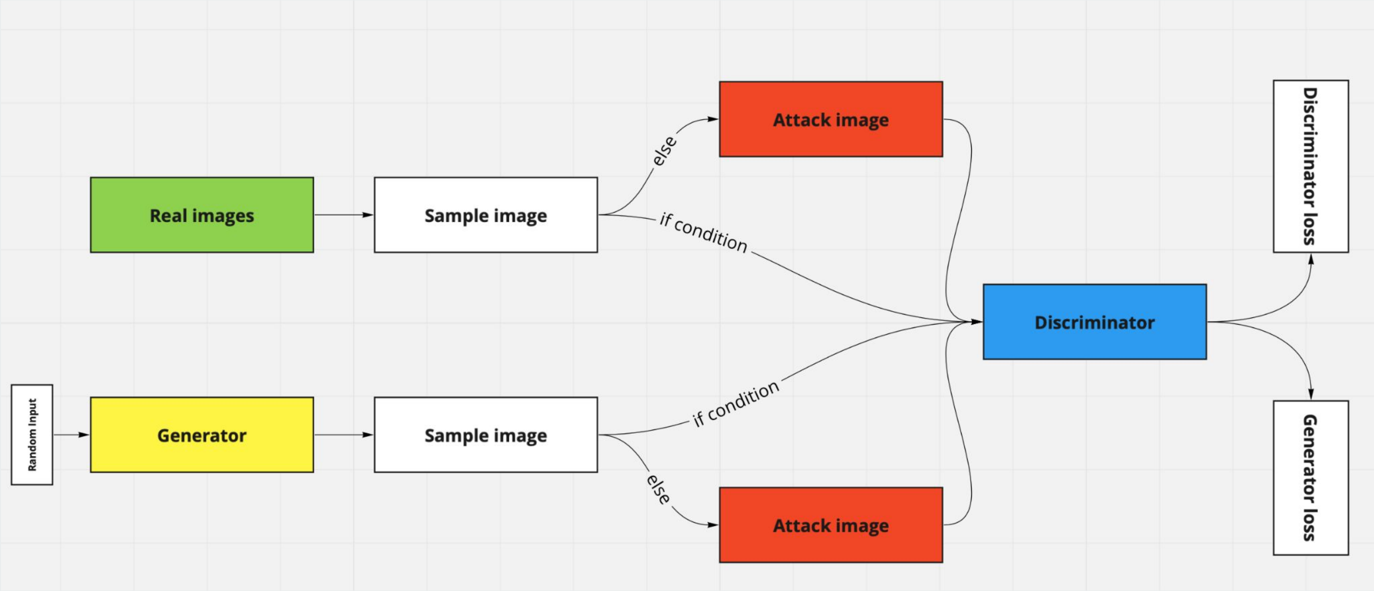
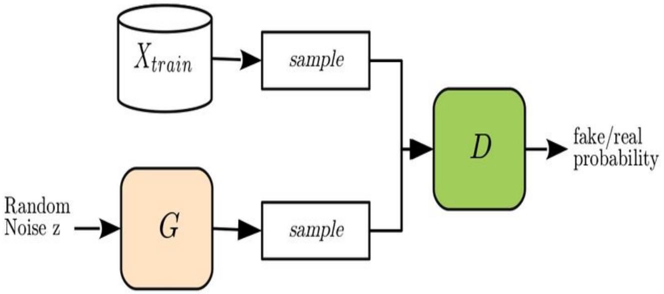
Training Generative Adversarial Networks with Adversarial Attacks



Coursework by Slava Pirogov, HSE AMI 2022
Supervised by Alanov Aibek, HSE visiting lecturer

Generative Adversarial Networks

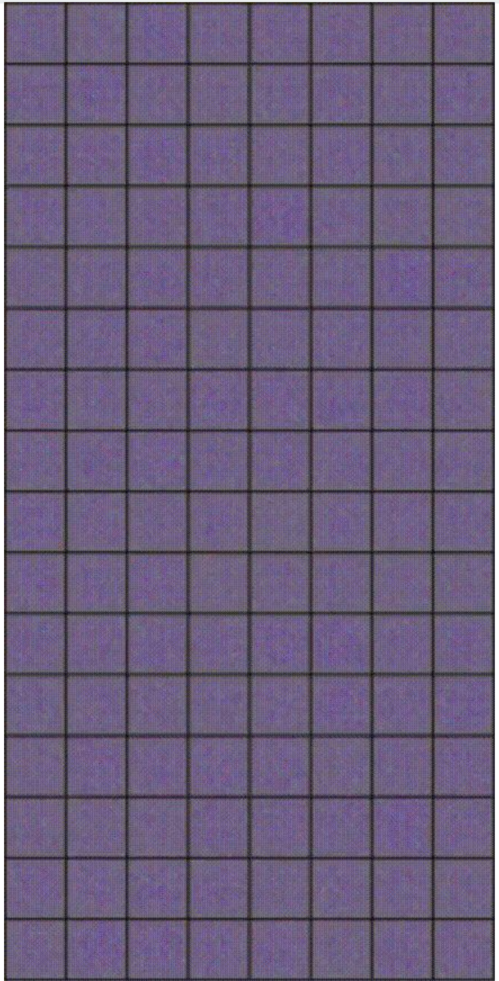
- Generative architecture, adversarial process
- Generator (G) and Discriminator (D)
- G aims to capture the distribution of the dataset
- D aims to estimate the probability that a sample came from the training data rather than G
- Minimax problem with value function V (G, D):



$$\min_G \max_D V(G, D) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$

Relevance of the task

- Training with Adversarial Attacks can be applied to any GAN
- GANs are still popular (StyleGAN 2019, StyleGAN3 2021, more than 700 papers published in 2022 on arxiv with word GAN in the abstract)
- Vanishing Gradients (research)
- Only one article! (Rob-GAN: Generator, Discriminator, and Adversarial Attacker by Liu and Hsieh (2019))



Goal and tasks

- Goal: explore different ways of building GANs and compare them with GANs that have been trained using Adversarial Attacks (first of all in terms of quality)
- Tasks:
  - Realization of few Adversarial Attacks methods on multiple datasets
  - Realization of some popular GANs, calculation and comparison of key metrics on CIFAR-10 dataset
  - Development of GAN Adversarial training theory, and implementation of it with different GANs and hyperparameters.

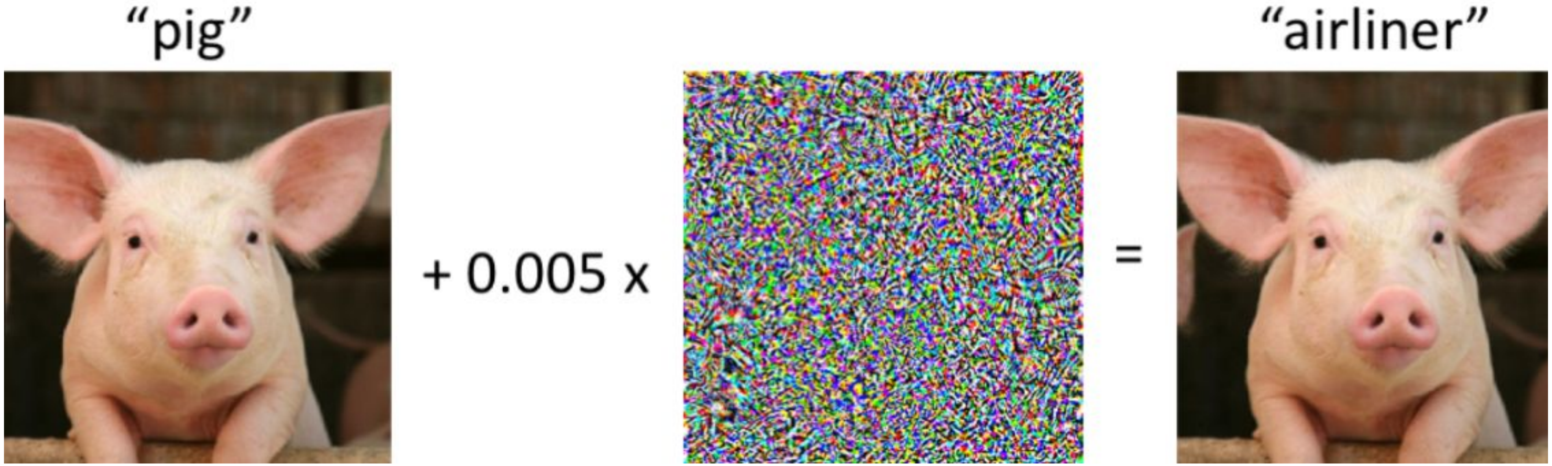
FGSM attack

J (θ, x, y) represents the loss of the network

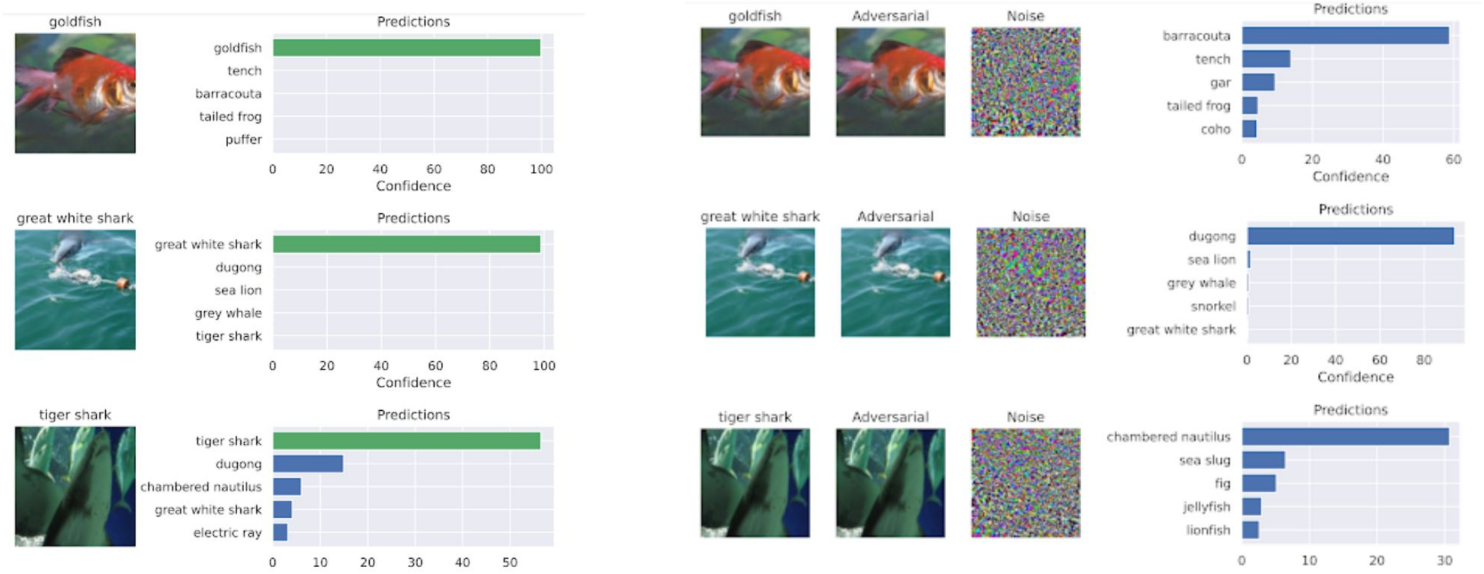
ε is the intensity of the noise

$\tilde{x}$ the final adversarial example

$$\tilde{x} = x + \epsilon \cdot \mathrm{sign}(\nabla_x J(\theta, x, y))$$



FGSM attack on ImageNet



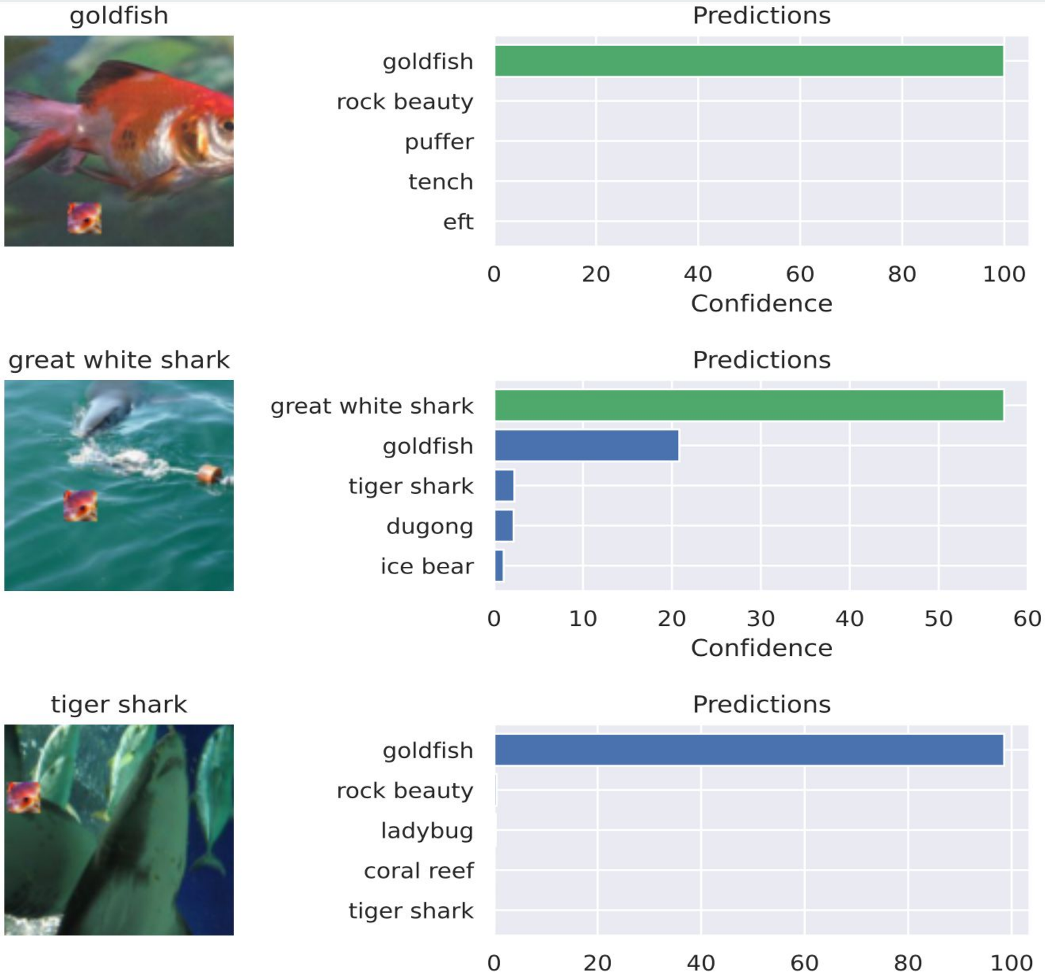
(a) dataset images

(b) FGSM images

Example of FGSM attacks on ImageNet

Adversarial Patches

Example of Adversarial Patches on ImageNet

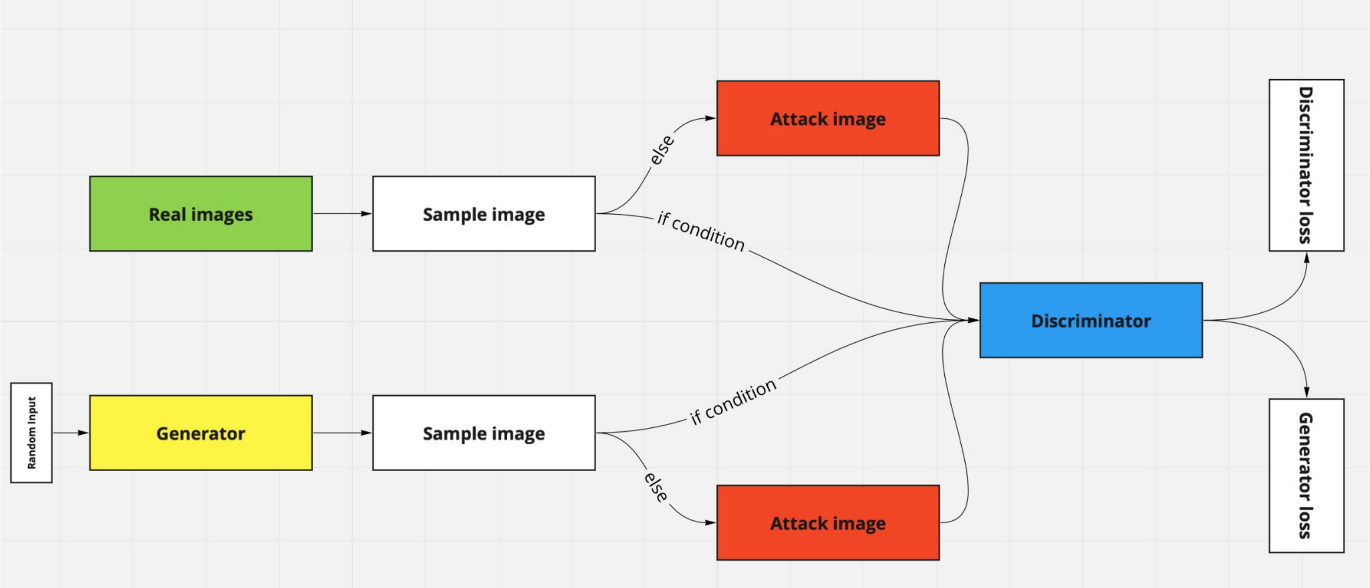


GANs

Model	Dataset	Inception Score ⇧	FID ⇩
our DCGAN	CIFAR10	6.40(0.06)	41.42
DCGAN	CIFAR10	6.26(0.06)	41.92
our WGAN-GP(CNN)	CIFAR10	7.71(0.11)	18.67
WGAN-GP(CNN)	CIFAR10	7.66(0.10)	19.83
our WGAN(CNN)	CIFAR10	6.00(0.08)	48.38
WGAN(CNN)	CIFAR10	6.62(0.09)	40.03
our SNGAN(CNN)	CIFAR10	7.76(0.13)	18.38
SNGAN(CNN)	CIFAR10	7.84(0.12)	17.81

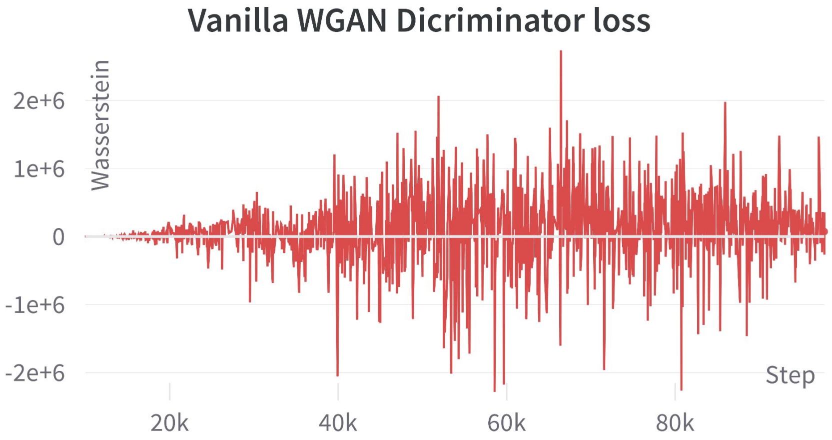
- DCGAN by Alec Radford (2015)
- SNGAN by Takeru Miyato (2018)
- WGAN by Martin Arjovsky (2017)
- WGAN-GP by Ishaan Gulrajanj (2017)

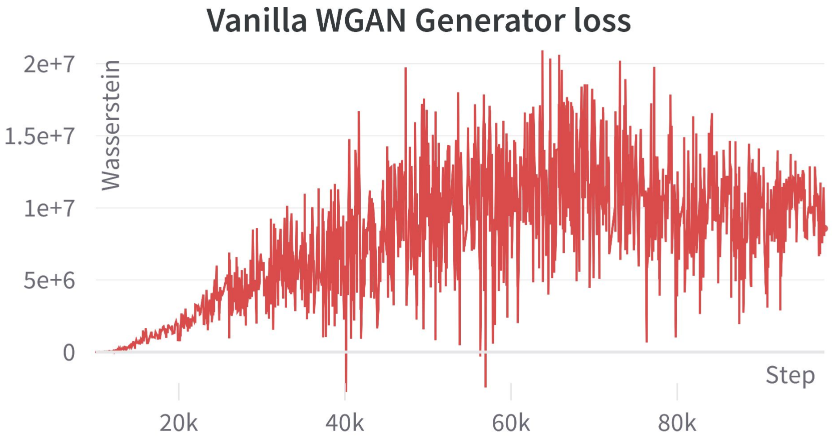
Theory of GAN Adversarial Training



- Start of attacks from 10% of epochs
- Chance to attack C
- ε in FGSM

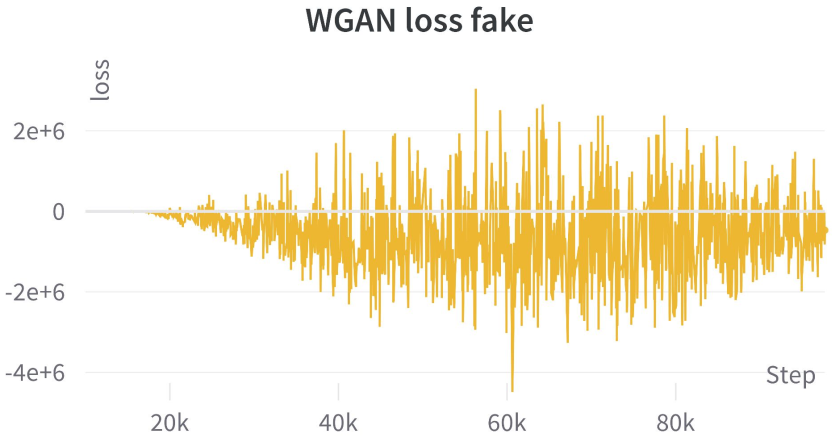
Theory of GAN Adversarial Training

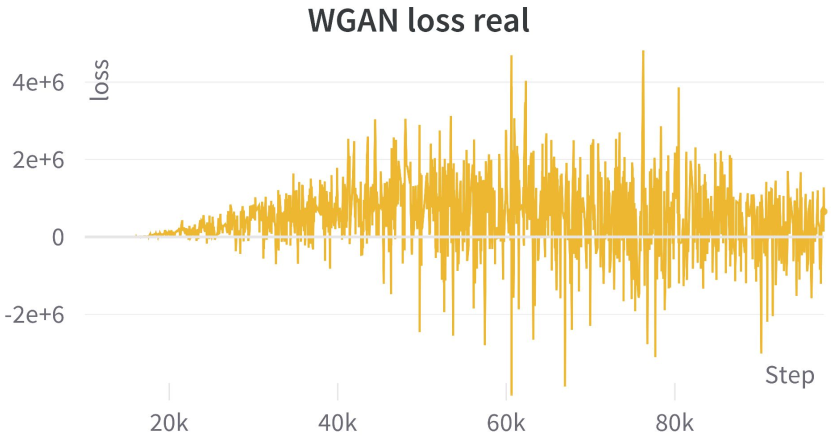




Monitor robustness and stability of the architecture

Theory of GAN Adversarial Training



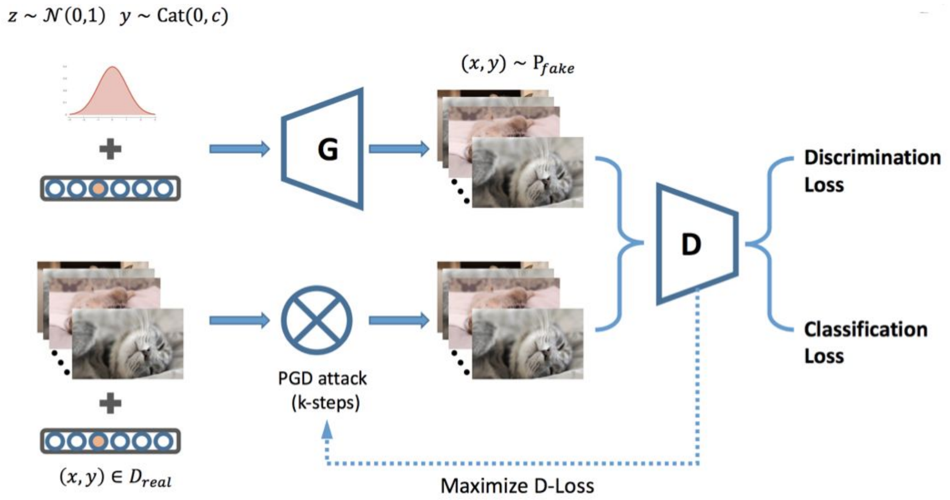


Split the Discriminator loss into real and fake parts

Related work

Rob-GAN: Generator, Discriminator, and Adversarial Attacker by Liu and Hsieh ([2019](#))

- Research about convergence speed of GAN training and the robustness of Discriminator
- Projected Gradient Descent attacks
- Auxiliary Classifier GAN
- Attack at every step



Experiments

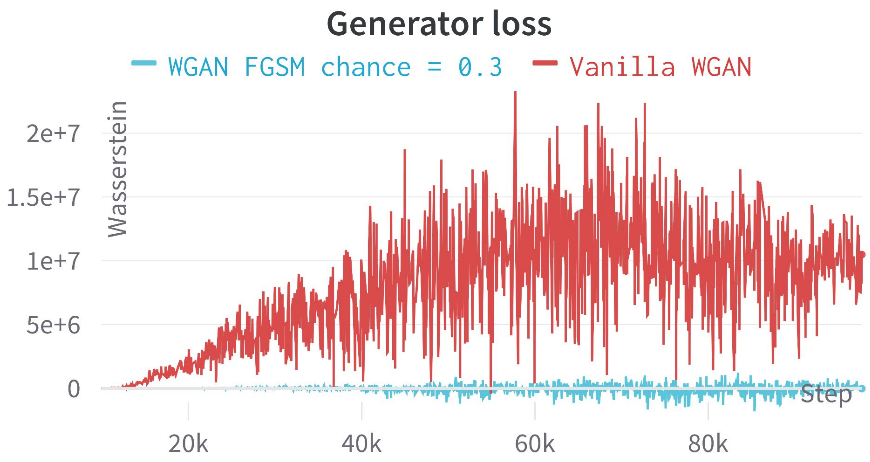
- CIFAR-10
- 1xV100 and 8xCPU
- Default ε = 0.02
- Left - real. Right - FGSM

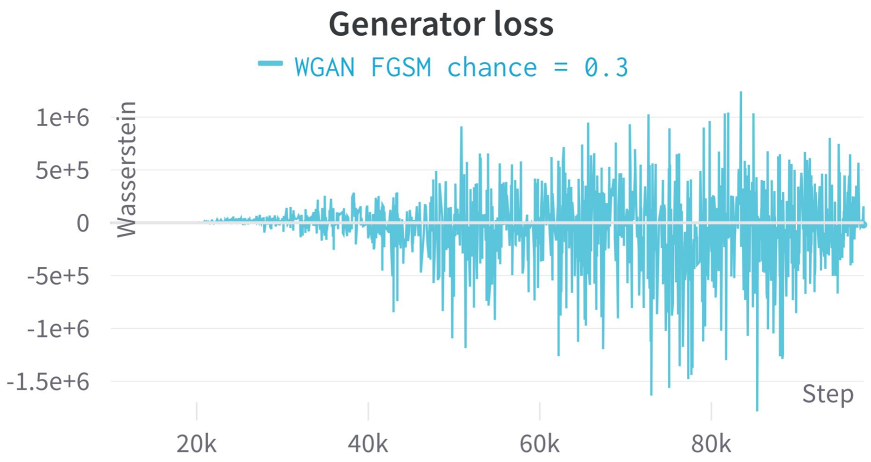
Experiments - WGAN

Model	FGSM chance	ϵ	Inception Score ⇧	FID ⇩	Time (min)
Baseline WGAN	–	–	6.00(0.09)	48.38	**502**
WGAN-FGSM	0.2	0.02	6.58(0.09)	35.21	538
WGAN-FGSM	0.3	0.02	6.53(0.06)	**33.60**	537
WGAN-FGSM	0.4	0.02	6.73(0.10)	35.56	595
WGAN-FGSM	0.3	0.01	**6.77(0.07)**	33.78	537

- IS improved by 10%
- FID improved by almost 30%
- Over 25 full experiments

Experiments - WGAN





Stabilizing of G loss

Red - vanilla version

Blue - FGSM version

Experiments - WGAN-GP

Model	FGSM chance	Inception Score	FID	Time (min)
Baseline WGAN-GP	–	**7.71(0.11)**	**18.67**	**613**
WGAN-FGSM	0.2	7.64(0.09)	19.97	645
WGAN-FGSM	0.4	3.35(0.03)	106.6	673
WGAN-FGSM	0.6	3.51(0.03)	112.57	693
WGAN-FGSM	0.8	4.19(0.07)	97.37	721

Several options for FGSM-attack on WGAN-GP

Only one explored

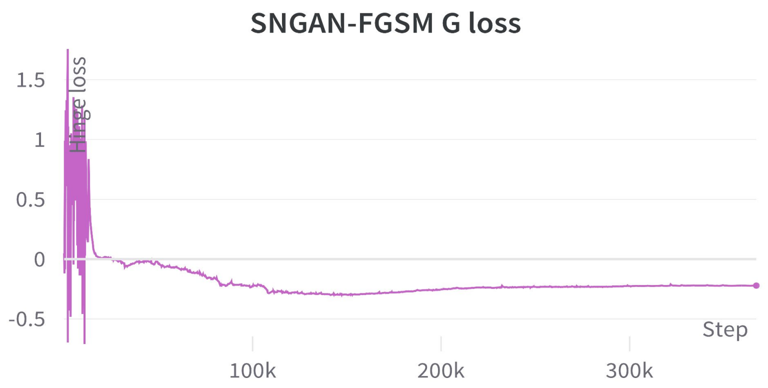
Experiments - DCGAN

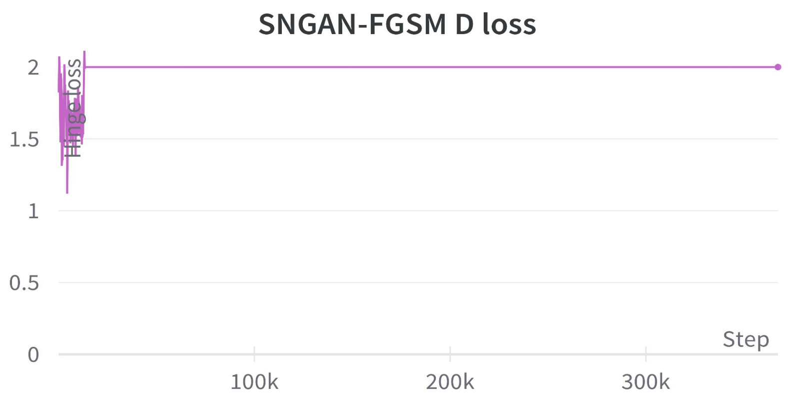
Model	FGSM chance	Inception Score ⇧	FID ⇩	Time (min)
Baseline DCGAN	–	6.40(0.06)	41.42	**590**
DCGAN-FGSM	0.2	**6.52(0.07)**	40.23	596
DCGAN-FGSM	0.4	6.15(0.09)	59.78	604
DCGAN-FGSM	0.6	6.34(0.04)	**39.10**	627
DCGAN-FGSM	0.8	5.97(0.05)	53.50	642

Metrics, loss behavior are similar

FGSM attacks don't necessarily improve weak GANs

Experiments - SNGAN





Many problems with the model at FGSM chance = 0.6, 0.8

Discriminator loss becomes constant

Decreasing start epoch of attack fix it

However, we learn different distribution

Experiments - SNGAN

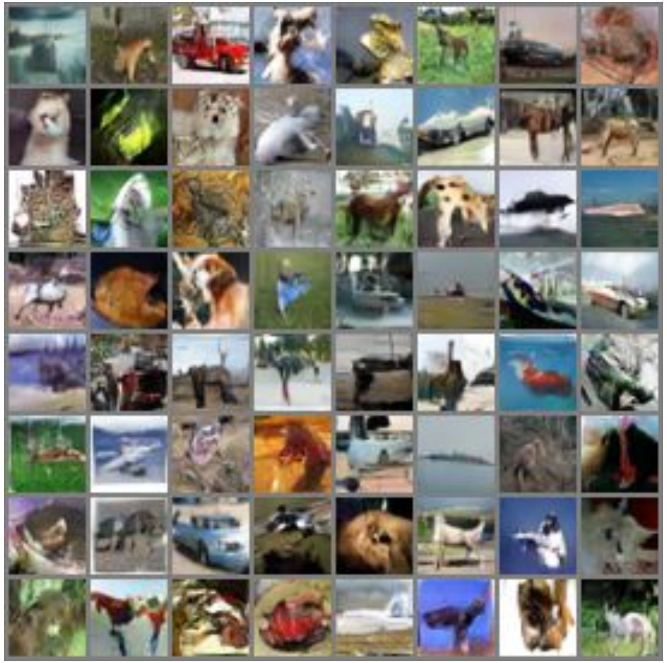
Model	FGSM chance	start FGSM	Inception Score ⇧	FID ⇩	Time (min)
Baseline SNGAN	–	–	**7.84(0.12)**	**17.81**	**503**
SNGAN-FGSM	0.2	10%	7.54(0.13)	19.20	750
SNGAN-FGSM	0.4	10%	7.36(0.03)	22.84	793
SNGAN-FGSM	0.6	5%	6.86(0.05)	28.34	580
SNGAN-FGSM	0.8	0%	6.64(0.08)	33.40	614

First model

Significant results were not achieved

Over 50 experiments were carried out

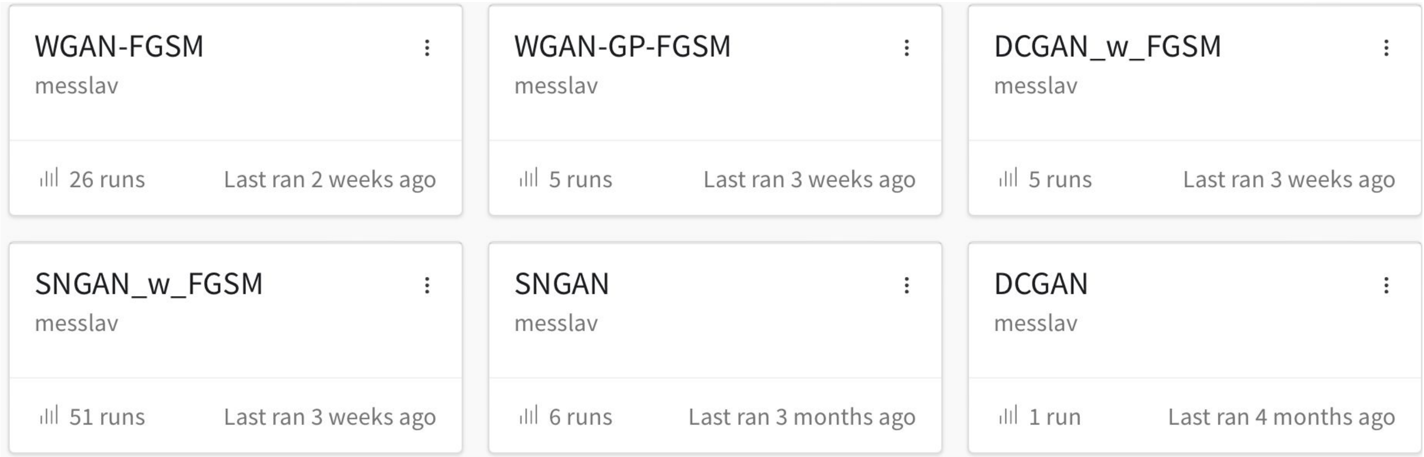
Experiments - SNGAN



Generated samples from SNGAN (left) and SNGAN-FGSM (right).
The same noise was in the input

Program Realization

- Github
- Python 3.7, PyTorch 1.10
- Almost 100 experiments, 30 days of computing resources



Results

- Implemented FGSM, Adversarial Patches attacks on ImageNet, MNIST, CIFAR-10 datasets
- Implemented DCGAN, SNGAN, WGAN, WGAN-GP with logging. Trained them on the CIFAR-10 dataset

★ Implemented GAN FGSM training with DCGAN, SNGAN, WGAN, WGAN-GP. Trained them on the CIFAR-10 dataset

★ Researched how FGSM attacks affects GAN losses and metrics

Acknowledgments

This research was supported in part through computational resources of HPC facilities at HSE University

Kostenetskiy P.S., Chulkevich R.A., Kozyrev V.I. HPC Resources of the Higher School of Economics // Journal of Physics: Conference Series. 2021. Vol. 1740, No. 1. P. 012050. DOI: https://doi.org/10.1088/1742-6596/1740/1/012050