Adversarial Image generation using GANs

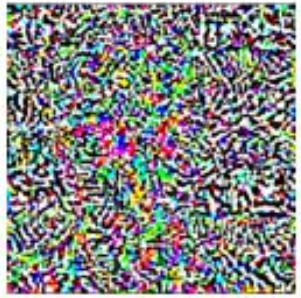
By

- Swaminathan Gurumurthy (sgurumur)
- Bhavan Jasani (bjasani)

What are Adversarial Images?

Input = Yaser

Very small in magnitude



Generated adversarial loss

Still looks like Yaser

Face recognizer

This is Simon

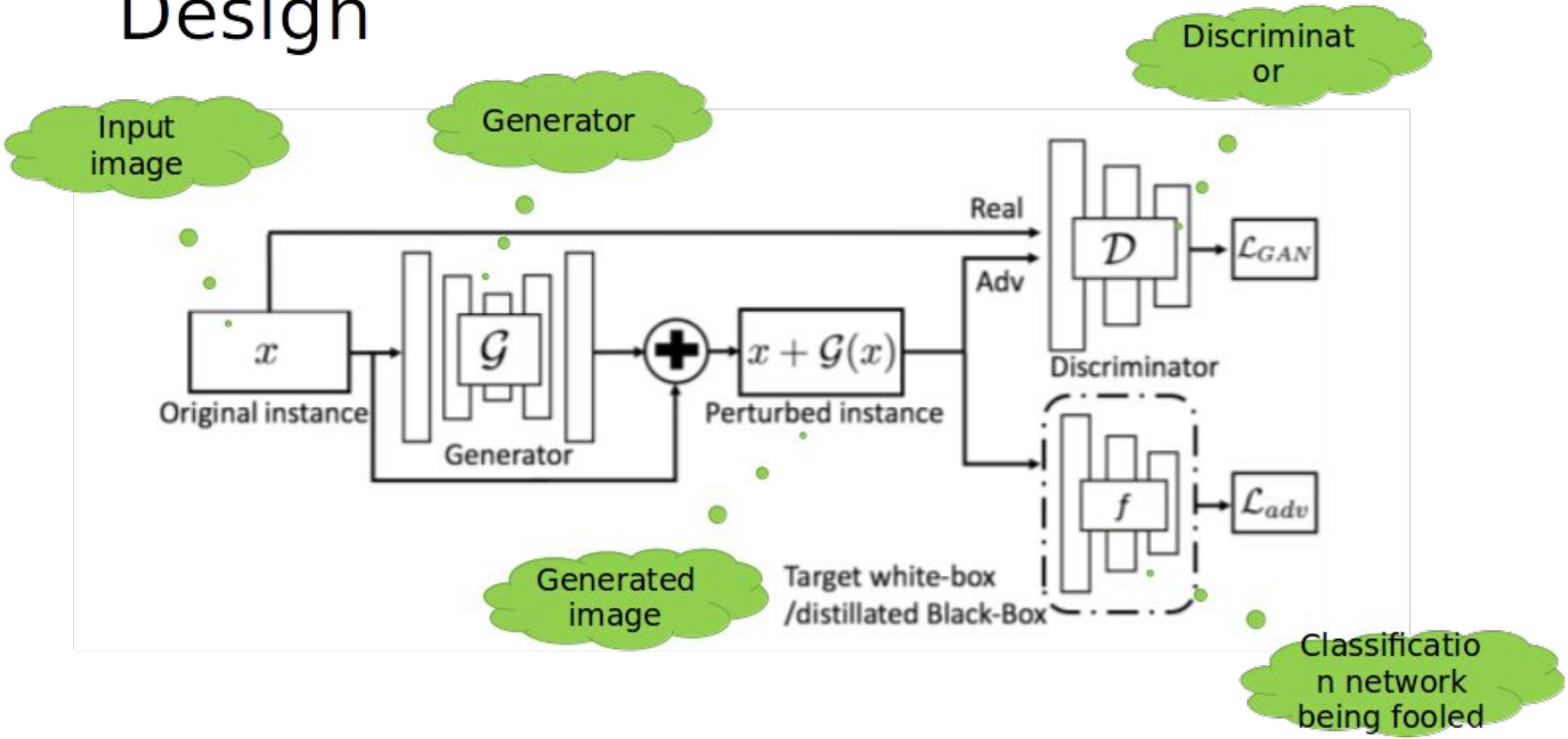
Really????

Implementation

- Based on architecture mentioned in - "***Generating Adversarial examples with adversarial networks***" (Anonymous Authors, ICLR 2018 submissions)
- Implemented and trained using **TensorFlow** on **CIFAR10** dataset
- Generates CIFAR10 adversarial images

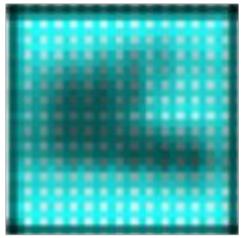
TensorFlow

Design


Discriminat or
Input image
Generator
Real
Adv
$\mathcal{D}$
$\mathcal{L}_{GAN}$
x
$\mathcal{G}$
$x + \mathcal{G}(x)$
Discriminator
Original instance
Generator
Perturbed instance
f
$\mathcal{L}_{adv}$
Generated image
Target white-box /distillated Black-Box
Classificatio n network being fooled

Results

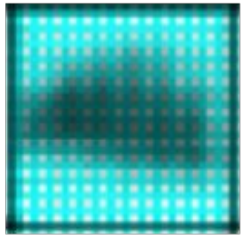
Bird



Noise

Cat

automobile



Noise

Bird

	FGSM	Carlini-Wagner	AdvGAN
time	~0.06s	~30s	~0.01s