Bypassing malware detectors with generative adversarial networks (GAN)

Kondah Mouad

University of Geneva

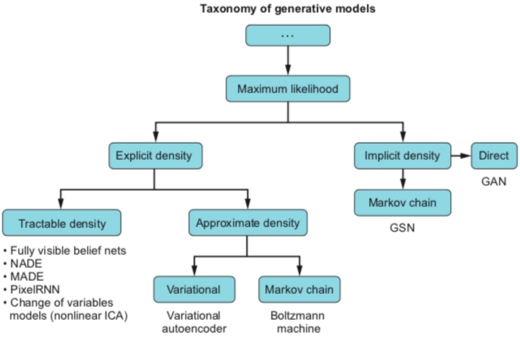
November 6, 2019

Plan

- Adversarial Deep Learning
  1. Generative Modeling ,Distribution Approximation
  2. GAN
- Bypassing malware detectors MalGAN

Distribution Approximation

- can we build a model to approximate a data distribution ?
- can we find $p_{model}(x;\theta) \sim p_{data}(x)$
- Maximizing Likelihood-VAE(here we learn some prior)
- Generate samples directly -GAN

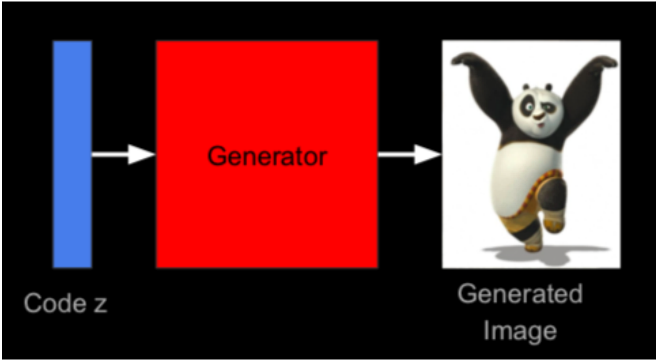


Representation Learning

- generate new samples follows the same probabilistic distribution of a given a training dataset
- the generator has a prior $p_\theta(z)$ and for maps each z to the observation space.

Generative models

- generate new samples follows the same probabilistic distribution of a given a training dataset
- the generator has a prior $p_\theta(z)$ and for maps each z to the observation space.

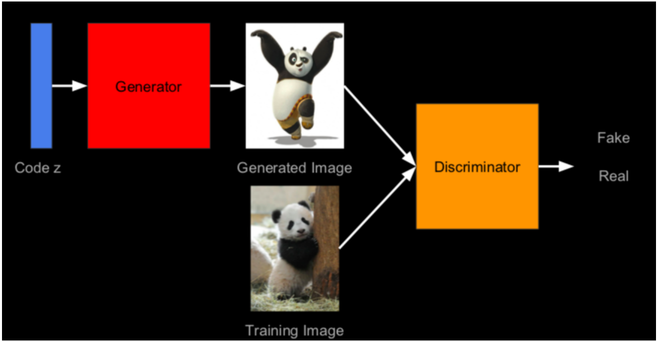


Generative models: mathematic formulation

Value of　　Expectation　　prob. of D(real)　　prob. of D(fake)

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

Minimize G　　Maximize D　　x is sampled from real data　　z is sampled from N(0, I)　　fake



GAN

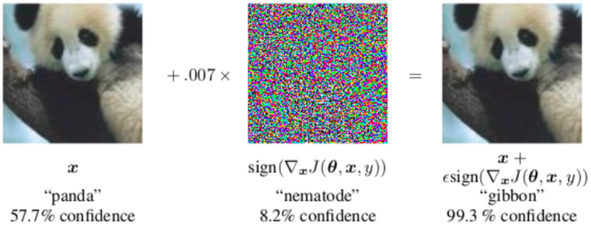
- Generative adversarial networks are composed of two parts: a generator and a discriminator.
- we generate from a latent space so that we map each z into the observation space that follow the distribution of the real data
- the generator is basically a neural network that we train in parallel of the discriminator.
- finding Nash equilibrium is challenging(D_x and G_θ needs to be good at the same time !)
- Very often we encounter a mode collapse, the latent space is mapped towards restricted space.
- when the two distribution are not disjoint it's hard to give a good measure of distance.

Malware detection

- Detecting Malware is still a challenge for many information security professional.
- Information security professionals are doing their best to come up with novel techniques to detect malware and malicious software

Direct gradient-based attacks

- perturbing the sample x in the direction that would most decrease the score.



Learn this sophisticated sample

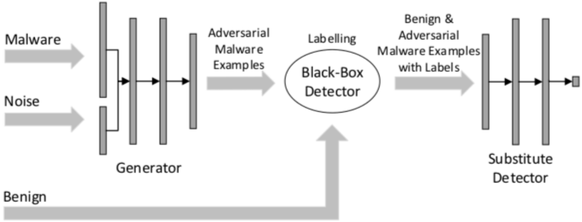


Figure 1: The architecture of MalGAN.

API Feature

1. We construct an M binary dimensional vector where the each value is set to 1 if the corresponding API is called by the program.
2. Program that call the WriteFile API only, $M = (1, 0, 0, 0..., 0)$.
3. Last layer we use Sigmoid function and binarization

Algorithm :

- While not converging do:
  1. Sample a minibatch of Malware M
  2. Generate adversarial samples M' from the generator
  3. Sample a minibatch of Goodware B
  4. Label M' and B using the detector
  5. Update the weight of the detector
  6. Update the generator weights

THANKS!