Fooling Image Classifier for Breast Cancer Detection

Nitin Pathania

Generative Adversarial Networks (GANs)

A new type of ML system (invented in 2014) that rely on two neural networks contesting in a game. Can be used to generate realistic images and videos of virtually anything.

Applications of GANs

AI-generated art

Human image synthesis

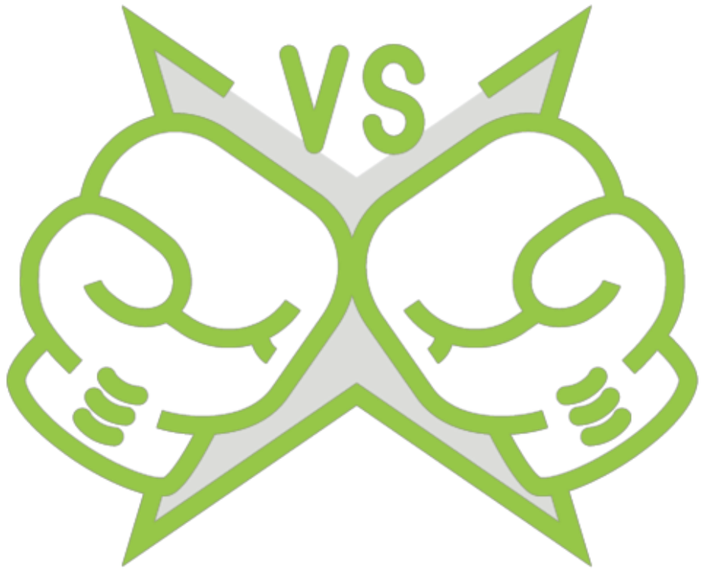
3D models from 2D images

Improve astronomical image

Traditional classification

Traditional regression

GANs



Considered part of unsupervised learning

Developed in 2014 by Ian Goodfellow et al

Two distinct contesting neural networks

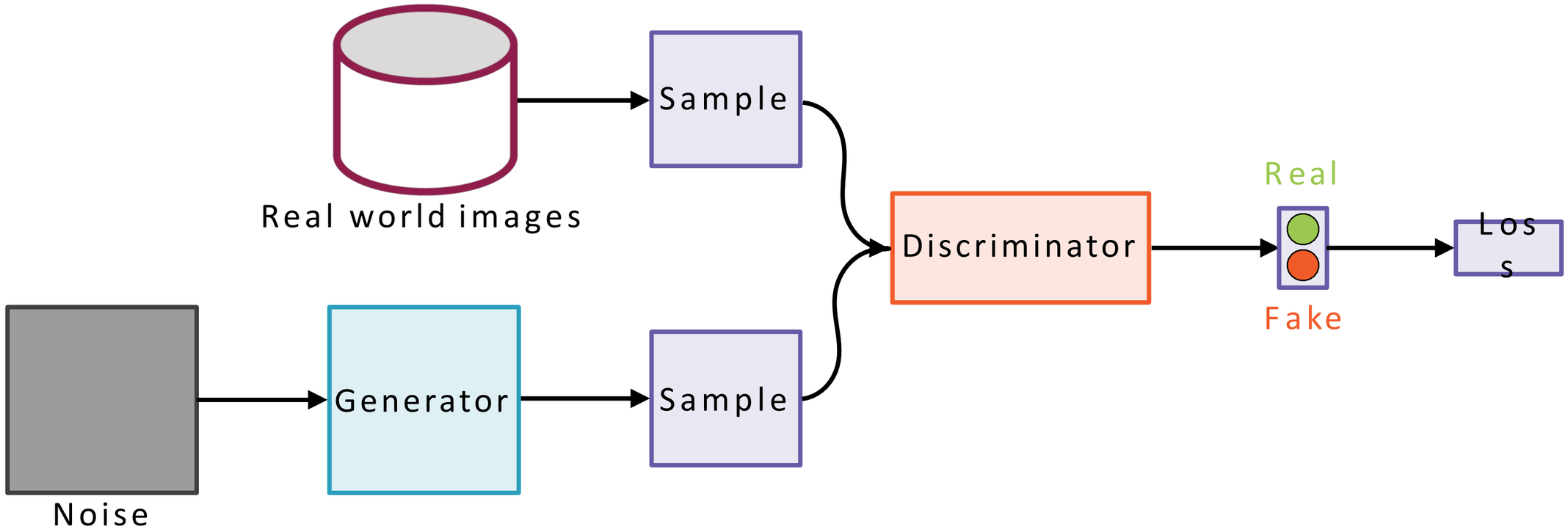
- Generative network
- Discriminative network

Two Neural Networks

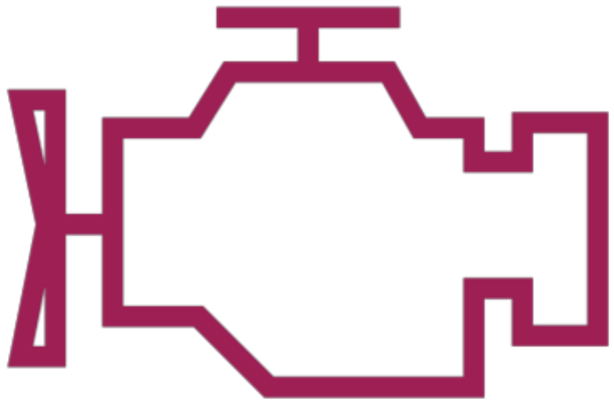
Generative Network
generates candidates

Discriminative Network
evaluates candidates

GANs


Real world images
Sample
Noise
Generator
Sample
Discriminator
Real
Fake
Los
s

Generator

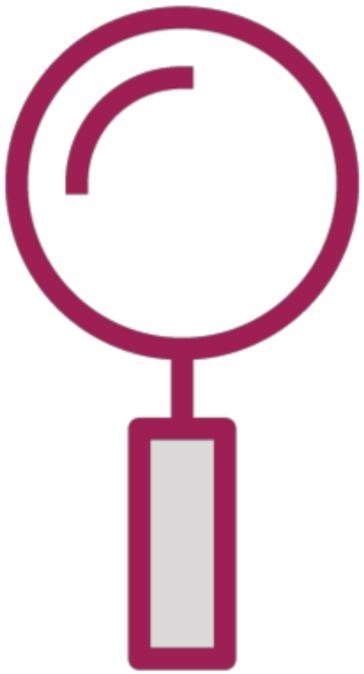


Generates data as realistically as possible

Trained to generate data similar to corpus

Seeks to fool discriminator

Discriminator

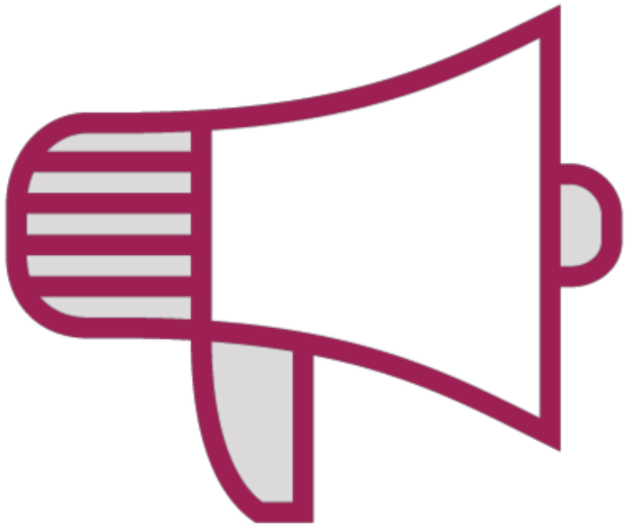


Generates probability that data is genuine

Classifies output of generator

Just like traditional classifier

Noise in GANs

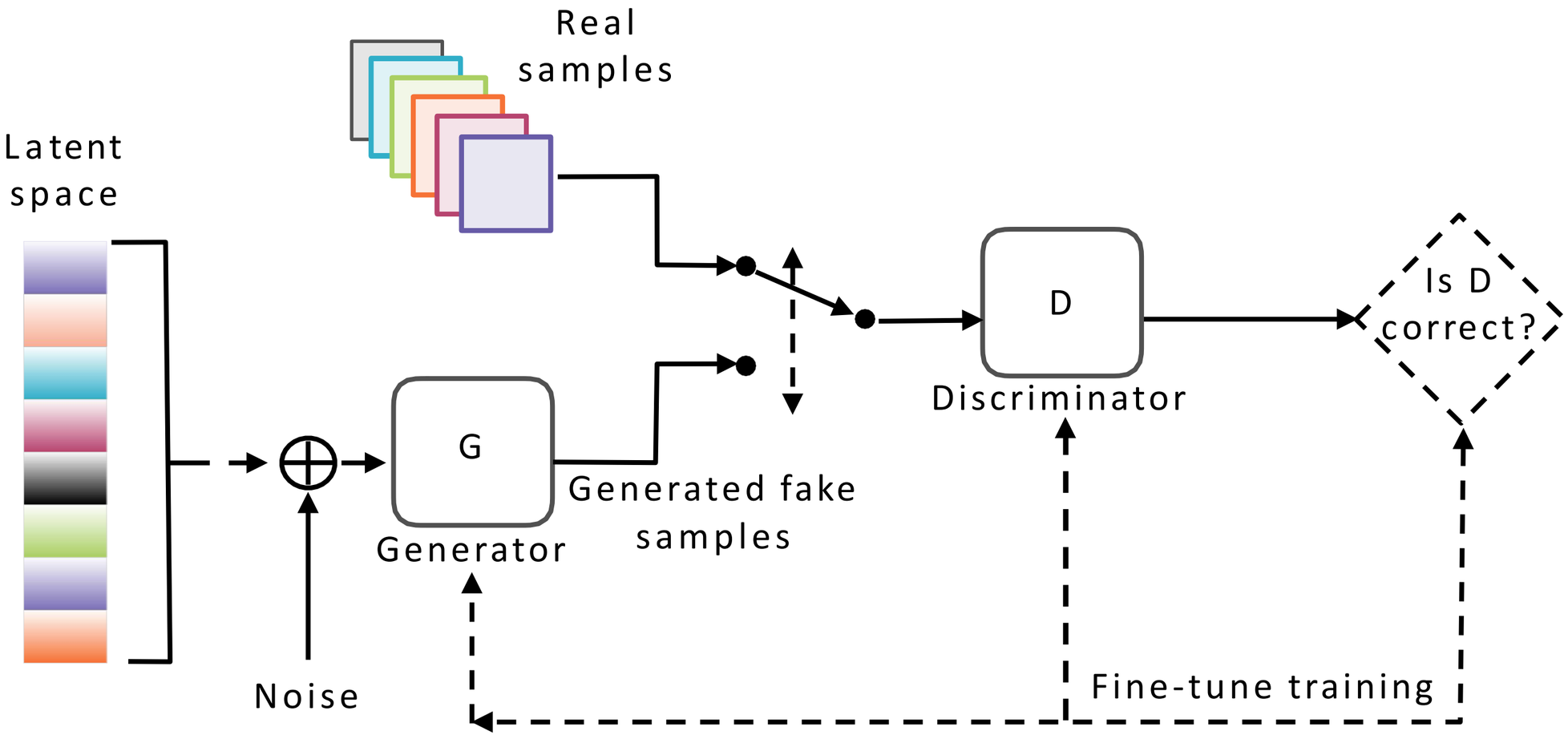


Requires function that generates noise

Create corpus of

- Real data points
- Noise function

GANs


Real
samples
Latent
space
G
Generator
Generated fake
samples
D
Discriminator
Is D
correct?
Noise
Fine-tune training

Training a GAN

Start with corpus of real points as well as noise

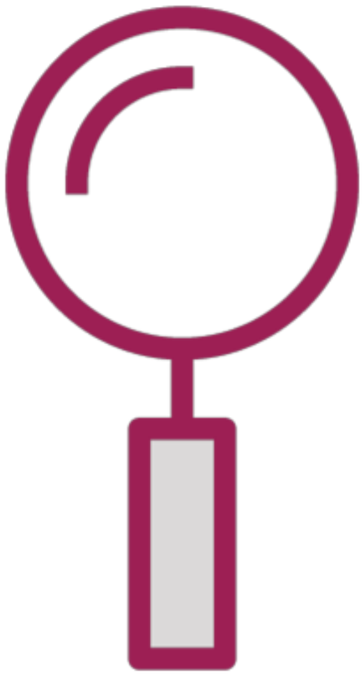
Train discriminator to tell them apart

Generate new noise points

Train generator to produce data that fools the discriminator

Repeat using optimizer

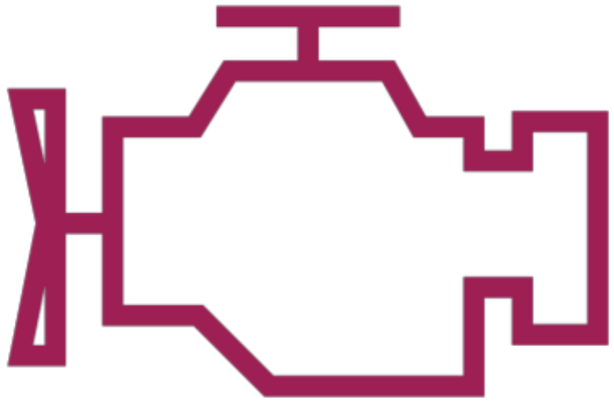
Discriminator



Maximizes probability of real data being classified as real

Minimizes probability of fake data being classified as real

Generator



Maximizes probability of fake data being classified as real

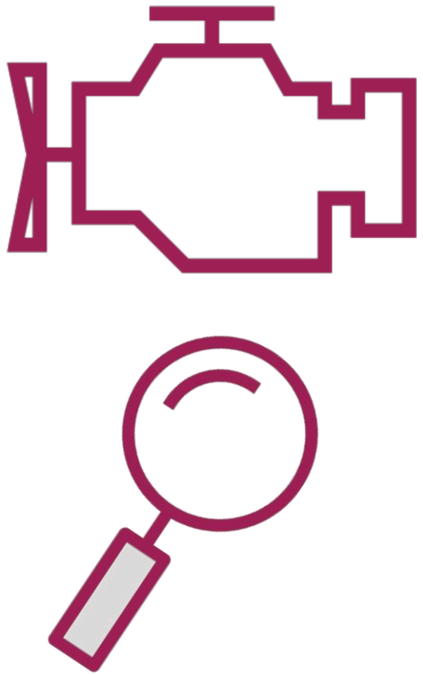
Loss Functions

Need optimizers for both networks

Loss function used is Binary Cross-Entropy (BCE) Loss

Used to heavily penalize incorrect classifications

Generator and Discriminator



Adversaries during training

At some point generator will generate realistic data

Consistently fool the discriminator

Using Generative Adversarial Networks (GANs) to generate histopathology images.

Thank You