# hearing the voice

# INFINITY

2017-2018

# Voice-Image Synthesiser

Asariri

# Asariri

# WHAT

## The Problem

Build a deep learning model that captures the voice fingerprint along with the user  auxiliary features like face, posture etc., and use that to synthesis new machine voice given a picture or recreate the auxiliary features given a voice recording.

# WHY

## The Purpose

As of now the use cases are fuzzy, following are few that popped-in  our head:
- We might end up with a model which can synthesise a voice for given picture
- Data  augmentation: creating new data set from the history of past recordings which are close to real world recordings, which then can be used to train other models
- When  above idea gets materialised, it can be wrapped as a fun mobile application to reach the end user, where we can collect more realistic data
- Crime Investigation

# HOW

## The Solution

- Collecting Video-Audio data on particular personalities
- By combining Tensorflow models like
    - Wave Net
    - GANS
    - Neural Encoders

# What does YOUR voice say about you?

# The Problem

At some point of time (mostly pre Facebook era :)) in our lives we had wished to see our close to heart RJs, whose voice had kept as awake in late nights or gave energetic early morning start, isn't?

How about solve that curious with latest technologies and internet data available, to build a deep learning model that can "GUESS" a person facial features based on the voice. Yes exactly that's
what you read! Let us use machines to do the sketching for our dearest Rjs who lives in our visualization. (of course these days they have their own FB pages ;))

# The Problem Cont.

Fun part apart, to deal with this challenge we need to be equipped with latest technologies both at Software and Hardware level.

Problem nature might not be directly coupled with any of immediate business values, however if this is taken up, this puts a strong reputation on Imaginea on what kind of complex problems we can deal with.

# Our Solution

There are lot of Deep Learning models that works well on Audio and Video data.
Out which we are passionate to go with Wavenet and GANs

GANs

- A image generator, which creates a **new** image by looking/training after 1000s of images

Wavenet:

- A paper and alibray published by Google DeepMind, which is a text to voice generator that **imitates** human voice

Our naive approach at the moment is to study this two different architecture/models from couple of Github repos and adapt it to our idea.

# Bottlenecks

The major hurdle at the moment what we think would be

- Data collection:
  - Any available video sources out there, eg.,
    - TED talks
    - Youtube videos (where only a single person in the video like tutorials)
    - Screenshots, voice samples, and possible audio to text data.
- Hardware Requirements
  - GPUs

# The Target Market

**- Data Generator:**

- Create new data set from past recordings which imitates the real world data to some degree of certainty
- When above idea gets materialised, it can be wrapped as a fun mobile application to reach the end user, where we can collect more realistic data

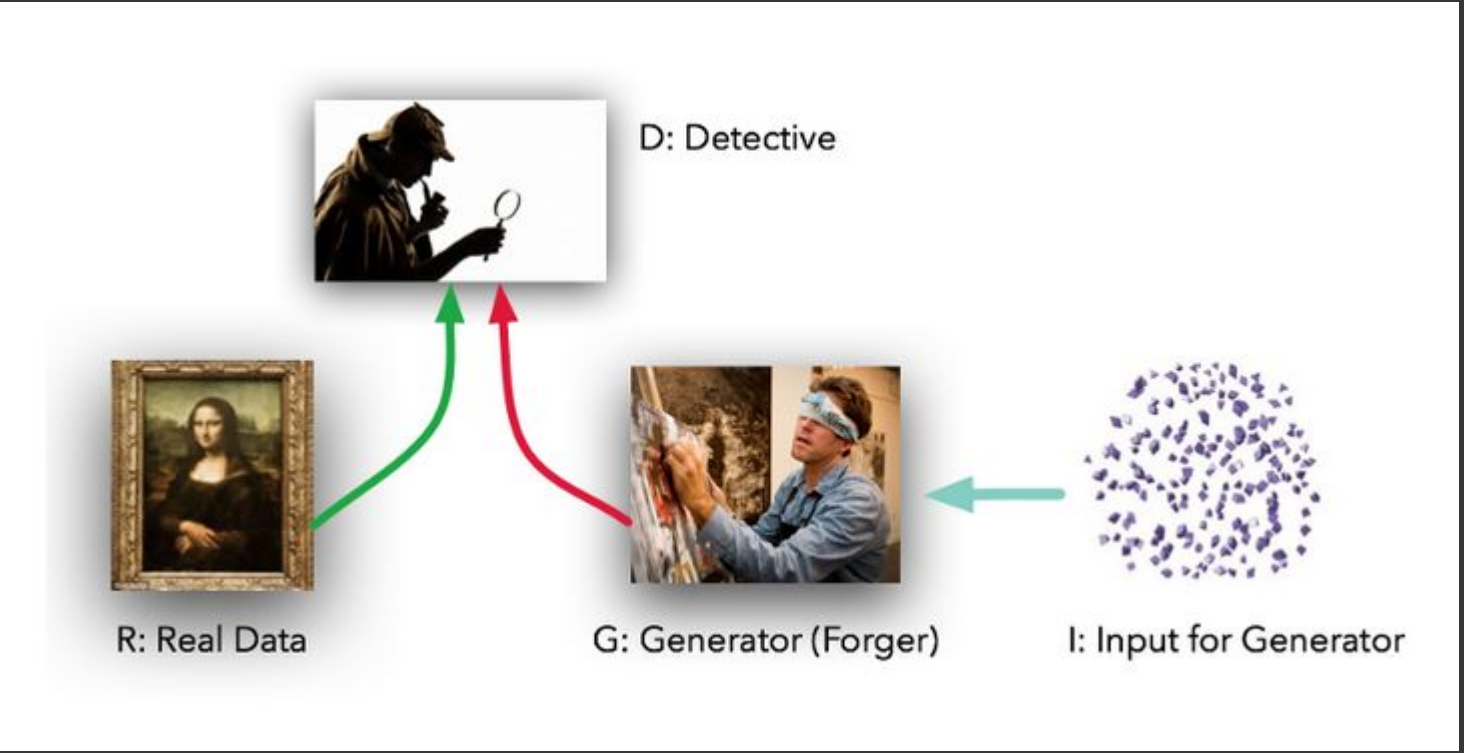
**- Synthesiser:**

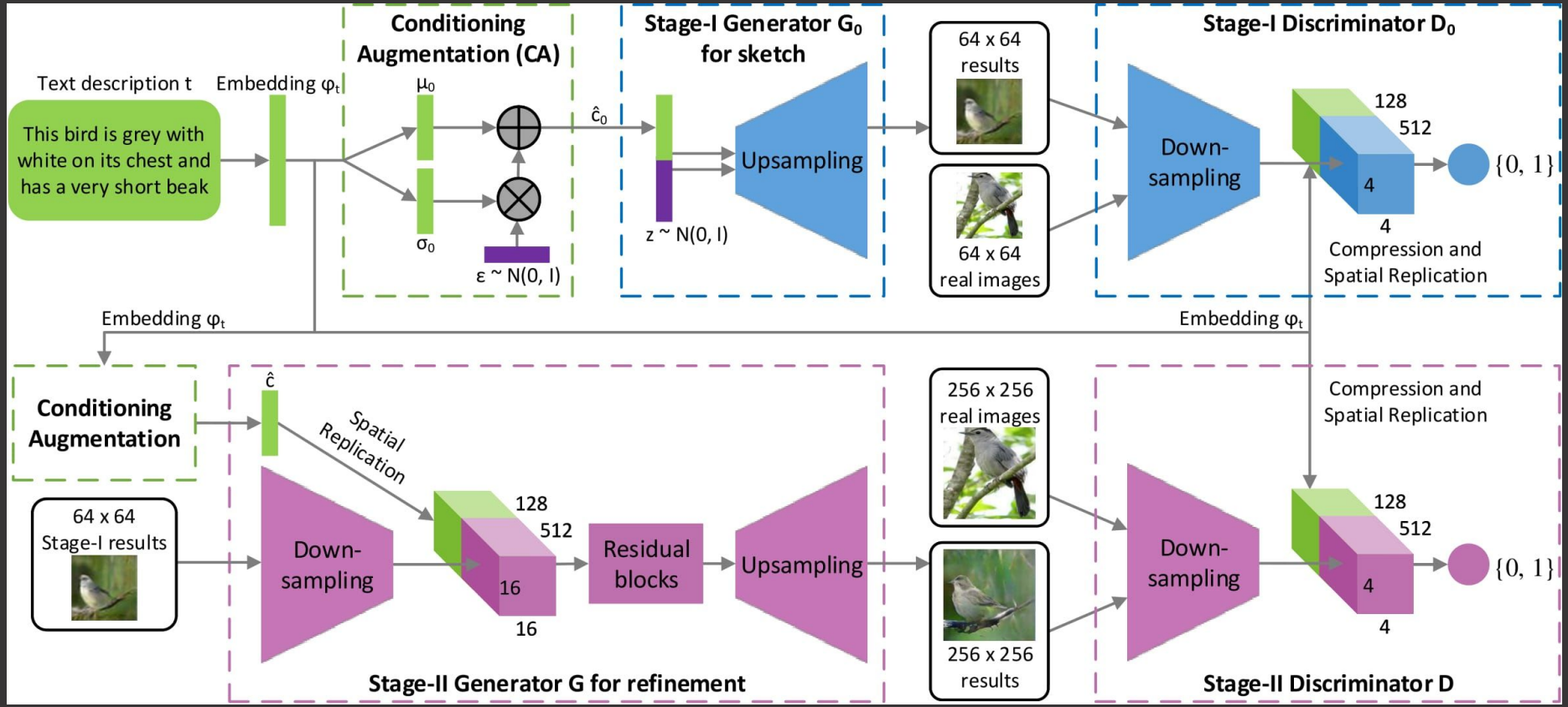
- We might end up with a model which can synthesise a voice for given picture
- In Crime Investigation, we could generate image for a given voice

# Similar Problem Solvers

- GAN
- StackGAN
- DeepMind Wavenet

# GAN


D: Detective
R: Real Data
G: Generator (Forger)
I: Input for Generator


Text description t
This bird is grey with white on its chest and has a very short beak
Embedding $\varphi_t$
Conditioning Augmentation (CA)
$\mu_0$
$\sigma_0$
$\varepsilon \sim N(0, I)$
$\hat{c}_0$
Stage-I Generator $G_0$ for sketch
$z \sim N(0, I)$
Upsampling
64 x 64 results
64 x 64 real images
Stage-I Discriminator $D_0$
Down-sampling
128
512
4
4
{0, 1}
Compression and Spatial Replication
Embedding $\varphi_t$
Embedding $\varphi_t$
Conditioning Augmentation
$\hat{c}$
Spatial Replication
64 x 64 Stage-I results
Down-sampling
128
512
16
16
Residual blocks
Upsampling
Stage-II Generator G for refinement
256 x 256 real images
256 x 256 results
Compression and Spatial Replication
Down-sampling
128
512
4
4
{0, 1}
Stage-II Discriminator D

This bird is white, black, and brown in color, with a brown beak

Stage-I

Stage-II

This flower has long thin yellow petals and a lot of yellow anthers in the center

Stage-I

Stage-II

# Concept Summary

In short, we would like pursue our R&D initiatives in Data Science team with an end goal, at the same time a fun filled project which can be turned down to be a product as the idea materializes and matures.