NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

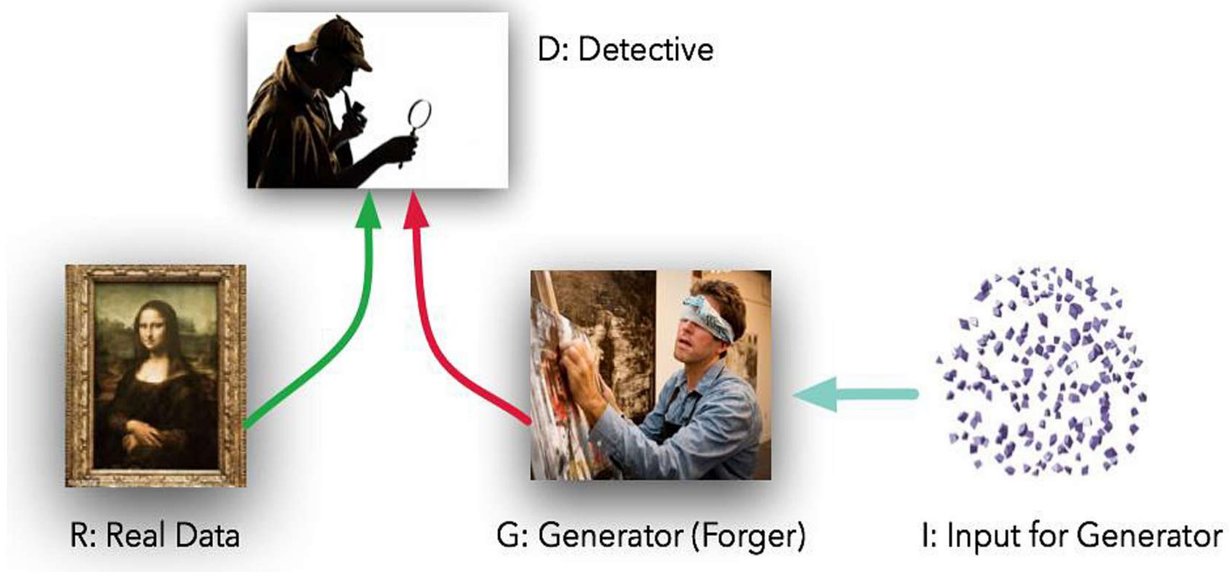
School of Computer Science and Engineering
College of Engineering

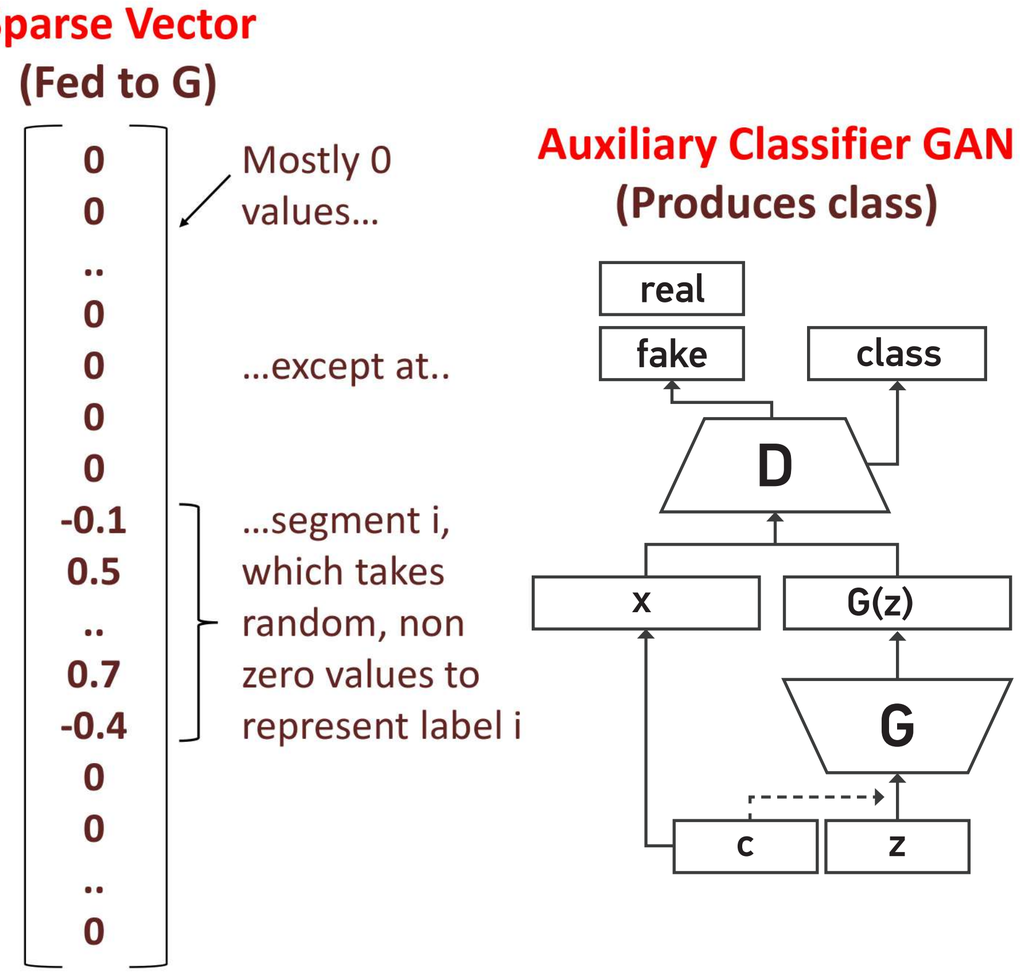
Emotion Speech Synthesis

Generative Models (GAN) for Emotion Speech Synthesis

What are GANs?

Generative Adversarial Networks (GAN) are neural networks that learns distributions of data. They consist of a generator, a 'forger producing counterfeit paintings', and a discriminator, a 'detective trying to distinguish the real from counterfeit paintings'. As each improves, the GAN learns to produce realistic output.





Producing Emotion Speech with GANs

We trained an audio GAN on data from RAVDESS which was an American database of emotions expressed in speech recordings. We taught the GAN to produce each type of emotion using three proposed conditioning techniques: Sparse Vector Conditioning, Auxiliary Classifiers and Dual Auxiliary Classifiers (with text labels).

Results & Findings

We found that using Dual Auxiliary Classifiers could best improve the ability of GAN to produce emotional speech accurately, while using a single classifier produced output which had highest quality and ease of intelligibility as judged by human listeners.

Model	Ease	Quality	Accuracy (Improvement)
Unmodified WaveGAN	2.63	2.13	-
Unmodified CWaveGAN	2.68	2.05	0.08
Sparse Vector CWaveGAN	3.08	2.35	0.15 (87.5%)
AC CWaveGAN	3.26	2.36	0.16 (100%)
Dual AC CWaveGAN	3.04	2.32	0.18 (125%)

Student: Nathanael S Raj | Supervisor: Professor Jagath C Rajapakse