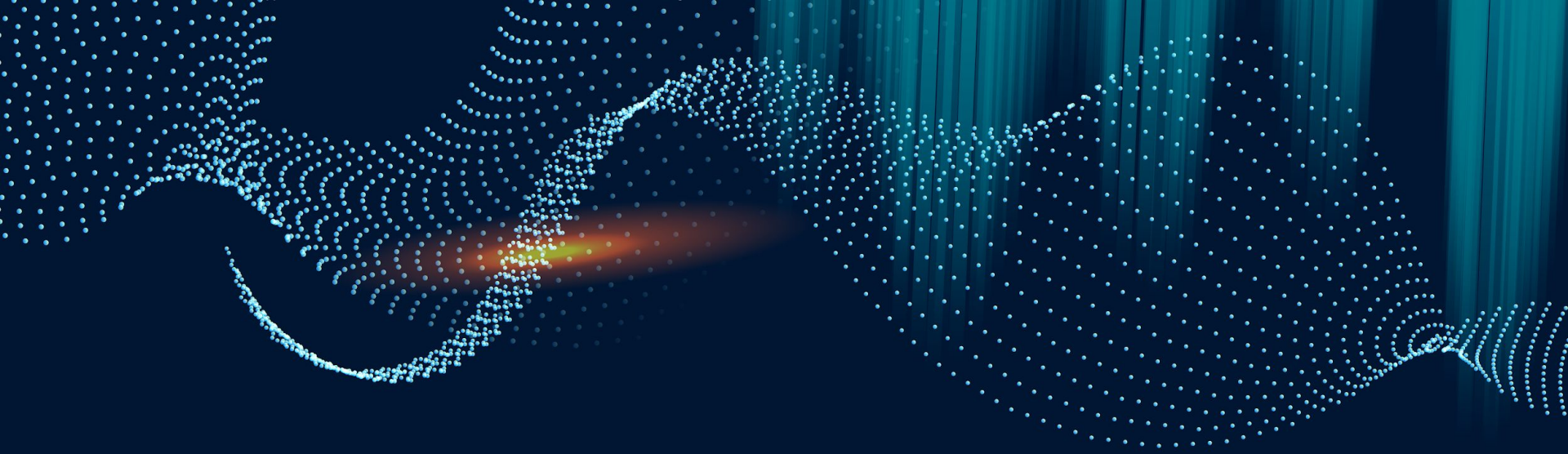# GANs for Speech Emotion Recognition & Translation

Deep Learning
Final Project

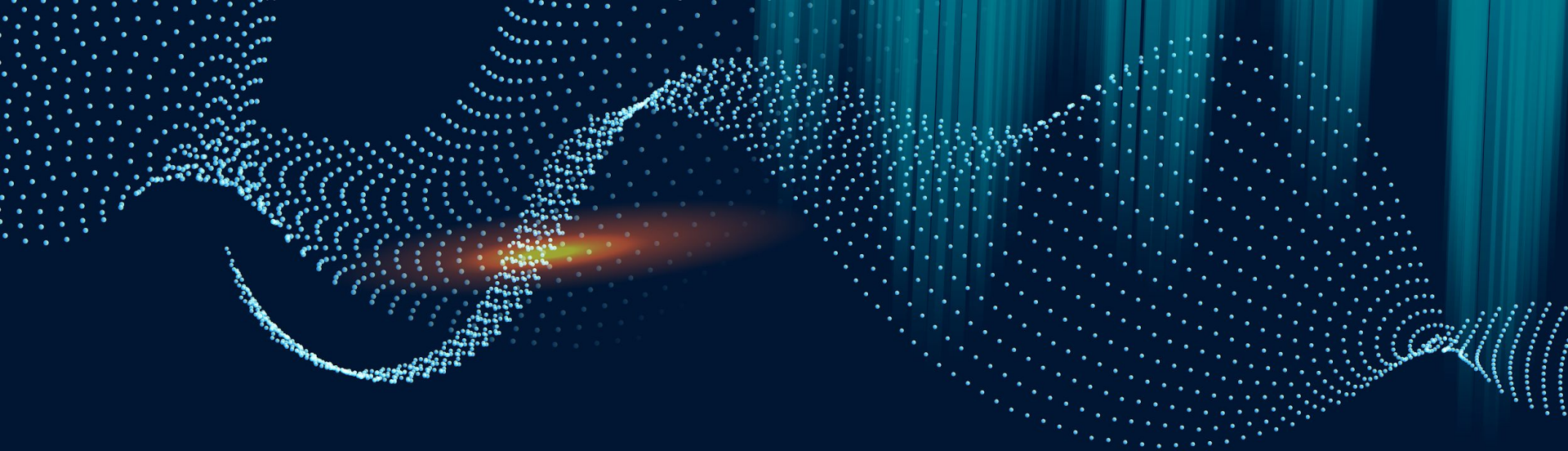By: Nestor Teodoro Chavez
& Charudatta Manwatkar

01 | Goal

Deep Learning Goal

- Paired audio-to-audio translation

- To take an audio recording of a sentence and translate its **emotion** content, while preserving the words.
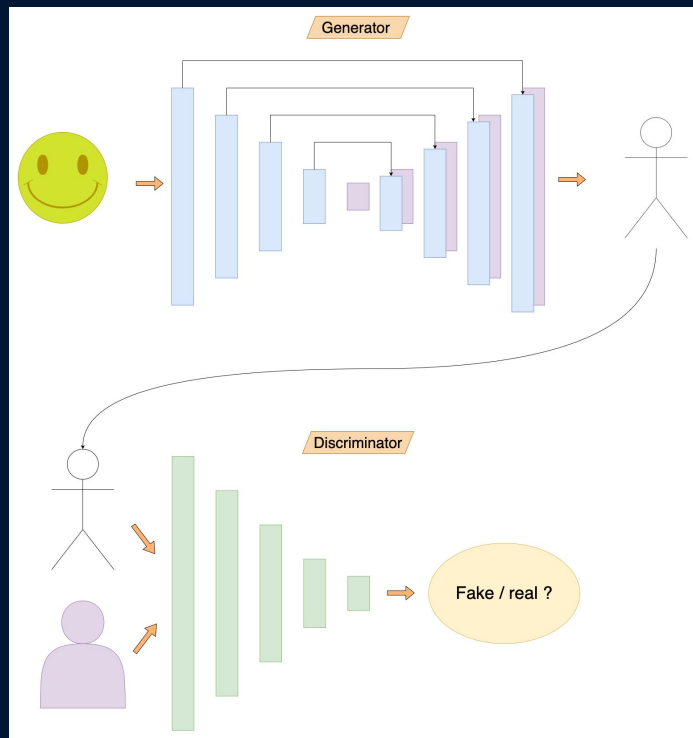
- E.g. convert a sentence from calm to angry

02 | Model Architecture

Model Architecture

03 | **Model Performance**
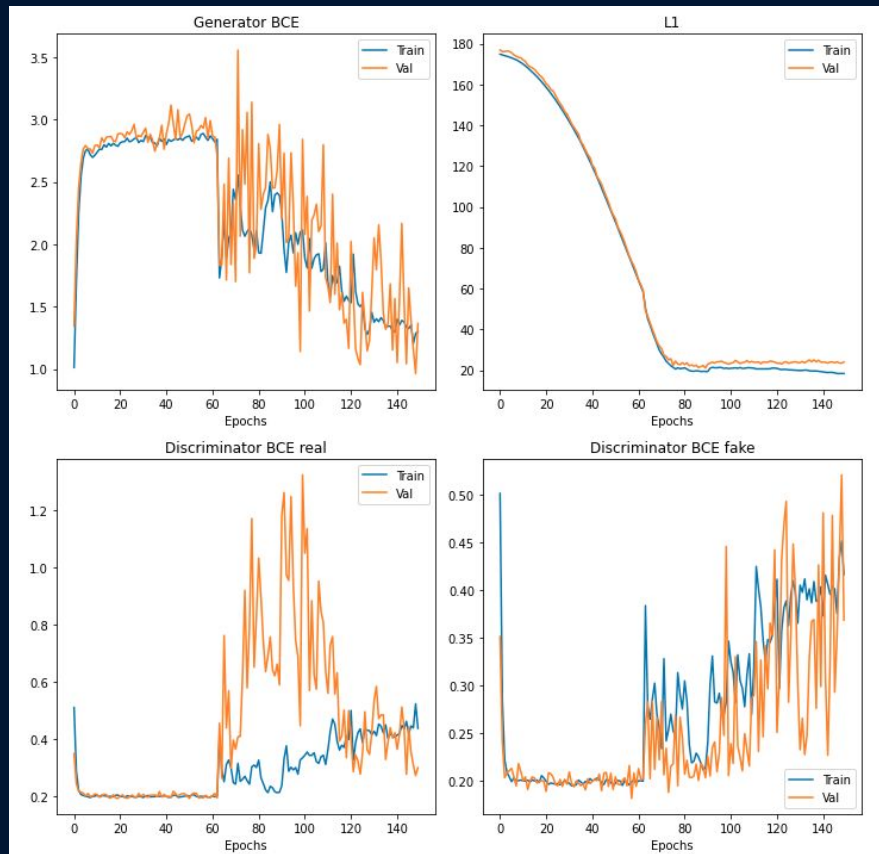
Model Performance

Evaluation

- Model is able to generate 8 different Mel Spectrograms – one for each emotion
- Compared generated Mel Spec to actual Mel Spec as our metric

Training

- We trained for 150 epochs
- With GANs, it is difficult to have proper metrics
- We used BCE for the discriminator and L1 for the Generator

Model Performance

04 | Techniques

GAN-specific techniques

- Label Smoothing
  - We can add small random noise in the labels for discriminator
  - Can be one sided or two-sided

- Asymmetric training
  - We can train Generator more than the discriminator (or vice versa)
  - Criterion can be number of epochs or loss values

04 | Successes

Sucessesses

What went well

- We were able to train the model for 150 epochs in 60 minutes on Google Colab with GPU

- L1 loss decreased steadily

- We learned a lot about audio processing and working with GANs

- We were able to:
  - Convert waveform audio files into images in the form of Mel Spectrogram
  - Make fair predictions for each emotion
  - Lay the foundation for future improvement
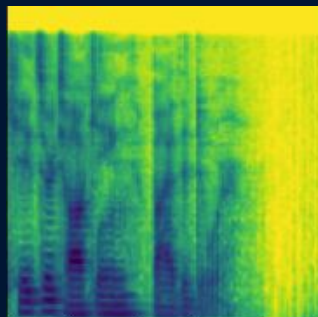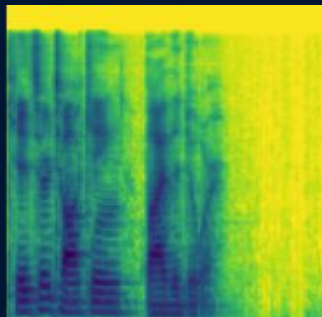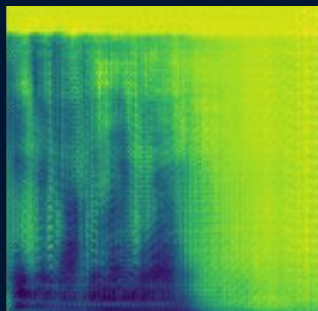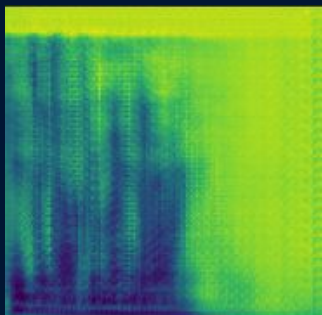
Sucessesses

Emotion	Calm 🔊	Angry 🔊
Original		
Generated		

For the sake of saving eardrums, opted out of what audio predictions are.

05 | Challenges

Challenges

What didn't go well

- The Discriminator learns really quickly and overpowers the Generator

- All emotion's visual representations were very similar

- Conversion from Mel Spectrogram predictions back into audio was abhorrent

Thank You!