Motivation

Audio Style Transfer with Voices
Fabian Boemer, Eric Gong, Youkow Homma
foemer@stanford.edu, ericgong@stanford.edu, yhomma@stanford.edu

NSynth Encoder [ERR+17]
- WaveNet-based autoencoder
- Learns temporal embeddings for audio

Data / Features
- NSynth dataset [ERR+17]
  - 3-4 second single-note pitches sampled at 64 kHz
  - Generated by neural networks in the style of various instruments
  - Used by the Magenta project to train the NSynth model weights [TEN]
- Content dataset
  - 2 NSynth acoustic vocal pitches
  - 1 kHz sine wave sound
  - Recording of a team member’s voice
- Style dataset
  - 3 synthetic flute pitches in the NSynth test set

Loss Function
- Content loss - taken from encoding layer
  \[
  \mathcal{L}_C(x_c, x_g) = \frac{1}{\text{number of notes}} \sum \mathbb{E}[(C(x_c) - C(x_g))^2]
  \]
- Style loss - linear combination of hidden layer embeddings
  - Gram matrix captures correlation between layers
    \[
    \mathcal{L}_s = \frac{1}{n} \sum_n \text{tr}((G_N(x_c) - G_N(x_g))(G_N(x_c) - G_N(x_g))^T)
    \]
  - \( L2 \) loss treats each layer independently
    \[
    \mathcal{L}_s = \frac{1}{n} \sum_n \| (G_N(x_c) - G_N(x_g)) \|^2
    \]
- Total cost - weighted combination of style and content cost
  \[
  \mathcal{L}(x_c, x_g) = \alpha \mathcal{L}_s(x_c, x_g) + \mathcal{L}_C(x_c, x_g)
  \]

Results

Pitch-Pitch Learning

Chord-Pitch Learning

Discussion

- Pitch-Pitch Learning: For \( \alpha = 0 \) with \( L2 \) loss, methodology interpolates between two pitches, since we can move from one pitch to another via gradient backups.
- Chord-Pitch Learning: For \( \alpha = 0 \) with \( L2 \) loss, methodology reconstructs a chord from a pitch after 30 iterations but further iterations result in white noise. Reconstructing a single pitch from a chord is unsuccessful.
- \( L2 \) losses, rather than Gram matrix, used for early style layers can act as a faster, noisier decoder on single tones
- For \( \alpha = 0.08 \) with Gram matrix loss, methodology preserves the content and adds additional frequencies for voice content.

Future Work

- Understand how matching the Maximum Mean Discrepancy via the Gram matrix affects NSynth layers/activations [LCO+17].
- Use histogram losses which minimized parameter tuning and blurring of images in the image Neural Style Transfer method [VWRB17].
- Include losses based on weighted energy contour and frequency energy contour, which stabilized output in Neural Style Transfer for audio spectrograms [V518].

References