Harmonic Combiner: Audio Style Transfer

Kate Pregler
Department of Electrical Engineering, Stanford University

Motivation

While audio generation and interpretation has traditionally relied on the use of recurrent neural networks (RNNs) such as the LSTM model, researchers are experimenting with applying deep CNN models traditionally used for image recognition tasks to audio recognition tasks, using the spectrogram of the audio data as the input image. Given these promising results, I was interested to see if I could replicate the results from Gatys et al "A Neural Algorithm of Artistic Style" with audio instead using the VGG16 model and transfer style from one song to another while retaining similar content (notes).

STFT Audio Encoding Method

Short time fourier transform (STFT) is the first step in creating a spectrogram. STFT estimates the frequency content contained in a small window of the audio by taking the discrete fourier transform (DFT) of the windowed data. Different window functions, overlap, and window lengths may be used, as long as the COLA criteria is satisfied: summing the window functions spaced by the chosen timestep should yield a constant 1.

$$X_m(\omega) = \sum_{n=-\infty}^{\infty} x(n)w(n-mR)e^{-j\omega n}$$

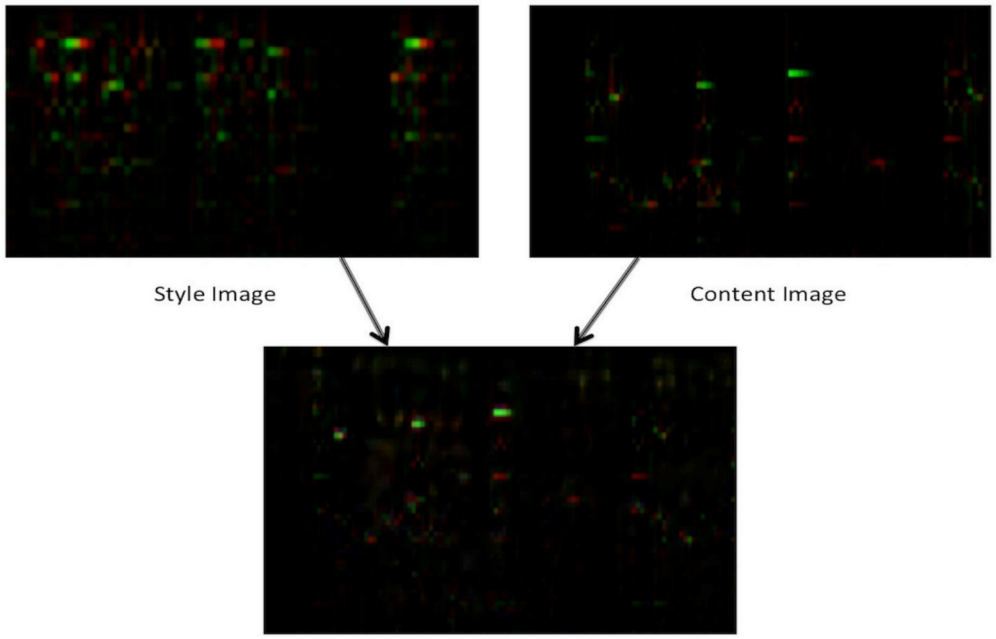
$$= \mathrm{DTFT}_\omega(x \cdot \textsc{Shift}_{mR}(w)),$$

$$\sum_{m=-\infty}^{\infty} w(n-mR) = 1, \ \forall n \in \mathbb{Z}$$

STFT and COLA requirements with window function *w*, column *m*, and timestep *R*

Choosing window length is a trade-off between accuracy in the frequency and time domains; while a longer window allows for better discretization of the audio frequency spectrum, it causes worse discretization in the time domain. I found a window length of 0.1s to be a good point for balancing the two.

Encoded Source Images and Sample Generated Image



Sample Results

Using alpha = 10 and beta = 40 on 3 minutes of data from each song, with 420 iterations.

Iteration Number	Content Cost	Style Cost	Total Cost
0	11353	6.054×10^8	2.422×10^{10}
200	1829	1.19×10^5	4.794×10^6
420	1261	5.84×10^4	2.35×10^6

Acknowledgements and References

Gatys, Ecker, Bethge. "A Neural Algorithm of Artistic Style," Sept. 2015.
STFT Implementation from CCRMA: https://ccrma.stanford.edu/~jos/sasp/Mathematical_Definition_STFT.html
Many thanks to the teaching staff of CS230