NEURAL STYLE TRANSFER ON AUDIO SIGNALS

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY VADODARA

Course co-ordinator: Dr Jignesh Bhatt

Contents

Problem Statement

Neural Style transfer is very popular for generating novel arts. In this project we try to extend the application of Neural Style transfer to audio signals as well. We intend to create new music genre by applying style of an existing genre on another.

Motivation

"Style Transfer" of images are very popular and an active research topic. Its everyone's favourite tool to make photos look like surreal paintings. Convolutional Neural Networks (CNNs) are adaptable to a great variety of tasks. We want to extend and modify the algorithm for audio signals and use the power of CNNs to generate new audio from a style audio. This can be a tune or a beat, or even simply someone speaking the lyrics of a song.

Dataset

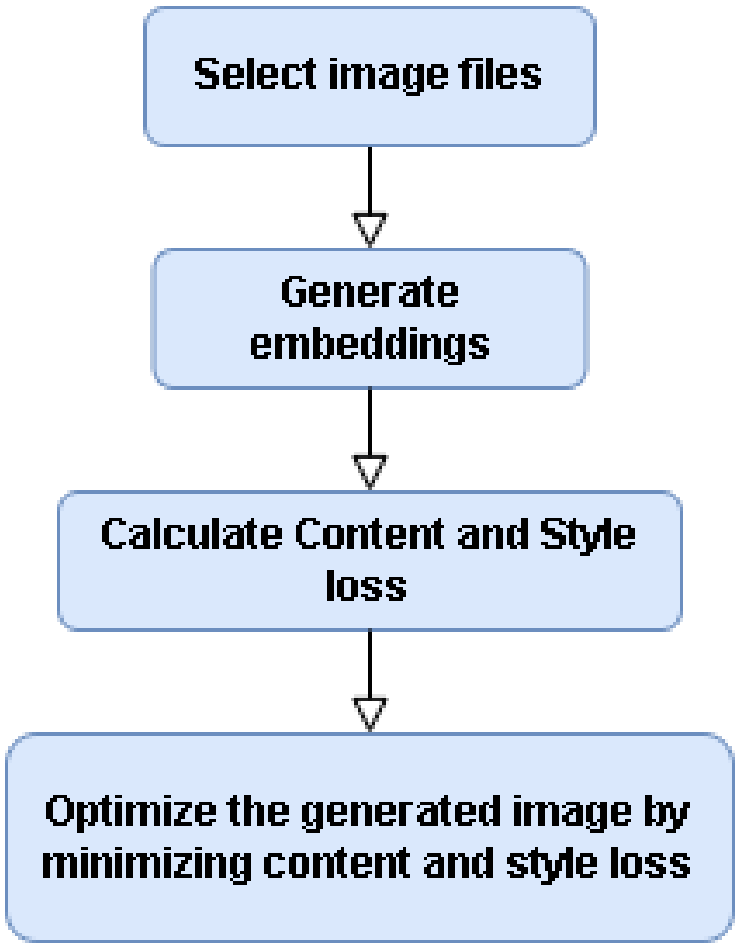
- GTZAN Dataset - Music Genre Classification
  - 30 seconds .wav files of Jazz and Classical genres
- Other audio clips taken from different sites to try out combinations
  - Music from games, ex Mario, etc
  - Popular movies theme songs, ex Interstellar, Harry Potter, etc

Neural Style Transfer on images

- Content Image
  - The image on which style transfer is applied.
- Style Image
  - The style we want to transfer to content image.
- Generated Image
  - The final blend of content and style image.
- Loss Function
  - The Loss Function is made up of Content Loss Function and Style Loss function.

Neural Style Transfer on images

Neural Style Transfer on images



Content Loss Function

- Using hidden layer 'l' to compute content cost.
- Use pre-trained ConvNet. (E.g., VGG network)
- Let $a^{[l](C)}$ and $a^{[l](G)}$ be the activation of layer l on the images
- If $a^{[l](C)}$ and $a^{[l](G)}$ are similar, both images have similar content

$$\mathcal{L}_{content}(\vec{p}, \vec{x}, l) = \frac{1}{2}\sum_{i,j}(F_{ij}^{l} - P_{ij}^{l})^2$$

$\vec{p}$ The original image
$\vec{x}$ The generated image
l Layer
F_{ij}^{l} Activation of the i^{th} filter at position j in the feature representation of $\vec{x}$ in l
P_{ij}^{l} Activation of the i^{th} filter at position j in the feature representation of $\vec{p}$ in l

Style Loss Function

- Style is defined as correlation between activations across channels
- Let $a^{[l]}_{i,j,k}$ = activation at (i, j, k) and layer 'l'. $G^{[l]}$ is $n^{[l]}_c \times n^{[l]}_c$
- Calculated as the distance between the gram matrices of the generated image and the style reference image.
- Difficult to compare intermediate features of both images hence Gram matrix used.
- Style loss defined as -

$$L_{style} = \sum_l w^l L^l_{style} \text{ where,}$$

$$L^l_{style} = \frac{1}{M^l} \sum_{ij} (G^l_{ij}(s) - G^l_{ij}g)^2 \text{ where,}$$

$$G^l_{ij}(I) = \sum_k A^l_{ik}(I) A^l_{jk}(I).$$

Total Loss Function

$$\mathcal{L}_{total}(\vec{p}, \vec{a}, \vec{x}) = \alpha \mathcal{L}_{content}(\vec{p}, \vec{x}) + \beta \mathcal{L}_{style}(\vec{a}, \vec{x})$$

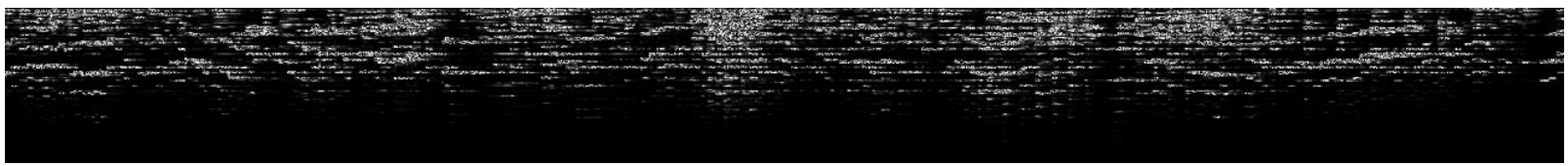
$\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the weighting factors for content and style reconstruction

Mel Spectogram

- What is Spectogram?
- Why Mel Spectogram?

$$m = 2595log(1 + \frac{f}{500})$$

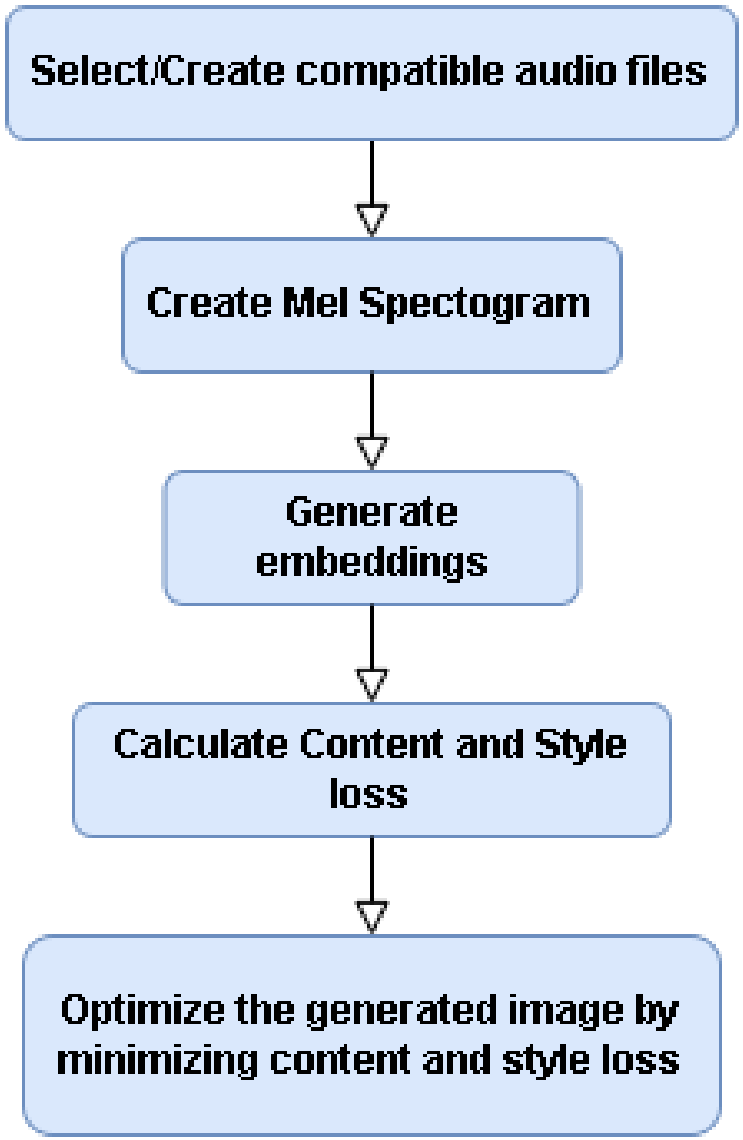
- How Mel Spectogram works?



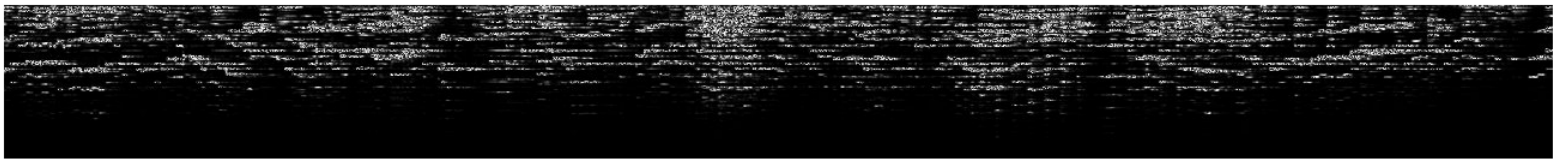
How Mel Spectogram works?

- Apply STFTs on audio signals.
- Convert Amplitude to Decibels(DBs).
- Convert frequencies to Mel Scale.

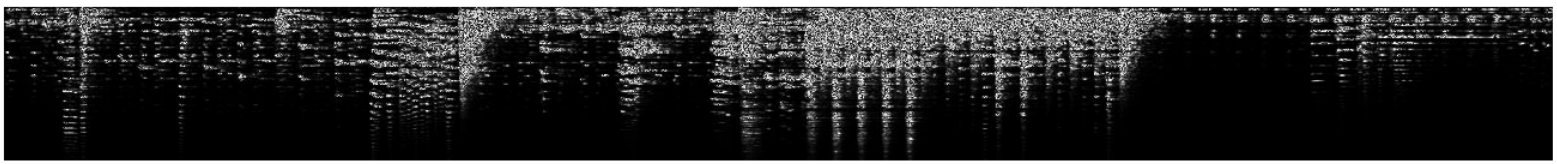
Style Transfer



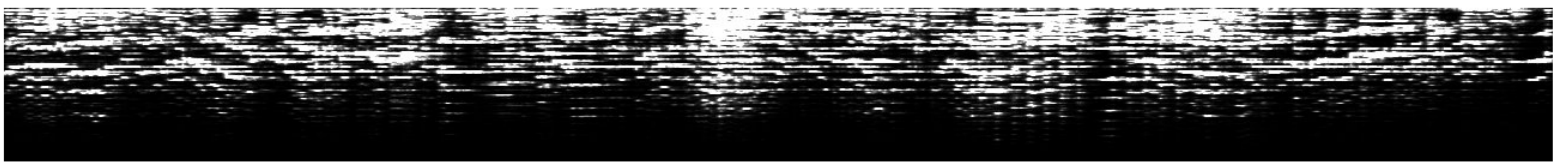
Results



Content



Style



Generated Audio

References

- Neural Style Transfer on Audio Signals
- Neural Style Transfer for Audio Spectograms
- TensorFlow

Any Questions?