DUONG HOANG
LUC TAN THO

AUDIO STYLE TRANSFER

TABLE OF CONTENTS

FROM IMAGE TO AUDIO STYLE

Style image

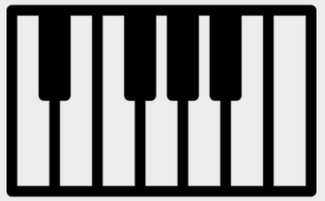
Content image

CNN based network

Generated image

Drum audio

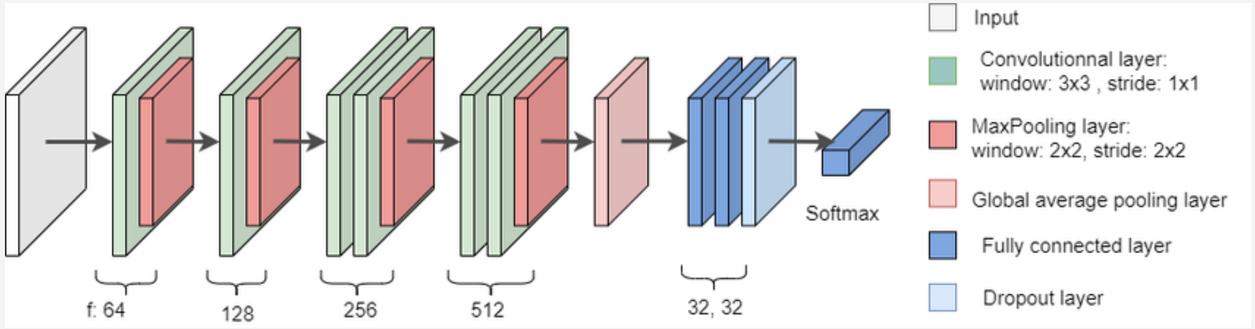
Piano audio



Drumbeat on piano melody

VGGISH FOR AUDIO TRANSFER

Vggish network



Same workflow as the original problem on image:

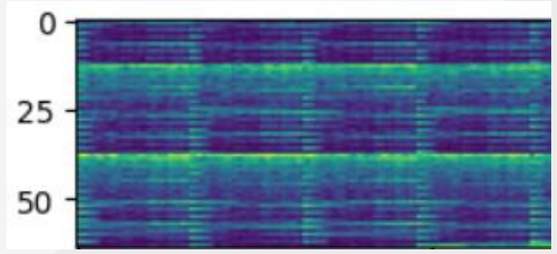
- Use a pre-trained CNN network for classification
- Choose layers for computing content and style loss
- Do backprop to generate stylized audio

Vggish:

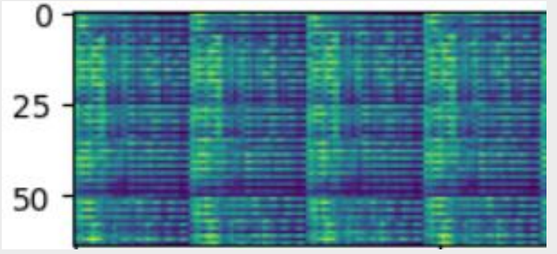
- Similar to Vgg
- Pre-trained model for audio classification
- Input: 2D mel spectrogram

Results

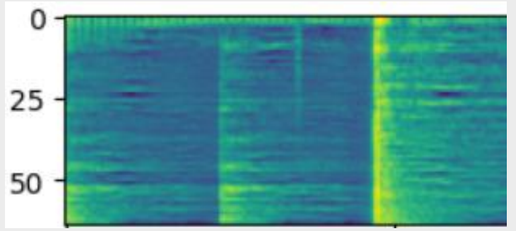
Content



Style



Generated audio



☹️ Not good, lost all content!

CNN14 FOR AUDIO TRANSFER

CNN14

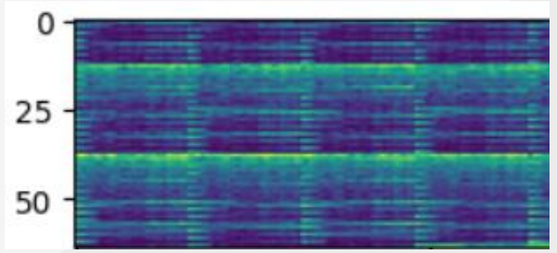
$\begin{pmatrix} 3 \times 3 \text{ @ } 64 \\ \text{BN, ReLU} \end{pmatrix} \times 2$
$\begin{pmatrix} 3 \times 3 \text{ @ } 128 \\ \text{BN, ReLU} \end{pmatrix} \times 2$
$\begin{pmatrix} 3 \times 3 \text{ @ } 256 \\ \text{BN, ReLU} \end{pmatrix} \times 2$
$\begin{pmatrix} 3 \times 3 \text{ @ } 512 \\ \text{BN, ReLU} \end{pmatrix} \times 2$
Pooling 2×2
$\begin{pmatrix} 3 \times 3 \text{ @ } 1024 \\ \text{BN, ReLU} \end{pmatrix} \times 2$
Pooling 2×2
$\begin{pmatrix} 3 \times 3 \text{ @ } 2048 \\ \text{BN, ReLU} \end{pmatrix} \times 2$
Global pooling
FC 2048, ReLU
FC 527, Sigmoid

CNN14:

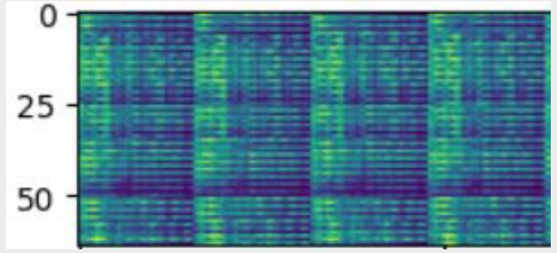
- Pre-trained model for audio classification
- Input: 2D log-mel spectrogram

Results

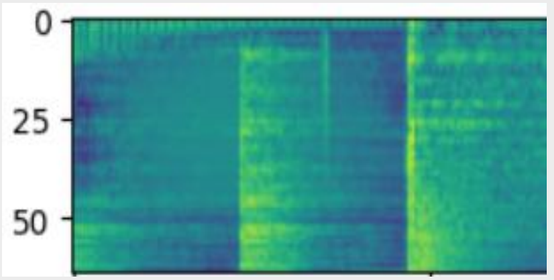
Content



Style



Generated audio



Still not good!

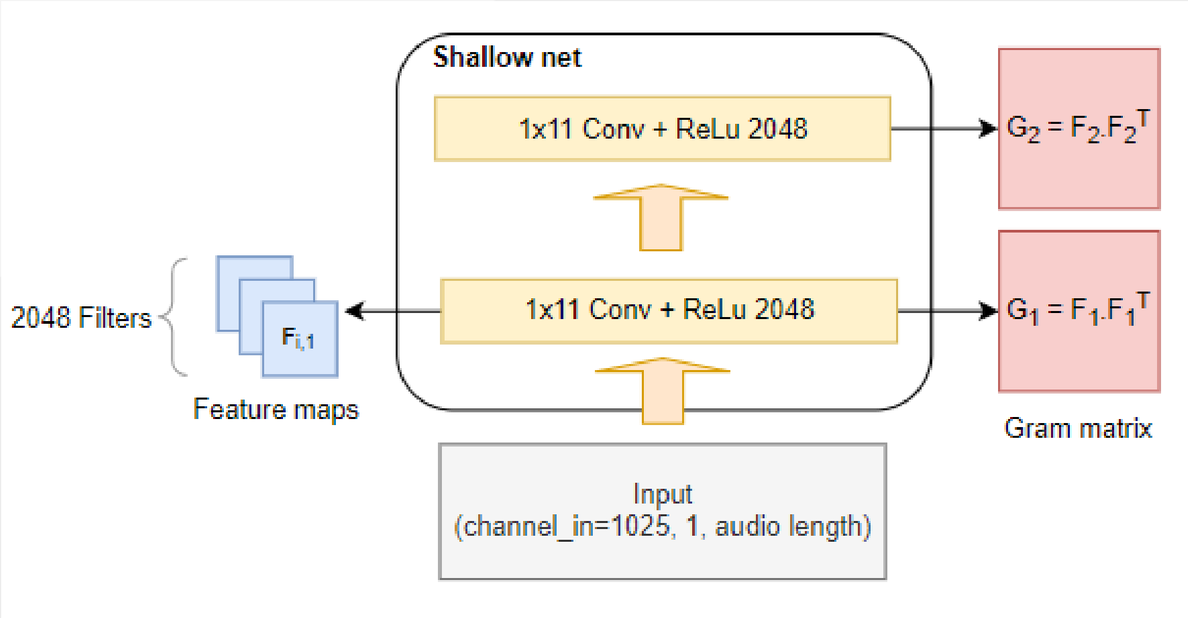
SHALLOW NET

Note:

- Visual perception and auditive perception may be different
- Pooling layer may not work for this audio problem
- Each frequency should not be interlinked with others

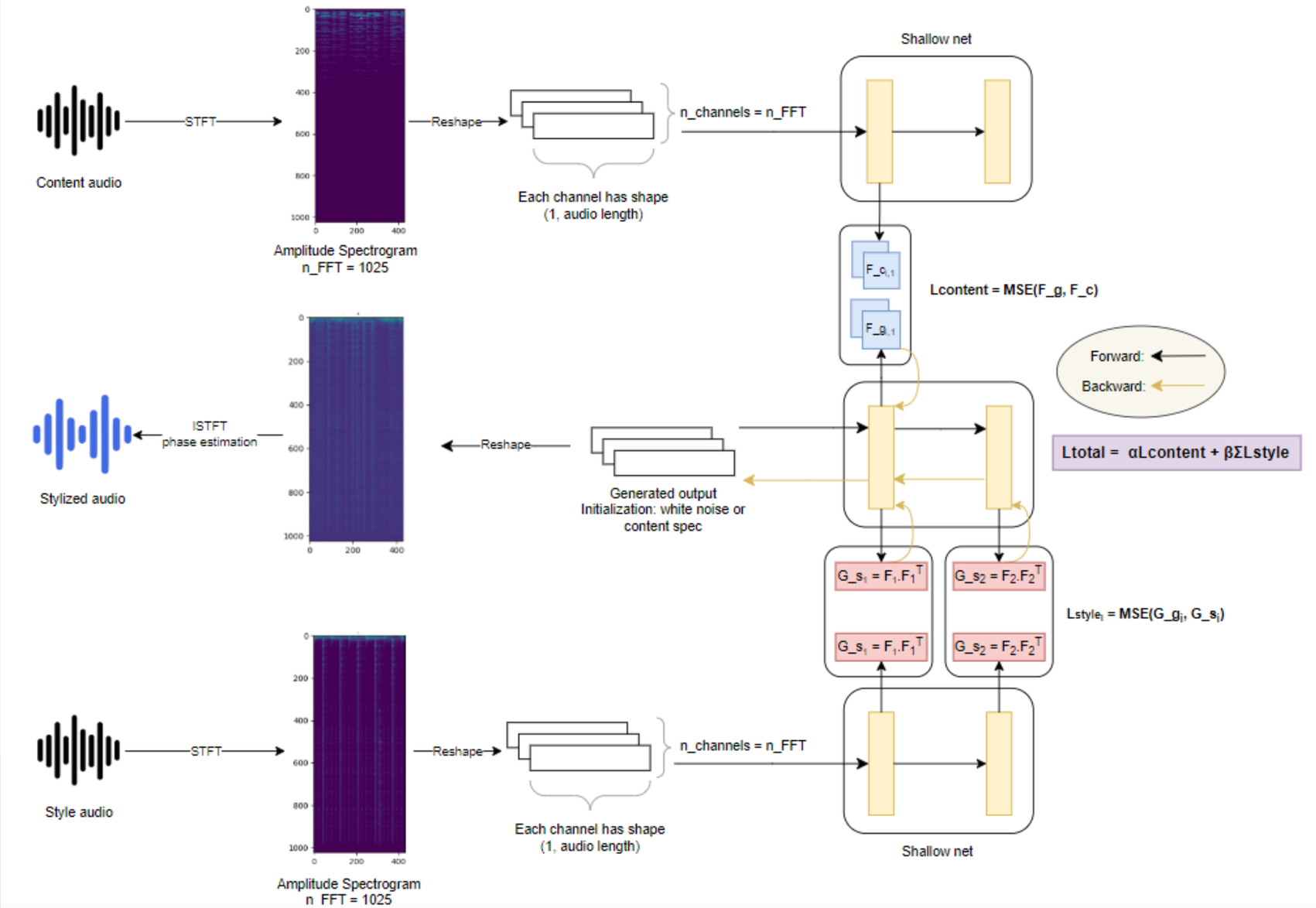
Shallow net

Based on Ulyanov's work, we define a shallow model:



1D conv instead of 2D conv
2 conv layers, 2048 filters for each layer
Do not need to be trained, all weights are random
Input: reshaped amplitude spectrogram

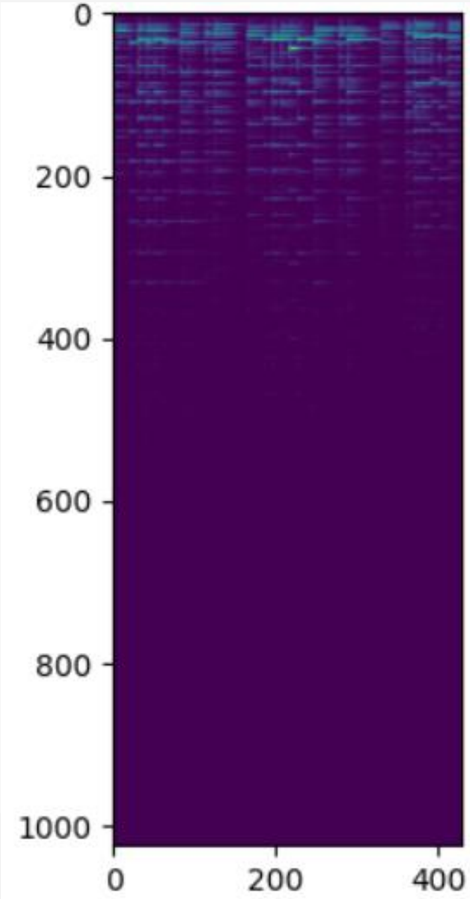
SHALLOW NET


Content audio
STFT
Amplitude Spectrogram
n_FFT = 1025
Reshape
Each channel has shape
(1, audio length)
n_channels = n_FFT
Shallow net
F_c,1
F_g,1
Lcontent = MSE(F_g, F_c)
Forward:
Backward:
Ltotal = αLcontent + βΣLstyle
G_s1 = F1.F1T
G_s2 = F2.F2T
Lstyle_l = MSE(G_g_i, G_s_i)
Generated output
Initialization: white noise or
content spec
Reshape
ISTFT
phase estimation
Stylized audio
Style audio
STFT
Amplitude Spectrogram
n_FFT = 1025
Reshape
Each channel has shape
(1, audio length)
n_channels = n_FFT
Shallow net

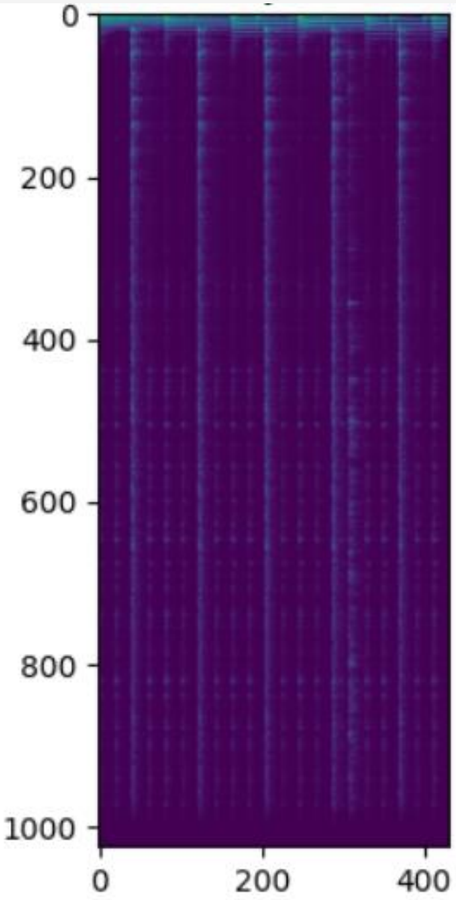
SHALLOW NET

Result

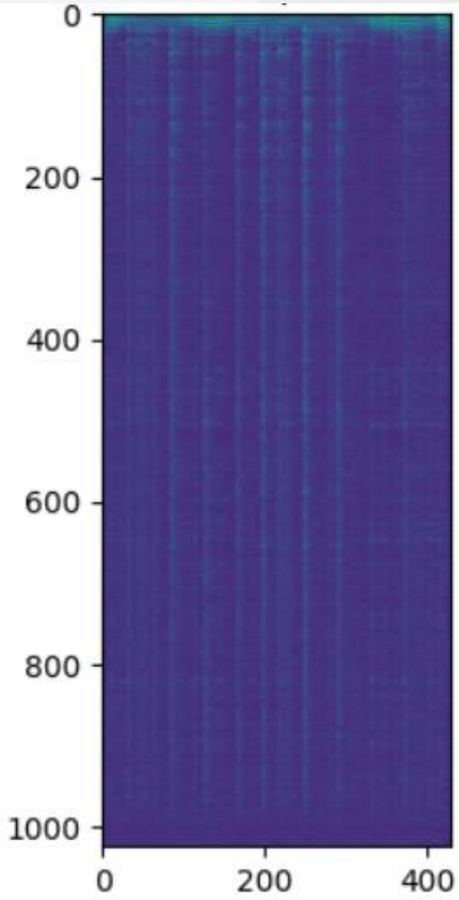
Content



Style



Generated output



SHALLOW NET

More results:

	Content	Style	Generated output
Piano to drum (same song)			
Drum to piano (same song)			
Drum to piano (different song)			
Piano to String (same song)			

Click here to access our github repository

REFERENCES

- Image Style Transfer Using Convolutional Neural Networks
- CNN ARCHITECTURES FOR LARGE-SCALE AUDIO CLASSIFICATION
- https://dmitryulyanov.github.io/
- AUDIO STYLE TRANSFER, Eric Grinstein, Ngoc Q.K. Duong, Alexey Ozerov and Patrick Pérez
- PANNs:Large-Scale Pretrained Audio Neural Networks for AudioPatternRecognition

THANK YOU/ MERCI BEAUCOUP