Audio Style Transfer

Yang Lu, Zhixian Huang, Bingjing Zhu

Overview

- Introduction
- Related work
  - Texture Network
  - StyleNet
  - WaveNet
- Future research direction

Introduction

- **What is audio style?**
  - Timbre/Instrumentation
  - Dynamics/Velocity(For different definition, music expressiveness)
- **Application**:
  - Assist in composing
  - Music education
  - Data augmentation

Image Style Transfer

- Sucess in image style transfer.
- How about audio style transfer?

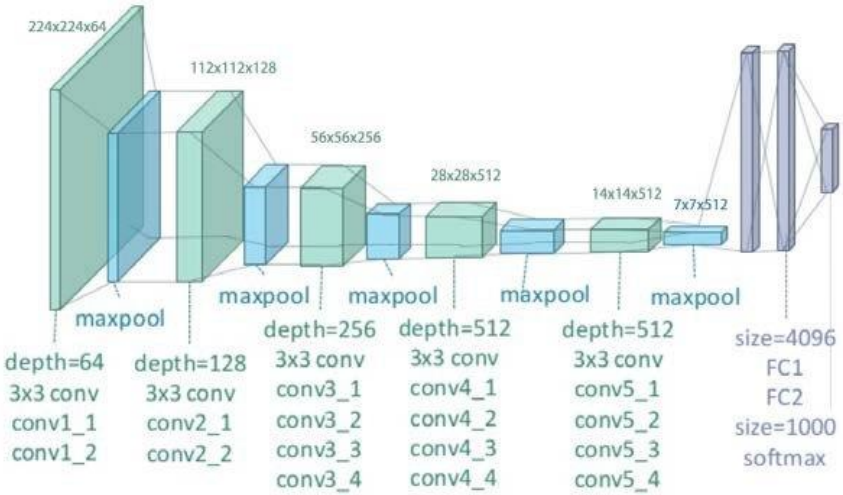


Texture Network: Iterative Optimazation

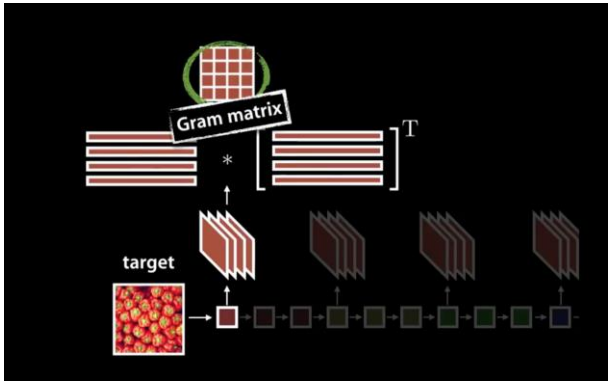
- [Texture Synthesis Using Convolutional Neural Networks, Gatys et al., NIPS'15]
- **Goal**: turn noise into similar texture image
- **Idea**:
  - Define a style loss function: L_style(x;y)
  - Use pre-trained CNN to extract features
  - Gradually optimize an initialization image by back propagation

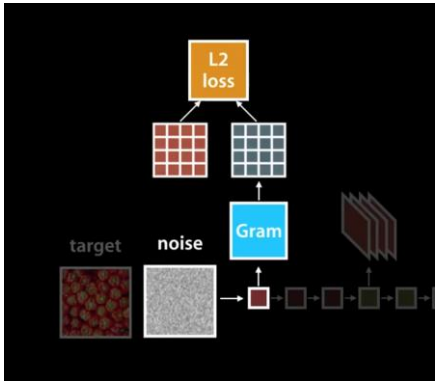


CNN for feature extraction:VGG19



Texture Network: Style Loss Details





Gram matrix:

$$G_{ij}^{l} = \sum_{k} F_{ik}^{l} F_{jk}^{l}$$

Texture Network: Style Transfer

- **Content loss:** based on the output feature maps of certain convolutional layers without computing further statistics
- Content loss compares feature activations at corresponding spatial locations, and therefore preserves spatial information.



Texture Network: Audio Style Transfer

- The raw audio is converted to a spectrogram via STFT
- Network: pre-trained network vs random weights
- Reconstruct a signal from its spectrogram
- Result: https://dmitryulyanov.github.io/audio-texture-synthesis-and-style-transfer/

Texture Network: Pros & Cons

- Pros:
  - Fast and simple
- Cons:
  - Lack of performance: sounds often as a mixture of the style and content sounds rather than as a stylized content
  - Fail in describing audio style and content

WaveNet:Generative Model

- [WaveNet: A Generative Model for Raw Audio]
- Goal: Generate raw audio waveforms from raw audio waveforms
- Idea:
  - Autogressive model.Predict distribution of audio sample based on previous samples.
  - Softmax distribution.

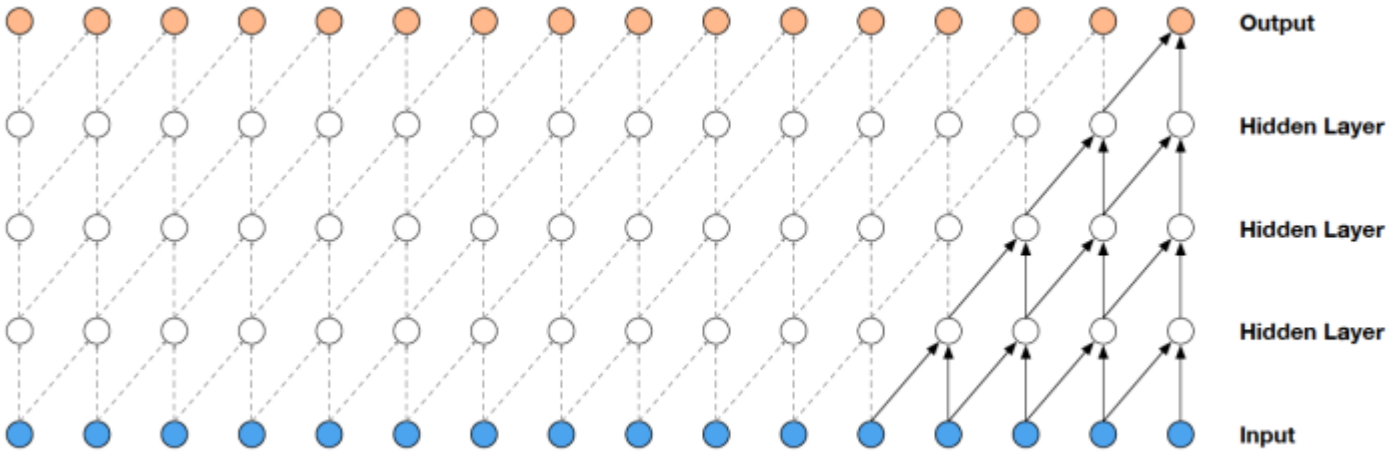
$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

Autoregressive model in images

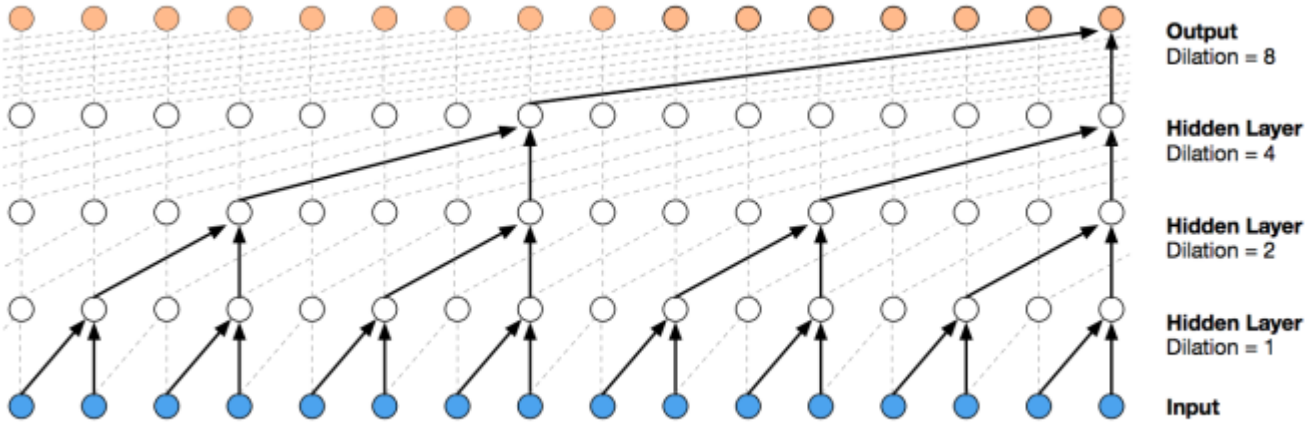


Figure 1. Image completions sampled from a PixelRNN.

WaveNet:Causal Convolutions



WaveNet:Dilated Causal Convolutions

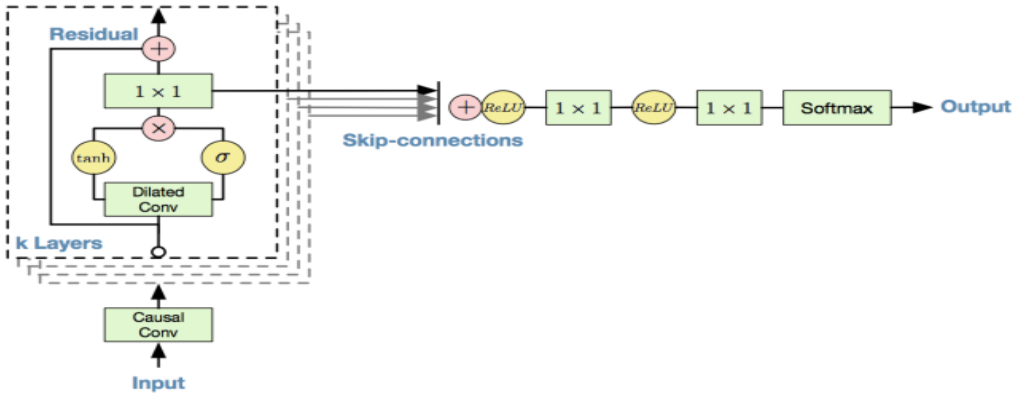


WaveNet:Softmax distribution

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad \text{for } j = 1, \ldots, K.$$

- The raw audio is stored as a sequence of 16-bit integer values(one per timesteps).
- Do a non-linear mapping(μ-law companding transformation) to 8-bit integer values

WaveNet:Entire Architecture



$$z = tanh(W_{f,k} * x) \odot \sigma(W_{g,k} * x)$$

WaveNet:Conditional WaveNet

- h can be :
- Speaker ID
- Music genre, instrument
- ...

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h})$$

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k}^T \mathbf{h}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k}^T \mathbf{h}\right)$$

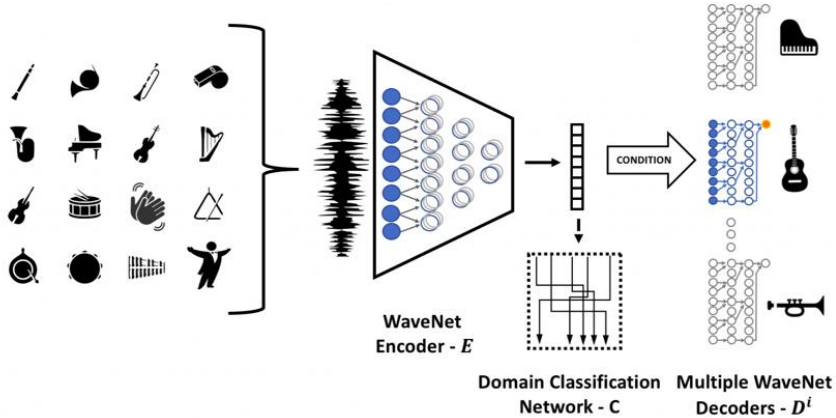
Demo

- https://deepmind.com/blog/wavenet-generative-model-raw-audio/

WaveNet:Music Translation Network

- [A Universal Music Translation Network, Noam et al.]
- Goal: Translate music across styles and instruments
- Idea:
  - Shared encoder, different decoder

WaveNet:Entire Architecture



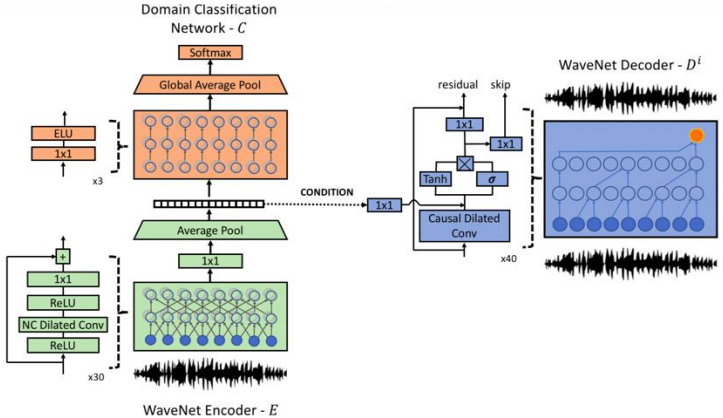
WaveNet:Classfication Network

- Loss for trianing
  - domain classifier: Loss of domain classification
  - autoencoder:Loss of generative output and target
- Transformation:

$$D^j(E(s)).$$

$$\Omega = \sum_j \sum_{s^j} \mathop{\mathbb{E}}_{r} \mathcal{L}(C(E(O(s^j, r))), j)$$

$$-\lambda\Omega + \sum_j \sum_{s^j} \mathop{\mathbb{E}}_{r} \mathcal{L}(D^j(E(O(s^j, r))), s^j)$$



Demo

- https://research.fb.com/publications/a-universal-music-translation-network/

WaveNet: Pros & Cons

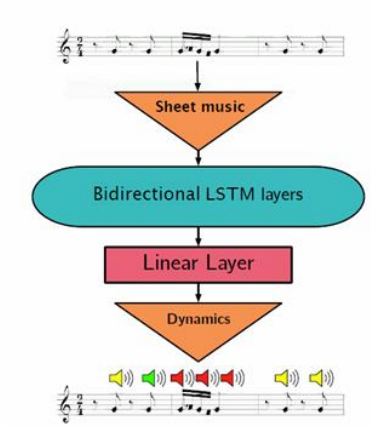
- Pros:
  - Good performance
  - Capture complex features
- Cons:
  - Hard to change dynamic and velocity
  - High computation cost
  - Still cannot work very good in multiple instruments music style

StyleNet: Concept

- Consider Dynamic as the representation of style

i.e. Jazz and Classical Music have their own identical velocities.

- Use MIDI note as the input and training data.

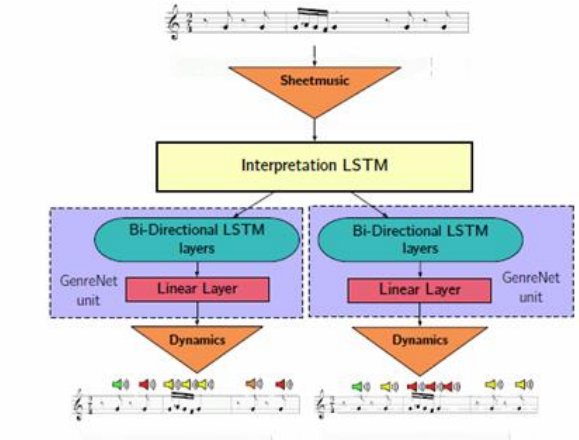


GenreNet

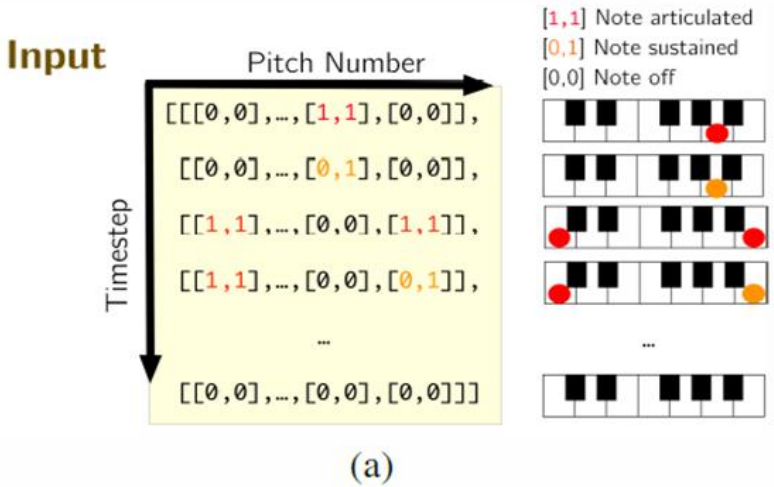
- Able to learn the dynamic of one specific style.

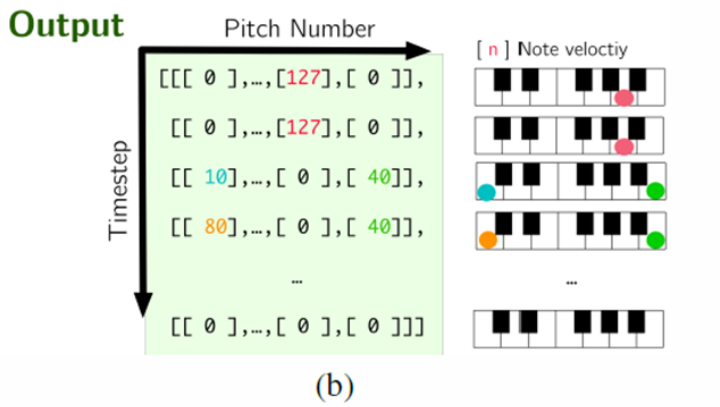
StyleNet

- Combining multiple GenreNets.



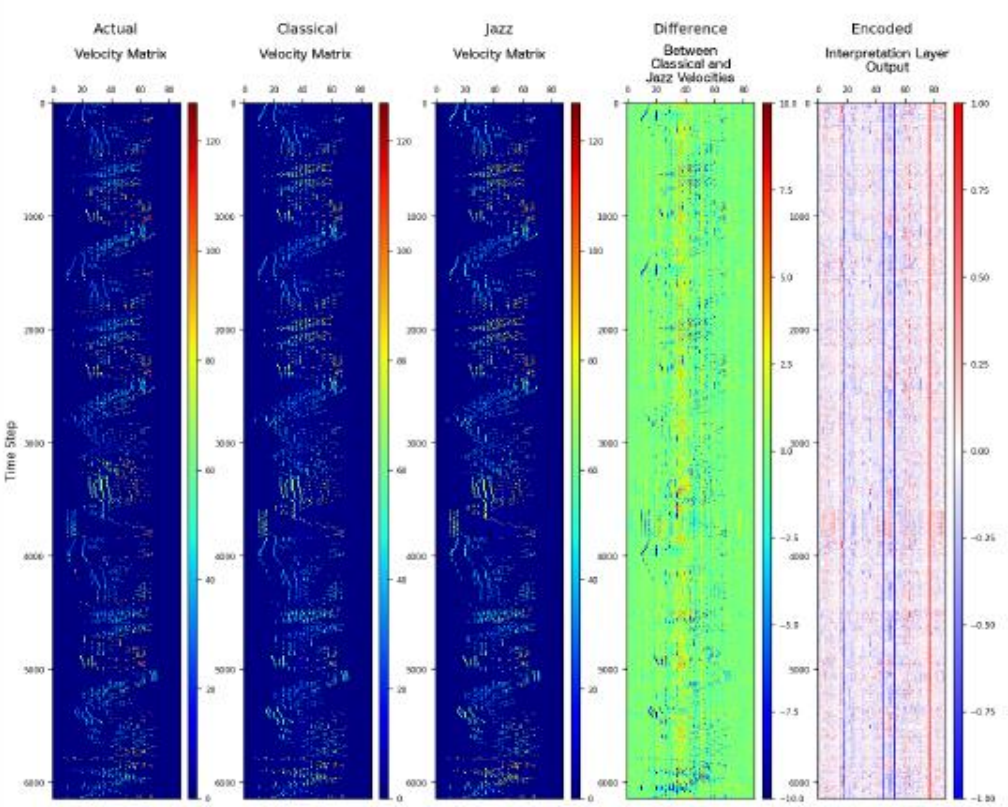
StyleNet: Input and Output Matrix





StyleNet

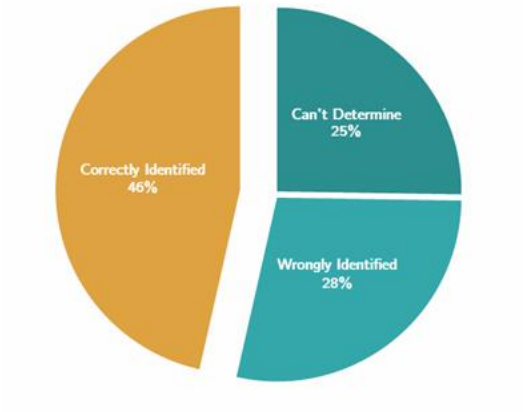
Result:
http://imanmalik.com/cs/2017/05/neural-style.html



StyleNet: Evaluation

“Identify the Human” Test : Turing Test

- Two parts with 9 questions, each question contains 2 audio clips of same performance.
  i.e. One played by actual human and one generated by model.
- Participants were asked to identify the one played by human.



StyleNet:Evaluation

“Identify the style” Test

- Two parts with 9 questions, each question contains 2 audio clips generated by different styleNnet i.e. Classical and Jazz.
- An average of 47.5% respondents selected the correct style.
- Not sufficient to separate the characteristic between two styles.

StyleNet: Pros & Cons

- Pros:
  - MIDI note provides smaller dataset for training.

- Cons:
  - Input is limited to the form of MIDI note, so as the style .
  - Model cannot generate noticeably stylized performances.
  - Defination of style is too strict,not flexible

Future Research Direction

- Improve the performance
- Reduce computational cost
- Change dynamic and velocity

Reference

- Gatys, Leon, Alexander S. Ecker, and Matthias Bethge. "Texture synthesis using convolutional neural networks." *Advances in Neural Information Processing Systems*. 2015.
- Ulyanov, Dmitry, et al. "Texture Networks: Feed-forward Synthesis of Textures and Stylized Images." *ICML*. 2016.
- Van Den Oord, Aäron, et al. "WaveNet: A generative model for raw audio." *SSW*. 2016.
- Mor, Noam, et al. "A Universal Music Translation Network." *arXiv preprint arXiv:1805.07848* (2018).
- Grinstein, Eric, et al. "Audio style transfer." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.
- Barry, Shaun, and Youngmoo Kim. ""“Style” Transfer for Musical Audio Using Multiple Time-Frequency Representations." (2018).
- Mital, Parag K. "Time domain neural audio style transfer." *arXiv preprint arXiv:1711.11160* (2017).
- Malik, Iman, and Carl Henrik Ek. "Neural translation of musical style." *arXiv preprint arXiv:1708.03535* (2017).

Q&A