Neural Audio Synthesis of Musical Notes

with WaveNet Autoencoders

2019-02-08

곽민구

00 | Paper

- Google Brain & DeepMind
- Google magenta 팀
- International Conference on Machine Learning 2017
- 인용: 87회

Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders

Jesse Engel [*1] Cinjon Resnick [*1] Adam Roberts [1] Sander Dieleman [2] Mohammad Norouzi [1] Douglas Eck [1]
Karen Simonyan [2]

Abstract

Generative models in vision have seen rapid progress due to algorithmic improvements and the availability of high-quality image datasets. In this paper, we offer contributions in both these areas to enable similar progress in audio modeling. First, we detail a powerful new WaveNet-style autoencoder model that conditions an autoregressive decoder on temporal codes learned from the raw audio waveform. Second, we introduce NSynth, a large-scale and high-quality dataset of musical notes that is an order of magnitude larger than comparable public datasets. Using NSynth, we demonstrate improved qualitative and quantitative performance of the WaveNet autoencoder over a well-tuned spectral autoencoder baseline. Finally, we show that the model learns a manifold of embeddings that allows for morphing between instruments, meaningfully interpolating in timbre to create new types of sounds that are realistic and expressive.

In this paper, we outline a data-driven approach to audio synthesis. Rather than specifying a specific arrangement of oscillators or an algorithm for sample playback, such as in FM Synthesis or Granular Synthesis (Chowning, 1973; Xenakis, 1971), we show that it is possible to generate new types of expressive and realistic instrument sounds with a neural network model. Further, we show that this model can learn a semantically meaningful hidden representation that can be used as a high-level control signal for manipulating tone, timbre, and dynamics during playback.

Explicitly, our two contributions to advance the state of generative audio modeling are:

- A WaveNet-style autoencoder that learns temporal hidden codes to effectively capture longer term structure without external conditioning.
- NSynth: a large-scale dataset for exploring neural audio synthesis of musical notes.

00 | NSynth Super

- Magenta collaborated with Google Creative Lab.

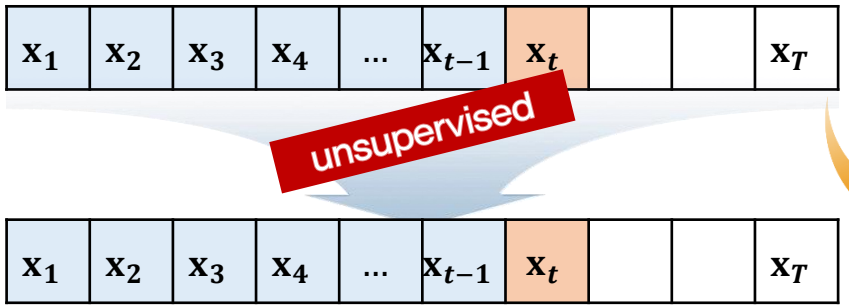
- https://www.youtube.com/watch?v=iTXU9Z0NYoU&t=1s

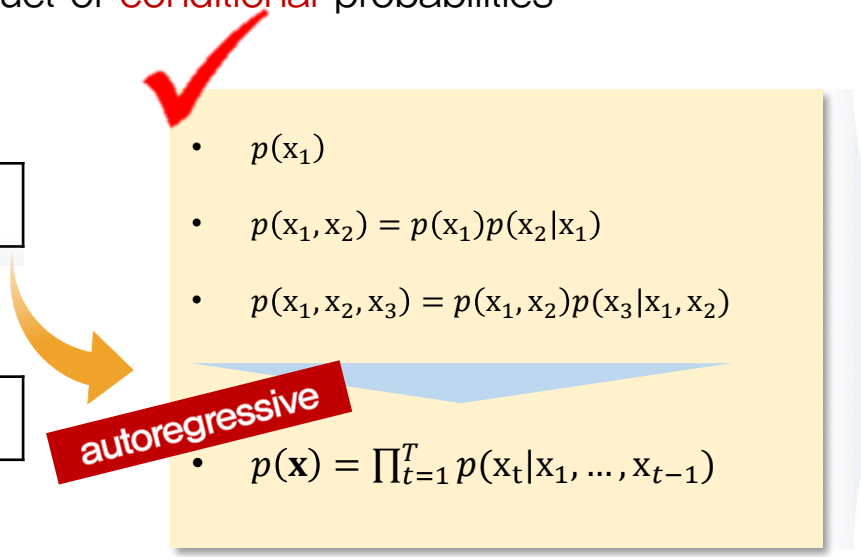
Contents

Contents

01 Autoregressive Model

- Audio 스스로의 분포를 학습하는 unsupervised learning
- Audio-sequence data의 statistical dependency를 반영하여 모델링을 해야 함
- T 시점까지의 audio waveform: $x = \{x_1, \dots, x_T\}$
- $p(x_t)$는 x_1부터 x_{t-1}까지의 데이터가 주어졌을 때의 conditional probability로 표현할 수 있음: autoregressive
  → $p(x_t) = p(x_t|x_1, x_2, \dots, x_{t-1})$
- Joint probability is factorized as a product of conditional probabilities

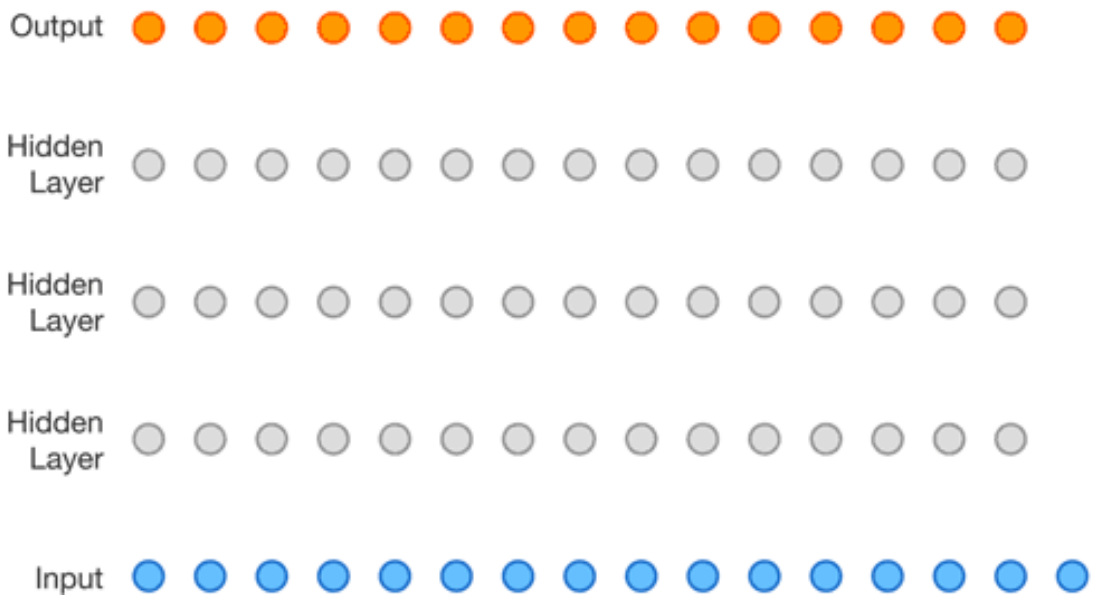


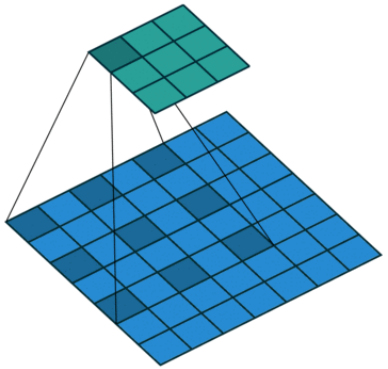


Stack of Convolution Layers로 표현하는 것
》 WaveNet

01 Dilated Causal Convolution

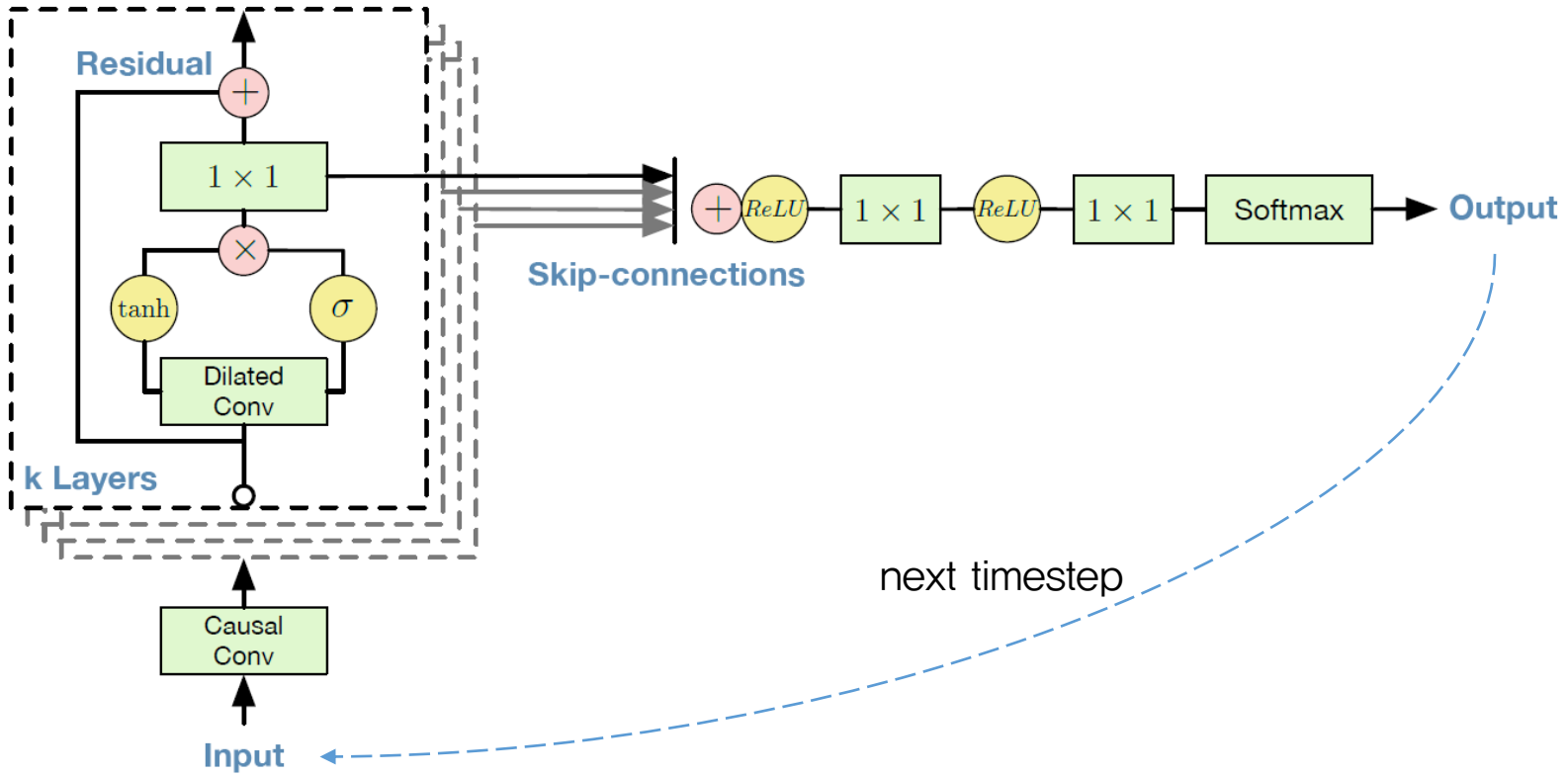
- Causal convolution: 미래의 값을 알 수 없기 때문에 t 시점 이후의 값들은 사용할 수 없다
  (t 시점의 input은 $t-1$ 시점의 output을 포함)
- Dilated: 적은 계산량으로 큰 receptive field를 얻음 (512 dilation → 1024 receptive field)





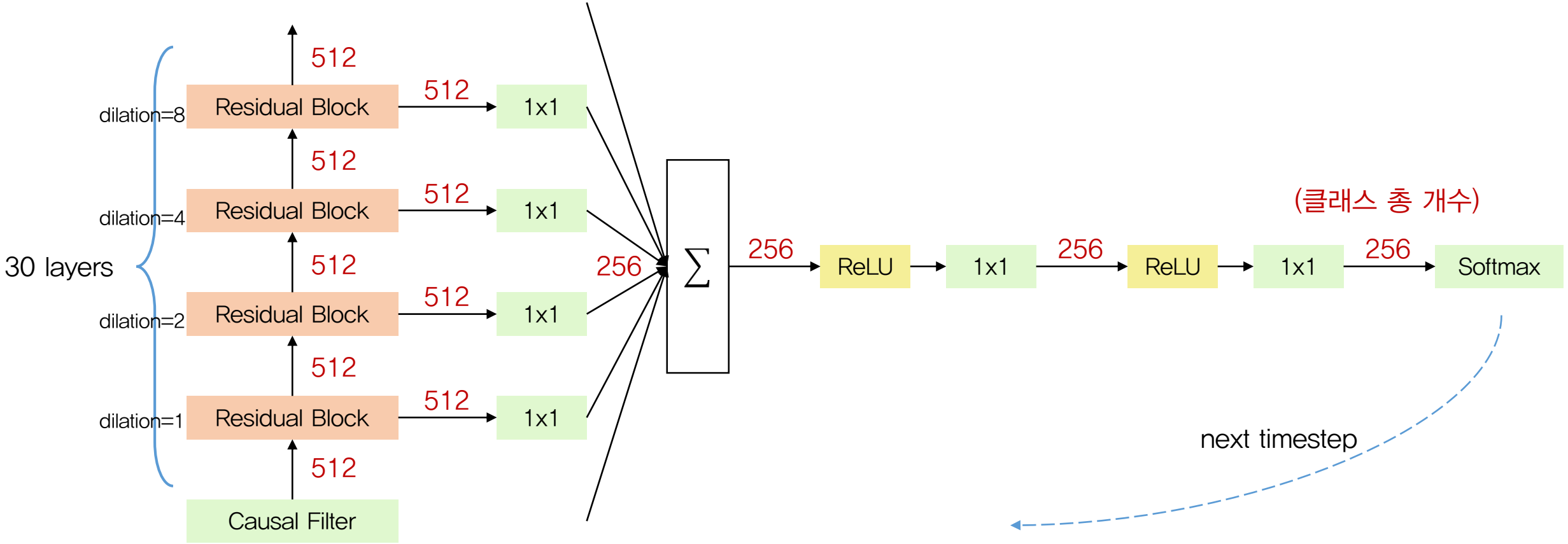
01 Residual and Skip Connections

- WaveNet: 한 timestep에서



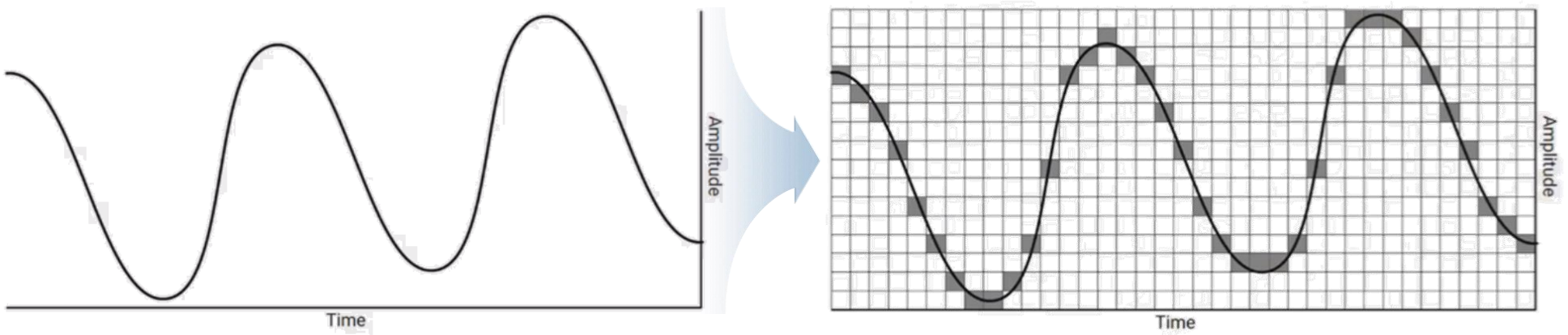
01 Residual and Skip Connections

- WaveNet: 한 timestep에서 펼쳤을 때 모습



01 Softmax Distributions

- 일반적으로 audio의 distribution을 표현하기 위해, Gaussian Mixture Model 등을 사용 (continuous)
- PixelRNN, PixelCNN과 유사하게, 각 amplitude 값을 하나의 class로 가정하여 softmax distribution을 사용함 (discrete)
- 음성 데이터를 256개 클래스로 묶어 regression 문제를 multi-class classification 문제로 변환



analogue: waveform

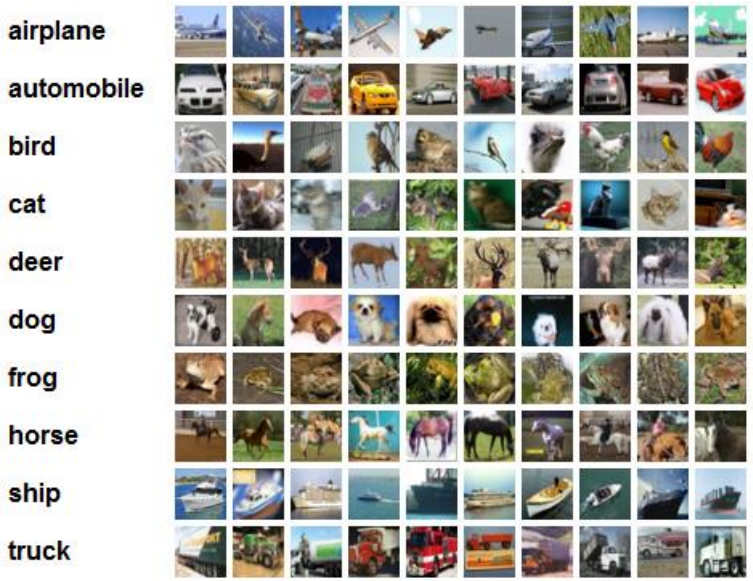
digital(computer): quantization

Contents

02 | Motivations

- Recent breakthroughs in generative modeling of images: high-quality and large-scale datasets
- Audio signals in the wild contain multi-scale dependencies → constrained domains

  * 사람이 직접 기타를 친 소리를 학습했을 경우 잡음이나 깨지는 소리가 많이 생성됨



[CIFAR 10]

[MNIST]

02 Dataset Description

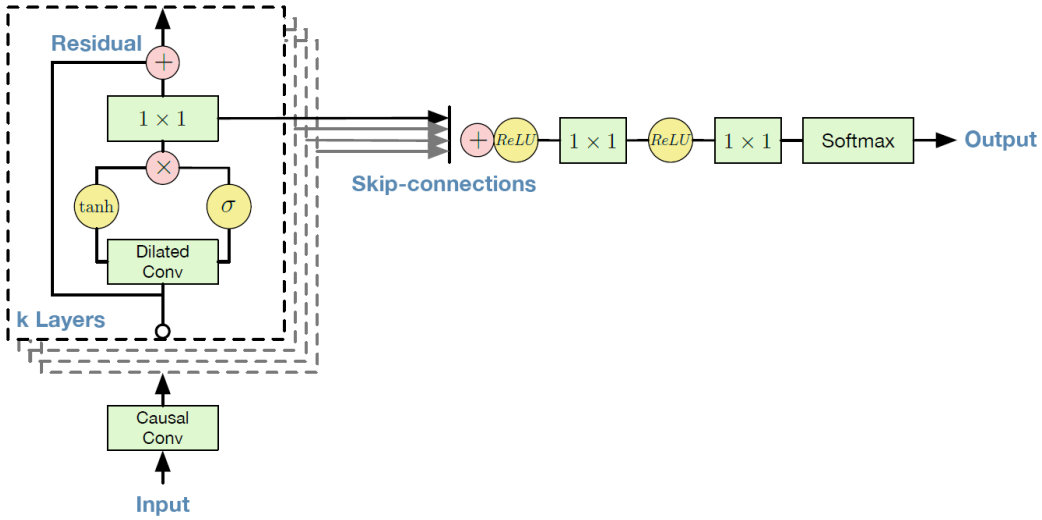
- 한 음에 대한 정보를 포함하고 있는 데이터셋
- 4 seconds, 16kHz → 64,000 data points
- 305,979 musical notes
- Unique pitch(음 높이), timbre(음색), and envelope(파형을 둘러싸듯이 그려진 선)
- Train – 289,205 & Validation – 12,678 & Test – 4,096
- Instrument, pitch, qualities, sample rate (samples or data points per second), source 등 다양한 정보를 제공

Family	Acoustic	Electronic	Synthetic	Total
Bass	200	8,387	60,368	68,955
Brass	13,760	70	0	13,830
Flute	6,572	35	2,816	9,423
Guitar	13,343	16,805	5,275	35,423
Keyboard	8,508	42,645	3,838	54,991
Mallet	27,722	5,581	1,763	35,066
Organ	176	36,401	0	36,577
Reed	14,262	76	528	14,866
String	20,510	84	0	20,594
Synth Lead	0	0	5,501	5,501
Vocal	3,925	140	6,688	10,753
Total	108,978	110,224	86,777	305,979

Contents

03 Motivations

- WaveNet & SampleRNN와 같은 autoregressive models에서 영감을 받아 개발
- 비교적 짧은 오디오 (~500ms)에 대해서는 좋은 성능을 보였으나, 긴 오디오를 생성하는데 있어서는 external condition에 많은 영향을 받는 것으로 확인됨



[WaveNet]

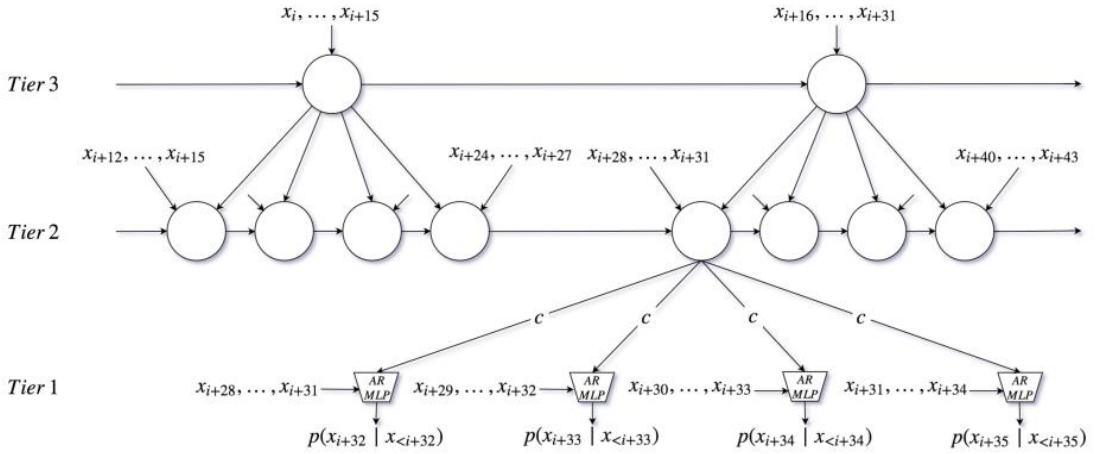


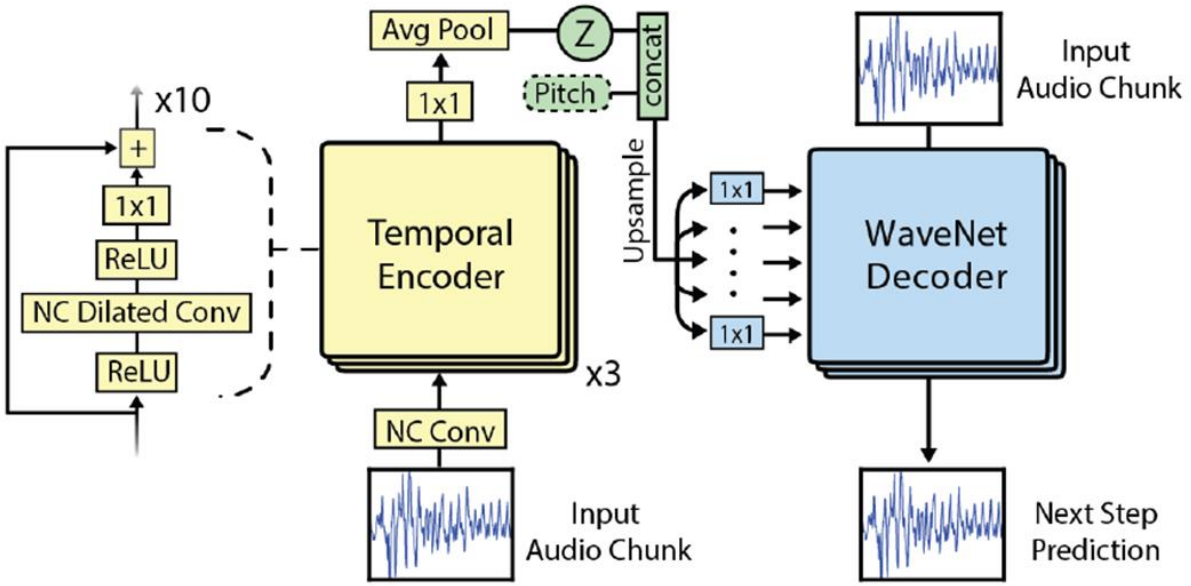
Figure 1: Snapshot of the unrolled model at timestep i with $K = 3$ tiers. As a simplification only one RNN and up-sampling ratio $r = 4$ is used for all tiers.

[Sample RNN]

03 Architecture

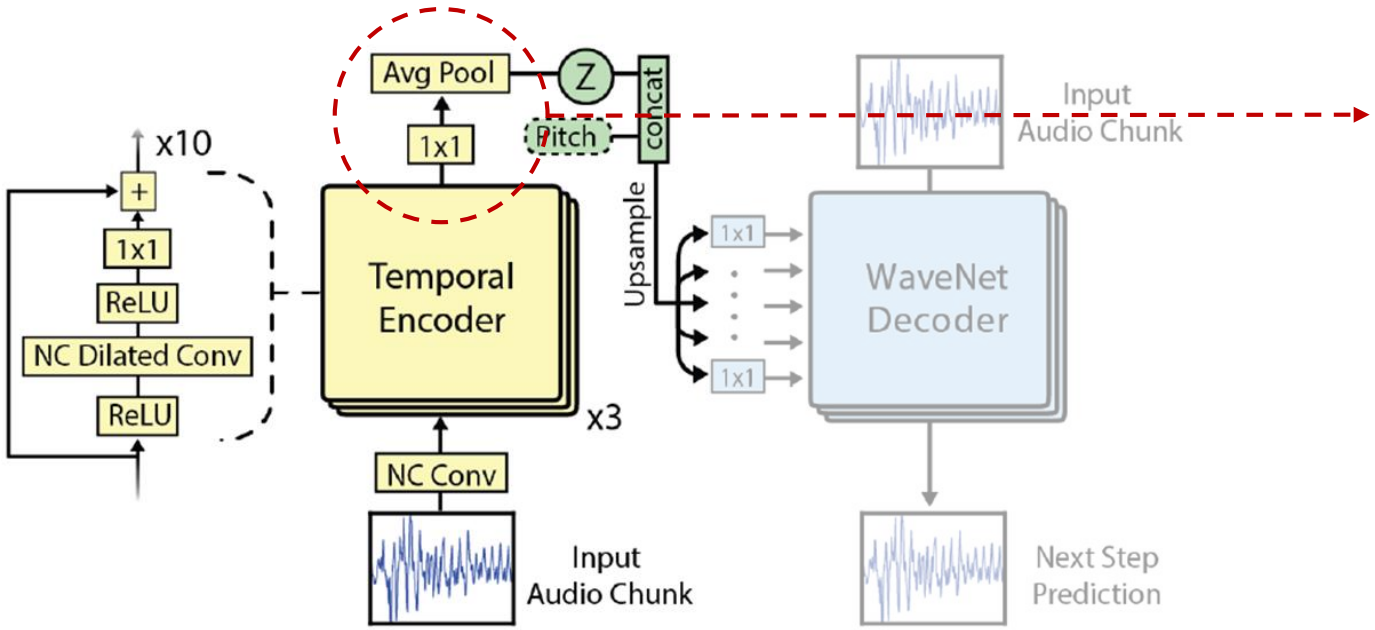
- Goal: 오토인코더 구조를 사용해서 external condition 없이 시그널 데이터를 학습 및 생성

  → embedding layer가 역할을 대신함
- (WaveNet-like) Encoder: infers hidden embeddings distributed in time
- (WaveNet) Decoder: use the embeddings to effectively reconstruct the original audio



03 Temporal Encoder

- 데이터의 latent variable을 학습하기 위해 causal convolution을 사용하지 않음
- Parameterized skip-connections를 사용하지 않음



① 1x1 Conv: 16차원 (channel) embedding

② Avg Pool: stride를 사용해서 temporal resolution을 결정할 수 있음

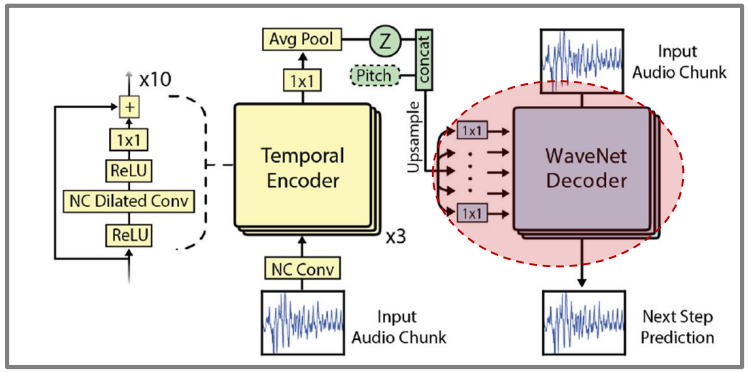
ex) 16,000 samples/sec인 4초짜리 오디오의 embedding을 구하려면?

→ 64,000 samples / 512 = 125

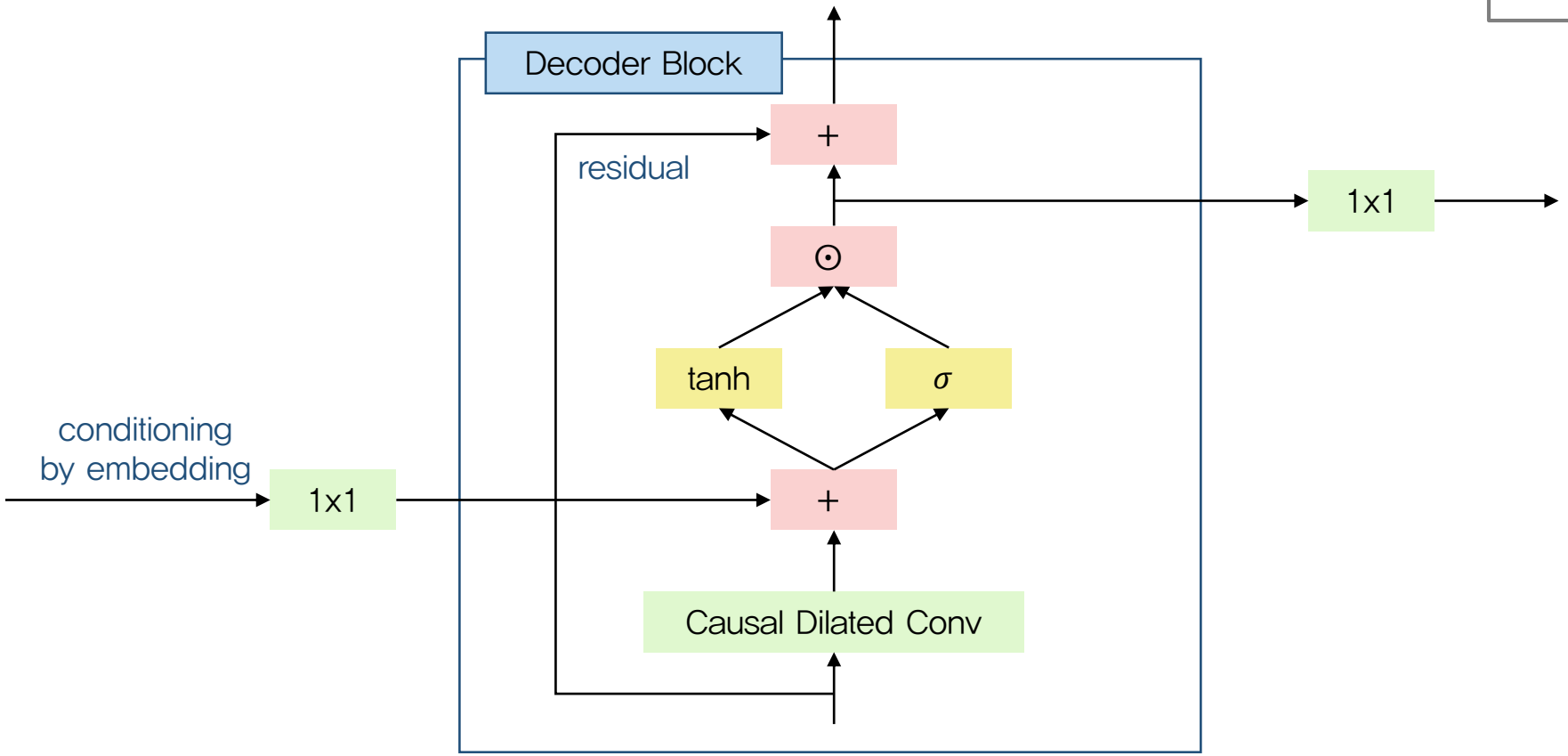
→ 125 x 16 embedding matrix

→ 오디오 압축률 = 64,000 / (125 x 16) = 32

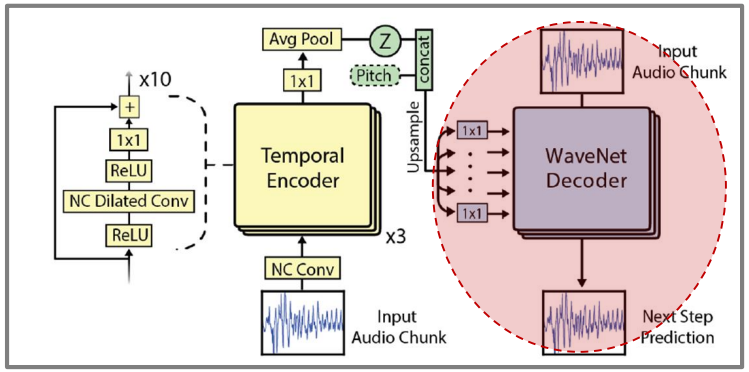
03 WaveNet Decoder



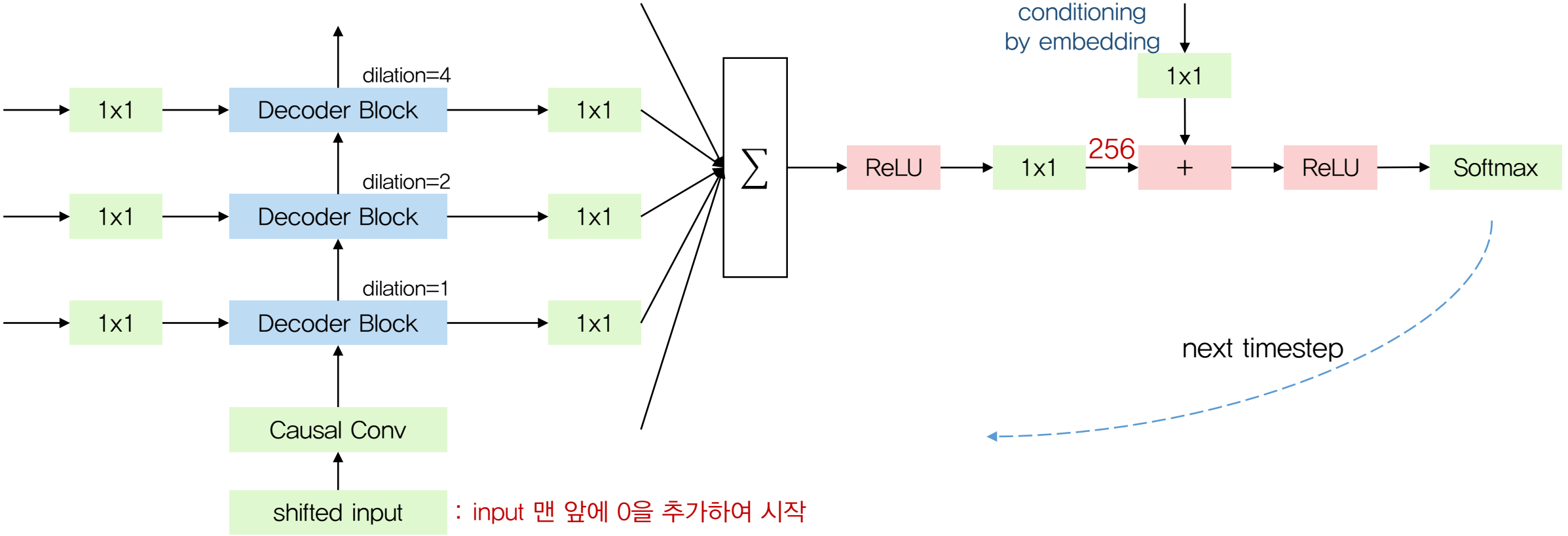
- WaveNet의 block과 같은 구조를 띄고 있으며, embedding matrix로부터 conditioned



03 WaveNet Decoder



- 매 시점마다 아래 과정을 반복: computationally expensive
- pitch: one-hot encoding이지만 실제로 사용하지는 않음

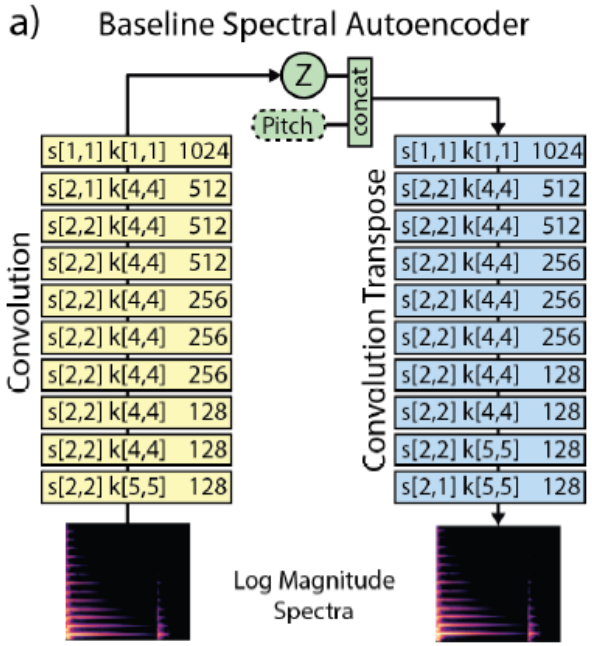


03 Experiments

- Audio is notoriously hard to represent visually
- Magnitude spectrograms: 오디오 분석에서 나오는 특징 대부분을 잡을 수 있으나, 비슷한 형태를 띄더라도 실제로 들었을 때 아예 다른 소리가 나올 수도 있음

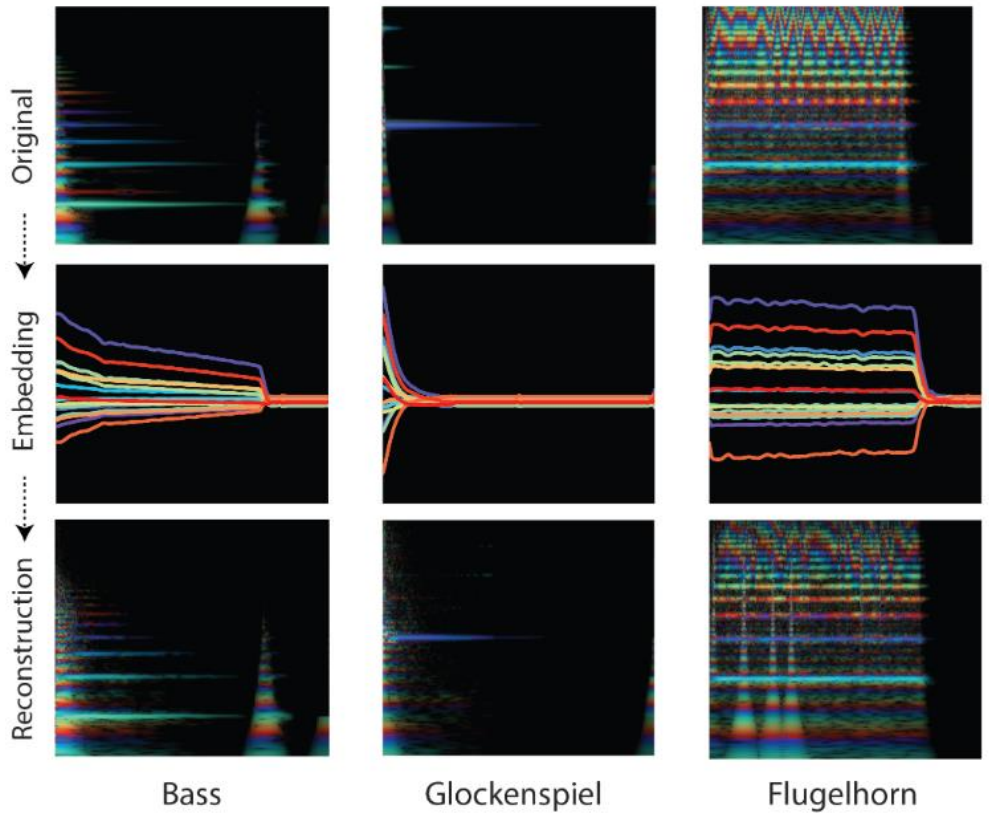
- Base model: convolutional autoencoder

- Adam optimizer, batch size = 32, epochs = 250,000



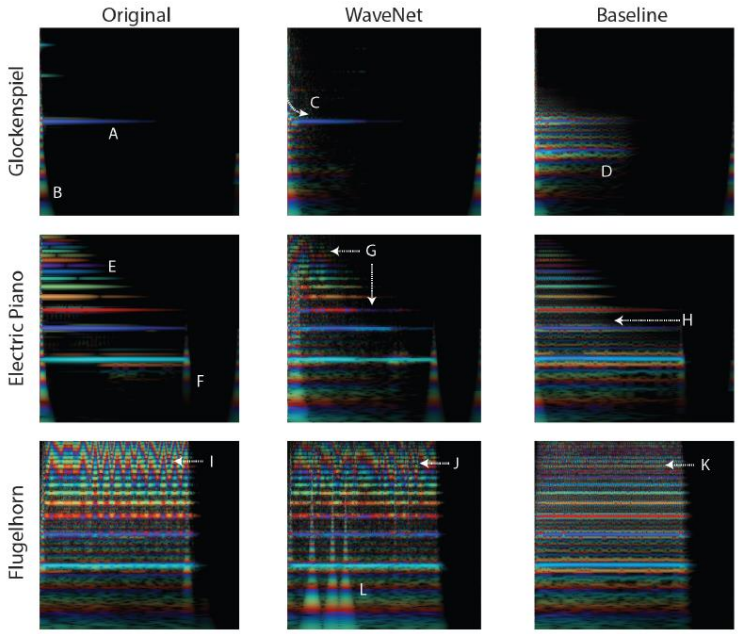
03 Experiments

- 크게 2가지 측면에서 실험을 진행
  - ① Reconstruction
  - ② Embedding



03 Experiments: Reconstruction

- CQT (constant-Q transform) spectrograms
  - → vertical = frequency
  - → intensity = magnitude
  - → horizontal = time
  - → color = instantaneous frequency



- Classification Model
  - → baseline encoder와 같은 구조를 사용
  - → 재구축 데이터를 기반으로 pitch, quality를 분류

Table 1. Classification accuracy of a deep nonlinear pitch and quality classifier on reconstructions of a test set.

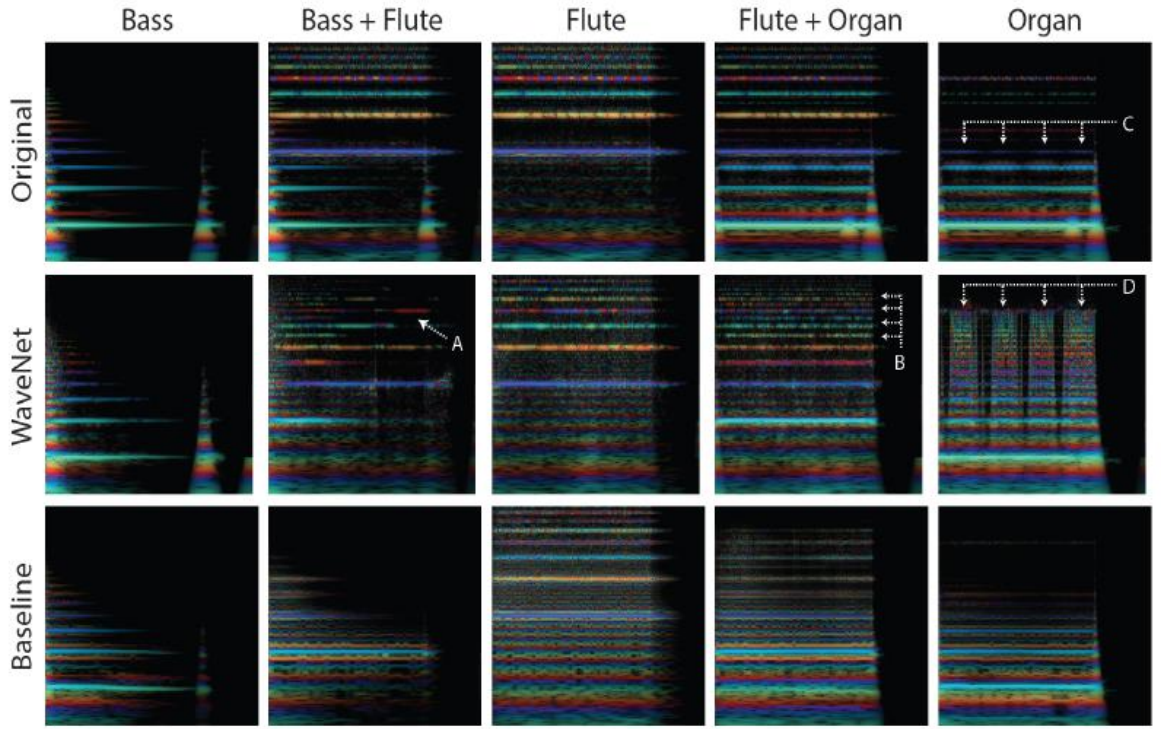
	Pitch	Quality
Original Audio	91.6%	90.1%
WaveNet Recon	79.6%	88.9%
Baseline Recon	46.9%	85.2%

03 Experiments: Embedding

- Linear interpolations between notes

  → Original: 단순 평균. 두 개의 음이 동시에 들리는 효과

  → Generative models: interpolate embeddings → combine semantic aspects



03 Links

- https://magenta.tensorflow.org/nsynth
- https://magenta.tensorflow.org/nsynth-fastgen

Contents

04 Other Research

- Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset
- Published as a conference paper at ICLR 2019
- Magenta team
- Using decoder of Transformer

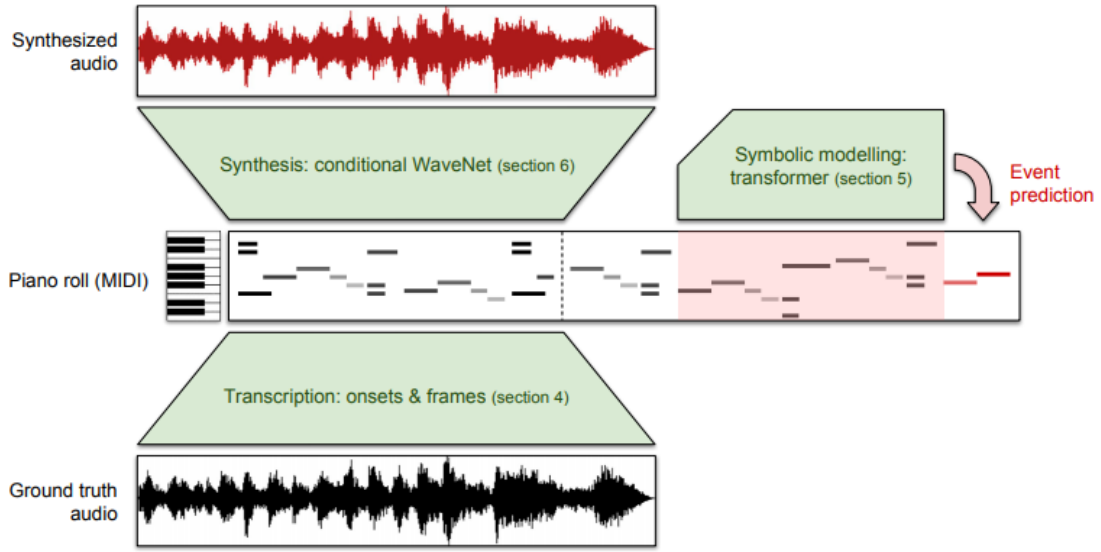


Figure 1: Wave2Midi2Wave system architecture for our suite of piano music models, consisting of (a) a conditional WaveNet model that generates audio from MIDI, (b) a Music Transformer language model that generates piano performance MIDI autoregressively, and (c) a piano transcription model that "encodes" piano performance audio as MIDI.

- https://magenta.tensorflow.org/maestro-wave2midi2wave

04 Other Research

- A Universal Music Translation Network
- Facebook AI Research
- Data augmentation, universal encoder and multiple decoders: DeepFake와 유사

  * DeepFake보다 도메인 수가 많다는 것에서 차이점이 있음

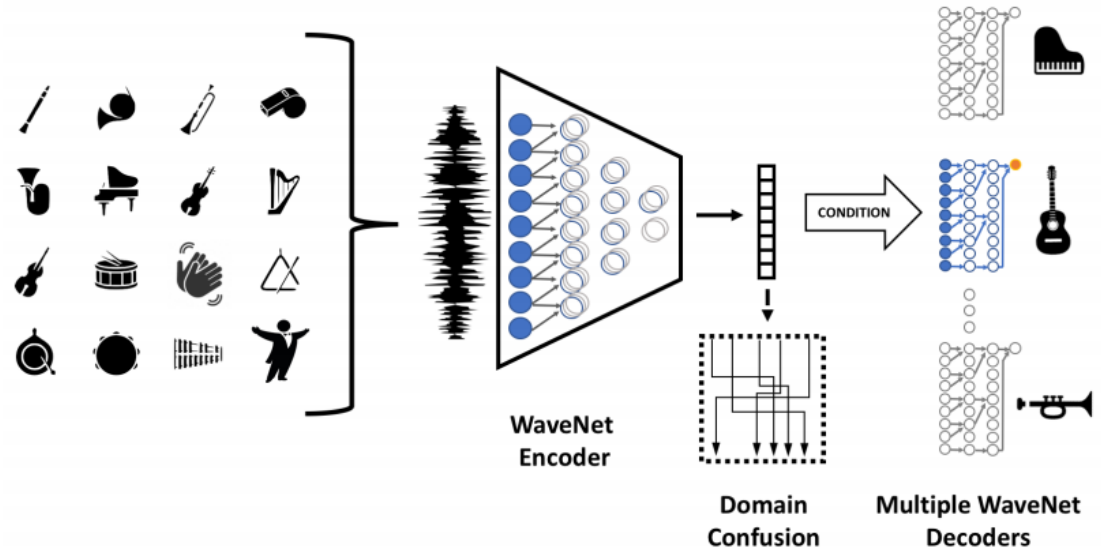


Figure 1: The architecture of our network. The confusion block (dashed line) is employed only during training.


Thank
you