SING: Symbol-to-Instrument Neural Generator

FAIR[1] / École Normale Supérieure[2] / INRIA[3] / PSL Research University[4]

Alexandre Défossez[1,2,3,4], Neil Zeghidour[1,2,3,4], Nicolas Usunier[1], Léon Bottou[1] and Francis Bach[2,3,4]

facebook Artificial Intelligence Research

inria inventors for the digital world

Abstract

SING is a deep learning based music notes synthetizer that can be trained on the NSynth dataset. NSynth is composed of 300,000 notes from over 1,000 instruments. Each note is a 4 seconds long waveform sampled at 16kHz. We obtain state-of-the-art results compared to the NSynth wavenet-like autoencoder [1][2] as measured by Mean Opinion Scores based on human evaluations, for a model that is 32 times faster to train and 2,500 faster for inference.

- We generate directly a waveform and introduce a differentiable regression spectral loss based on the log-power spectrogram of the generated audio.
- Architecture based on standard modules: LSTM-based sequence generator with a convolutional decoder.
- Specific pre-training procedures based on matching the embedding obtained from an autoencoder.
- State-of-the-art MOS (Mean Opinion Scores) and ABX on the NSynth dataset.
- Input is disentangled representation of the pitch, instrument and velocity. Generalizes to unseen combination of pitch and instrument.

https://github.com/facebookresearch/SING

NSynth dataset

300,000 notes from a 1000 instruments, all pitches at 5 velocities (= intensity). Each note $x_{V,I,P} \in [-1, 1]^{64{,}000}$ is 4 seconds at 16,000 Hz indexed by a triplet $(V, I, P) \in \{0, \dots, 4\} \times \{0, \dots, 1005\} \times \{0, \dots .120\}$. For each instrument, keep 10% of the pitches for the test set.

Reconstruction losses

We want to evaluate the distance between the generated waveform $\hat{x}$ and the ground truth x. Either MSE on the waveform:

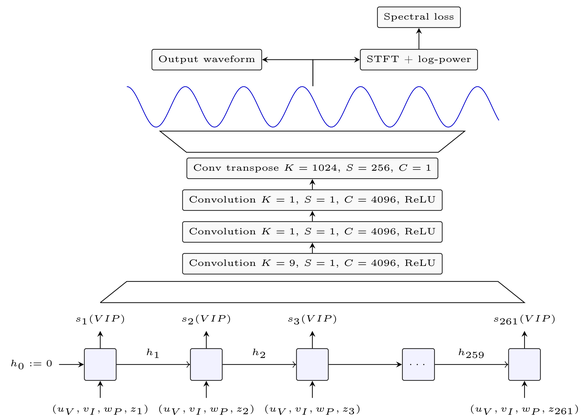
$$\mathrm{L}_{\mathrm{wav}}(x, \hat{x}) := \|x - \hat{x}\|^2,$$

or using a spectral loss:

$$\mathrm{L}_{\mathrm{stft},1}(x, \hat{x}) := \|l(x) - l(\hat{x})\|_1$$

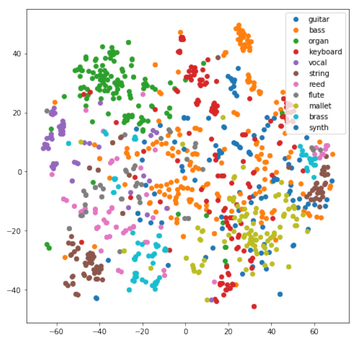
where $l(x) := \log\left(\epsilon + |\mathrm{STFT}[x]|^2\right)$.

Architecture



- Input of the LSTM: embeddings $(u_V, v_I, w_P) \in \mathbb{R}^2 \times \mathbb{R}^{16} \times \mathbb{R}^8$ from look-up tables, time embedding $z_T \in \mathbb{R}^4$.
- Output = temporal representation of the sound $s_i(VIP) \in \mathbb{R}^{128}$ at 62.5 Hz. Convolutional decoder upsample it from 62.5Hz to 16kHz.

Instrument embeddings



Instrument embeddings from the look-up table projected in 2D using T-SNE [3].

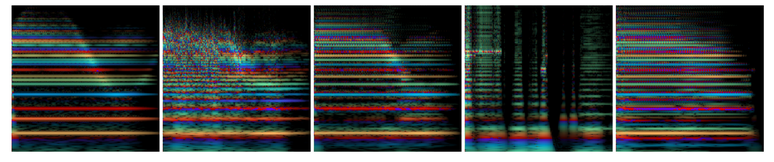
Training

I - autoencoder: take the symetric of the decoder and train an autoencoder for 50 epochs: 12 hours on 4 GPUs.
II - sequence generator: match the output of the LSTM-based RNN with the output of the frozen encoder using MSE. 50 epochs using truncated backpropagation with length 32, takes 10 hours.
III - end-to-end sing: fine tune the whole architecture end-to-end for 20 epochs, takes 8 hours on 4 GPUs.
In total: 30 hours on 4 GPUs to train.

Ablation study

		Spectral loss		Wav MSE	
Model	**training loss**	*train*	*test*	*train*	*test*
Autoencoder	waveform	0.026	0.028	0.0002	0.0003
SING	waveform	0.075	0.084	0.006	0.039
Autoencoder	spectral	0.028	0.032	N/A	N/A
SING	spectral	0.039	0.051	N/A	N/A
SING no time embedding	spectral	0.050	0.063	N/A	N/A

Comparison of generated rainbowgrams



From left to right: ground truth, nsynth, SING with spectral loss, SING with wavform loss, SING with spectral loss and no time embedding. Rainbowgram [1] computed from the waveform, the intensity of the color is proportional to the log-power spectrogram while the color itself encode the derivative of the phase. The vertical axis represents frequencies in logarithmic scale, horizontal axis is time.

Human evaluations

Model	MOS	Training time (hrs * GPU)	Generation speed	Compression factor	Model size
Ground Truth	3.86 ± 0.24	-	-	-	-
Wavenet	2.85 ± 0.24	3840^*	0.2 sec/sec	32	948 MB
SING	3.55 ± 0.23	120	512 sec/sec	2133	243 MB

(*): adjusted to account for difference in FLOPs of GPUs used.

MOS: for each model, 100 samples are evaluated by 60 humans on a scale from 1 ("Very annoying and objectionable distortion") to 5 ("Imperceptible distortion") using Crowdmos toolkit [4] for removing outliers.
ABX: Ask 10 humans to evaluate for a 100 examples if Wavenet or SING is closest to ground truth. 69.7% are in favor of SING over Wavenet.

References

[1] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders. Technical Report 1704.01279, arXiv, 2017.
[2] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. Technical Report 1609.03499, arXiv, 2016.
[3] L. Maaten, G. Hinton. Visualizing Data using t-SNE. In JMLR 2008.

[4] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer. Crowdmos: An approach for crowdsourcing mean opinion score study. ICASSP 2011.