Google Research

Multi-instrument Music Synthesis with Spectrogram Diffusion

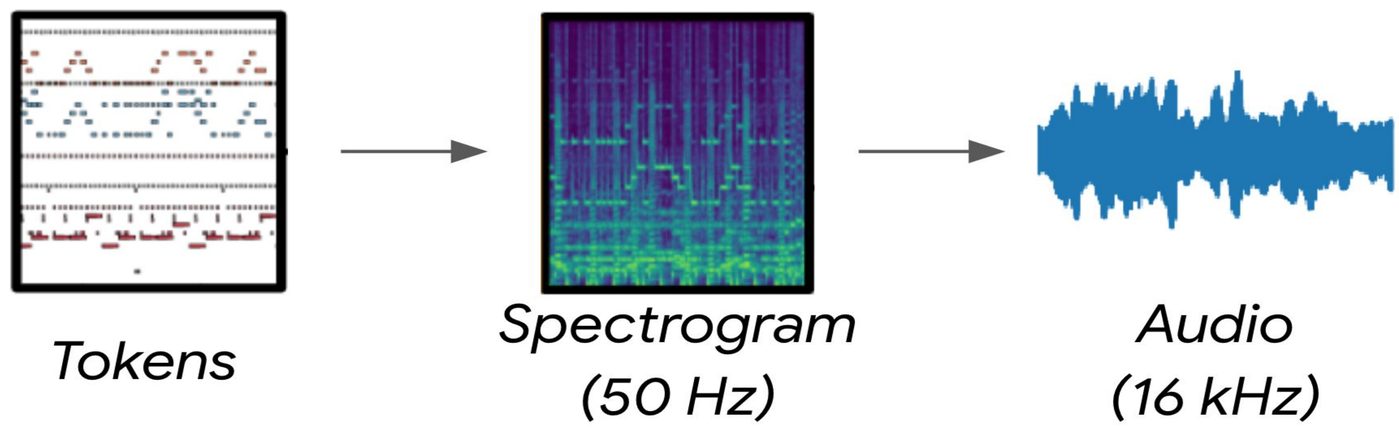
Curtis Hawthorne, Ian Simon, Adam Roberts, Neil Zeghidour, Josh Gardner, Ethan Manilow, Jesse Engel
{fjord, iansimon, adarob, neilz, jpgard, emanilow, jesseengel}@google.com

Overview

Music synthesis model with MIDI as input, audio as output:

- Note-level pitch and instrument control
- No specific restrictions on polyphony or number of instruments
- Trains on any dataset with paired MIDI and audio
- Synthesizes tracks of arbitrary length
- Realtime inference speed

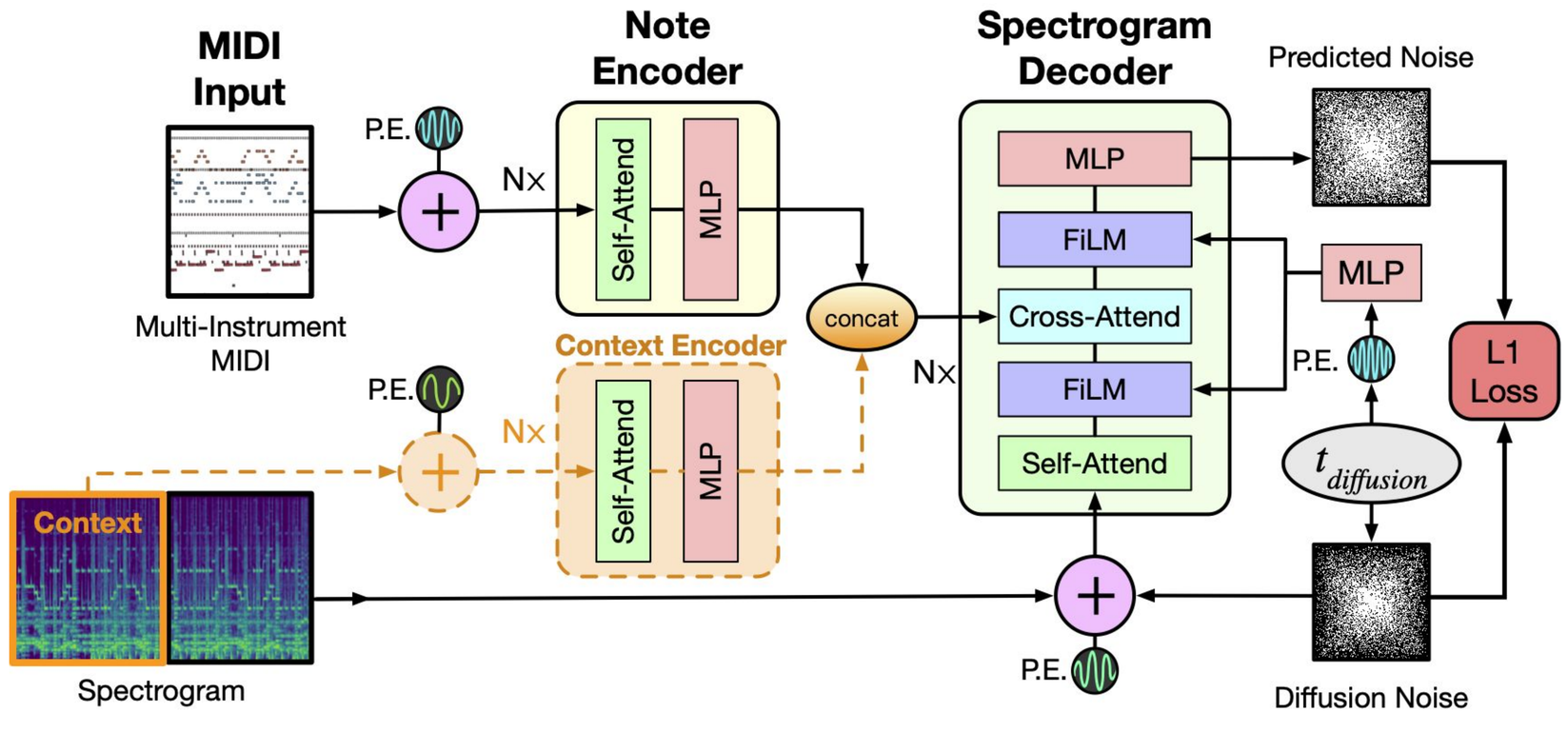
Tokens → Spectrogram, Spectrogram → Audio



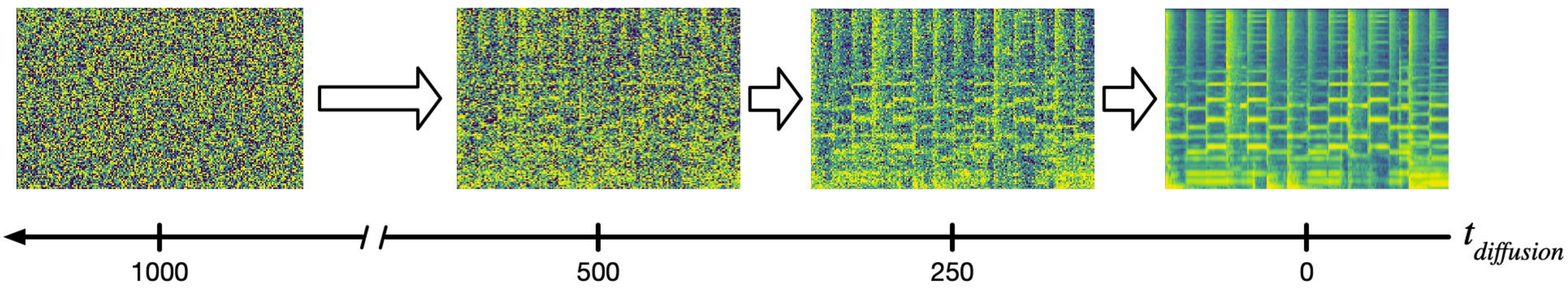
The best diffusion models achieve high fidelity from several stages. We take a similar approach by using a DDPM to predict spectrograms and training a separate GAN spectrogram inverter to generate audio.

T5-Style Transformer Encoder-Decoder Architecture

The first encoder stack takes a sequence of note events as input. We train the decoder stack as a Denoising Diffusion Probabilistic Model (DDPM), where the model learns to iteratively refine Gaussian noise into a target spectrogram. We generate ~5 second spectrogram segments, so to ensure a smooth transition between these segments we (optionally) encode the previously generated segment in a second encoder stack.



Spectrogram Diffusion Process Example



Links

Audio Examples https://g.co/magenta/spec-diff-ex	Render your own MIDI in Colab https://g.co/magenta/spec-diff-demo	Code and Pretrained Models https://g.co/magenta/spec-diff-code