

Deep Learning for Audio

Lecture 10

Pavel Severilov

AI Masters

2024

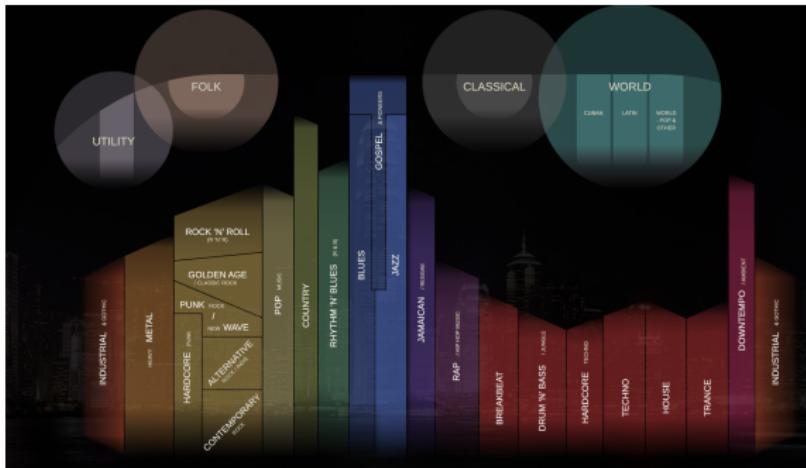
Outline

1. Music generation
2. Jukebox
3. Diffsound
4. MusicLM
5. AudioCraft

Outline

1. Music generation
2. Jukebox
3. Diffsound
4. MusicLM
5. AudioCraft

Problems with generating music



- ▶ Should NNs compose music by following the same logic and process as humans do?
- ▶ Are current evaluation methods good enough to compare and measure the creativity of the composed music?
- ▶ Difficult to collect data due to copyright
- ▶ Music is subjective
- ▶ Each music Genre has its own rules

Datasets

- ▶ JSB Chorales Dataset (chorales by Johann Sebastian Bach)
- ▶ Maestro Dataset (200 hours of virtuosic piano performances with fine alignment between note labels and audio waveforms)
- ▶ The Lakh MIDI Dataset (176,581 unique MIDI files)
- ▶ MetaMIDI Dataset (436,631 MIDI files)

Labs, apps, ideas

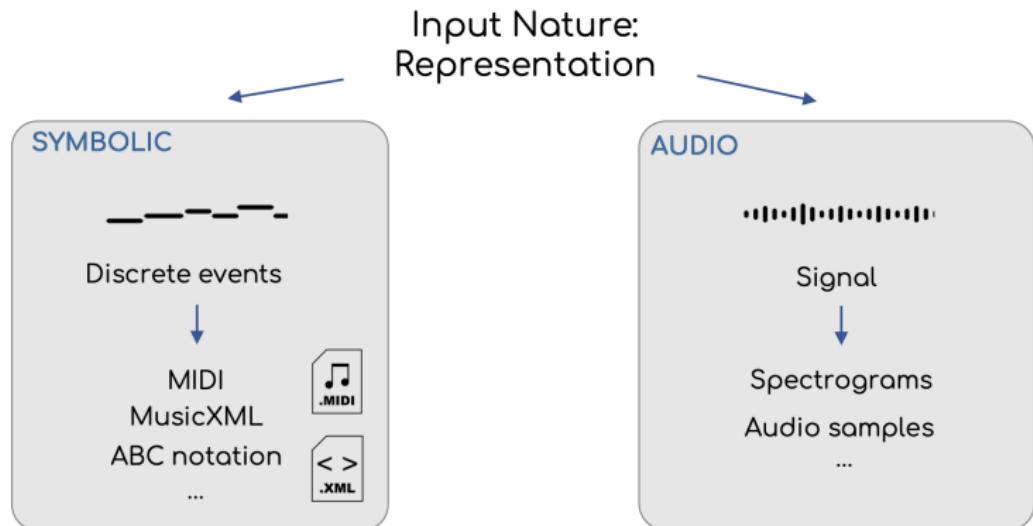
Research Groups and Labs:

- ▶ Google Magenta
- ▶ Audiolabs Erlangen
- ▶ Music Informatics Group
- ▶ Music and Artificial Intelligence Lab
- ▶ Metacreation Lab

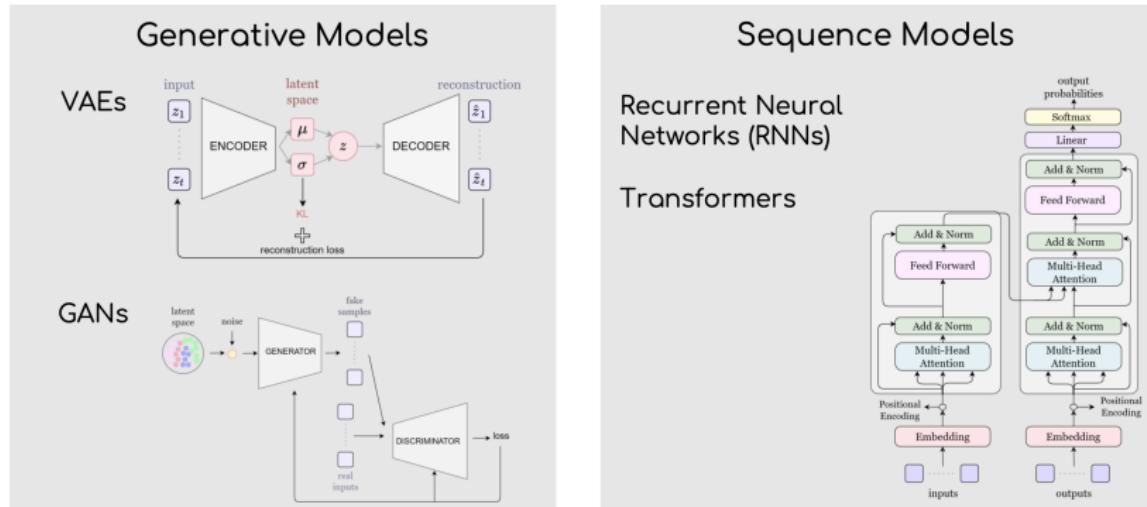
Apps:

- ▶ AIVA
- ▶ Amper Music
- ▶ Ecrett Music
- ▶ Humtap
- ▶ Amadeus Code
- ▶ Computoser
- ▶ Brain.fm

Input representations



History: before 2019



MuseNet

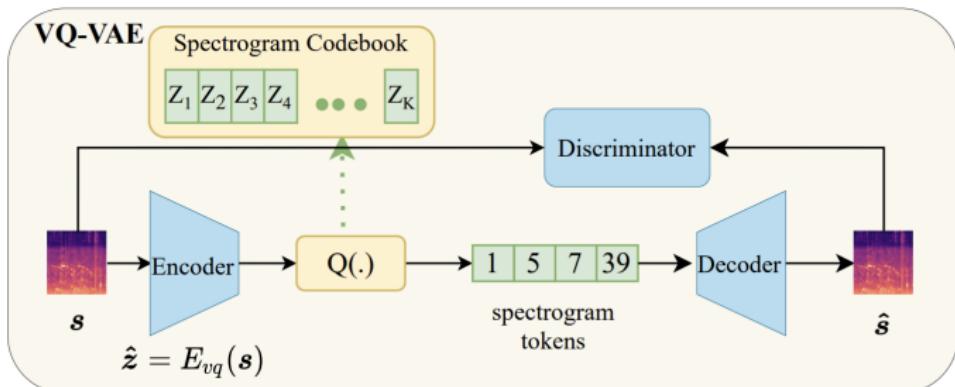


- ▶ Learn to predict the next token
- ▶ Uses GPT-2
- ▶ MIDI format
- ▶ Sparse Transformer to train a 72-layer network with 24 attention heads
- ▶ Long context: full attention over a context of 4096 tokens

Outline

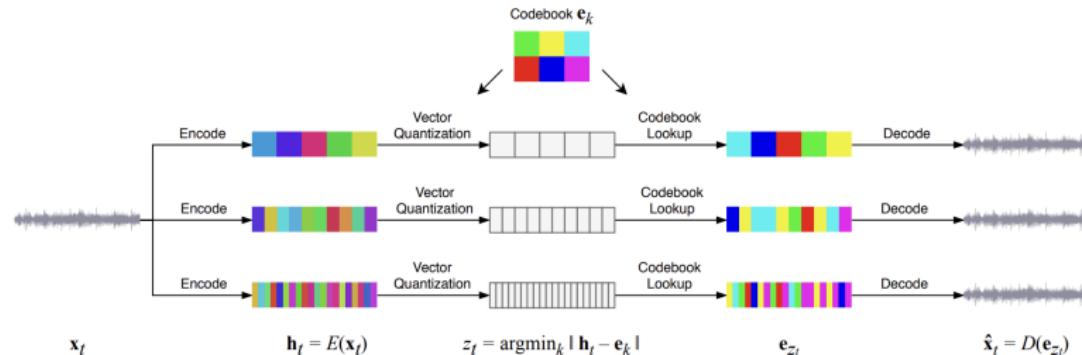
1. Music generation
2. Jukebox
3. Diffsound
4. MusicLM
5. AudioCraft

VQ-VAE



- ▶ Encoder extracts the representation z from the mel-spectrogram
- ▶ Codebook contains a finite number of embedding vectors
- ▶ Decoder reconstructs the mel-spectrogram based on mel-spectrogram tokens
- ▶ Discriminator distinguishes the mel-spectrogram is original or reconstructed
- ▶ $Q(\cdot)$ denotes a spatial-wise quantizer that maps each features z_{ij} into its closest codebook entry z_k

Jukebox

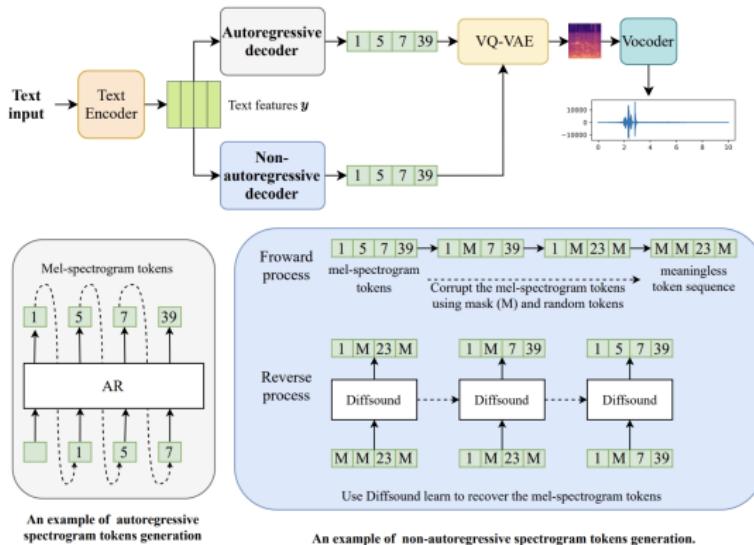


- ▶ Based on VQ-VAE
- ▶ Separated autoencoders with different temporal resolutions
- ▶ Spectral loss

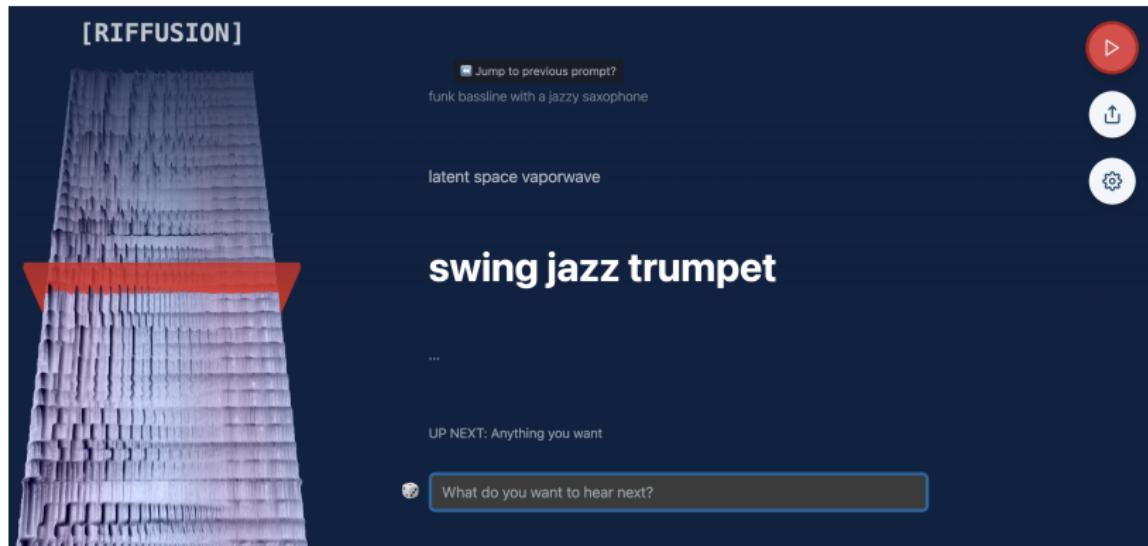
Outline

1. Music generation
2. Jukebox
3. Diffsound
4. MusicLM
5. AudioCraft

Diffsound



- ▶ Text encoder (CLIP) extracts text features from the text input
- ▶ Decoder generates mel-spectrogram tokens
- ▶ Pre-trained VQ-VAE transforms the tokens into mel-spectrogram
- ▶ Vocoder transforms mel-spectrogram into waveform

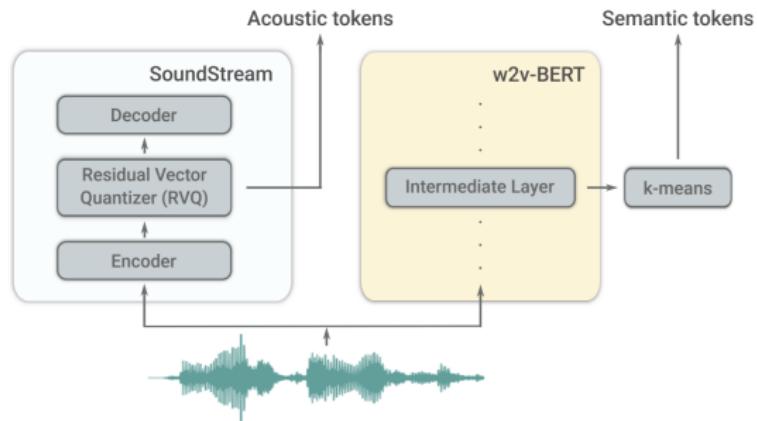


- ▶ Real-time music generation with stable diffusion
- ▶ Fine-tuned the model to generate images of spectrograms
- ▶ Combine short clips by sampling the latent space between a prompt with two different seeds, or two different prompts with the same seed

Outline

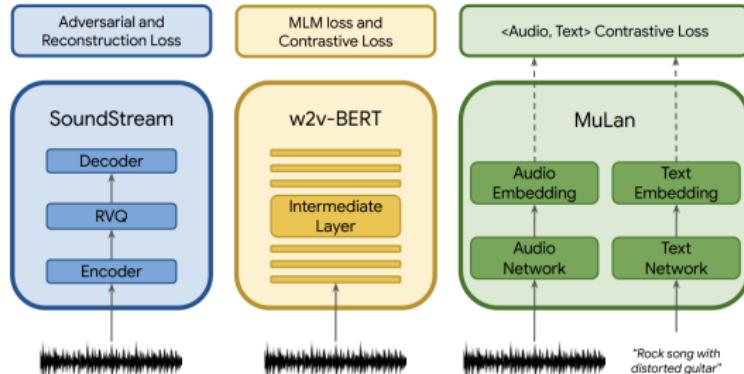
1. Music generation
2. Jukebox
3. Diffsound
4. MusicLM
5. AudioCraft

AudioLM



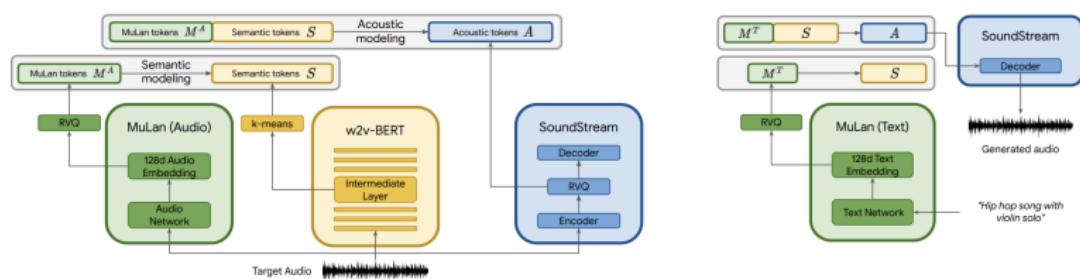
- ▶ SoundStream: 24 kHz audio to 50 Hz embeddings (Autoencoder codec with vector quantizers)
- ▶ w2v-BERT: extract embeddings from the 7th layer and quantize them using the centroids of a learned k-means over the embeddings

MusicLM



- ▶ Based on AudioLM: SoundStream + w2v-BERT + MuLan
- ▶ MuLan tokens
- ▶ Metrics Frechet Audio Distance (FAD), KL

MusicLM: training



► training

Suno: Bark

The screenshot shows the Suno app interface. At the top left is the Suno logo. Below it is a section titled "Explore new styles of music with Suno". A subtitle reads "Here's a small taste (or whatever the listening equivalent is) of what's possible". In the center, there's a button labeled "Pick a style, or roll the dice...". Below this are three small icons: a volume icon, a previous track icon, a dice icon, and a next track icon. To the right of the dice icon is a large grid of text labels representing various music styles. The styles listed include:

- accordion
- french dubstep
- bedroom pop semis
- j-pop blues
- lo-fi afro house
- drum and bass
- accordion drill
- disco chillsynth
- afrobeat rock
- dance
- acoustic slushwave
- russian flamenco
- hip hop baba
- french dubmow
- Japanese swing
- carnatic glitch hop
- classical baba
- bedroom pop new jack swing
- boogie caribbean
- p-funk samba
- psychedelic acid trance
- j-pop acid jazz
- liquid drum and bass slushwave
- gospel disco
- harpsichord klezmer
- dakar raga
- disco chilstep
- drum and bass acoustic rock
- ambient trance alternative rock
- accordion delta blues
- acoustic rockabilly
- afrobeat new jack swing
- sitar sertanejo
- french big band
- japanese surf rock
- russian dubstep
- hip hop techno
- rumba cape verdean
- bedroom pop
- boogie
- carnatic acid trance
- classical
- rumba
- j-pop acid breaks
- liquid drum and bass rockabilly
- p-funk mariachi
- psybient new wave
- rumba
- dakar new wave
- disco alternative r&b
- drill sertanejo
- gospel acid breaks
- harpsichord hip hop
- accordion ambient techno
- acoustic rock chillsynth
- afrobeat griot
- ambient techno mento
- french afro-rock
- japanese surf
- russian celtic
- hip hop
- sitar rumba
- barbershop breakbeat
- blues rock american primitivism
- carnatic
- city pop symphonic metal
- indie grunge
- liquid drum and bass new jack swing
- p-funk jazz
- psybient griot
- roots reggae house
- dakar math rock
- dirty south boom bap
- drill regga
- goa trance afro-cuban jazz
- harpsichord g-funk
- accordion afro trap
- acoustic rock afro-jazz
- afrobeat garage
- ambient techno hyphy
- fife and drum blues
- p-funk
- japanese surf

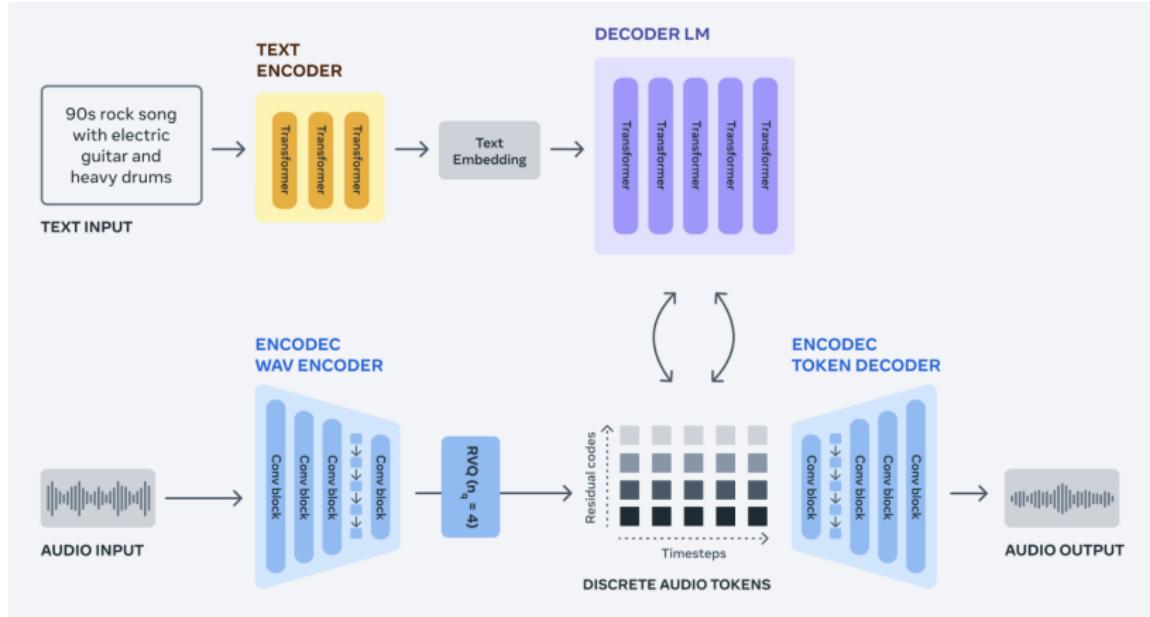
At the bottom left is a pink button with the text "What will you create?".

- architecturally very similar to Google's AudioLM

Outline

1. Music generation
2. Jukebox
3. Diffsound
4. MusicLM
5. AudioCraft

AudioCraft



- ▶ model from Meta