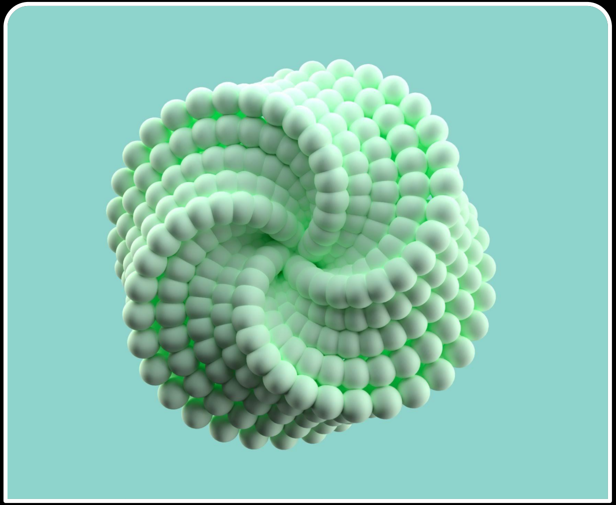


CS109b

Transformer-based music generation

Group Members:
Yanyan Zeng
Bright Liu
Dewey To
Anthony Rodriguez-Miranda
John Vizhco-Leon

TF:
Daniel Nurieli

Outline

01

Introduction and Problem Statement

Problem Statement

Although transformer models were originally used for text-based tasks, how effective are they at interpreting and generating music?

Our goal is to take advantage of the self-attention mechanism in Transformers and adapt it to generate the next pitch, duration, and instrument to create coherent music comparable to the input data.

02

EDA & Visualizations

The Data:

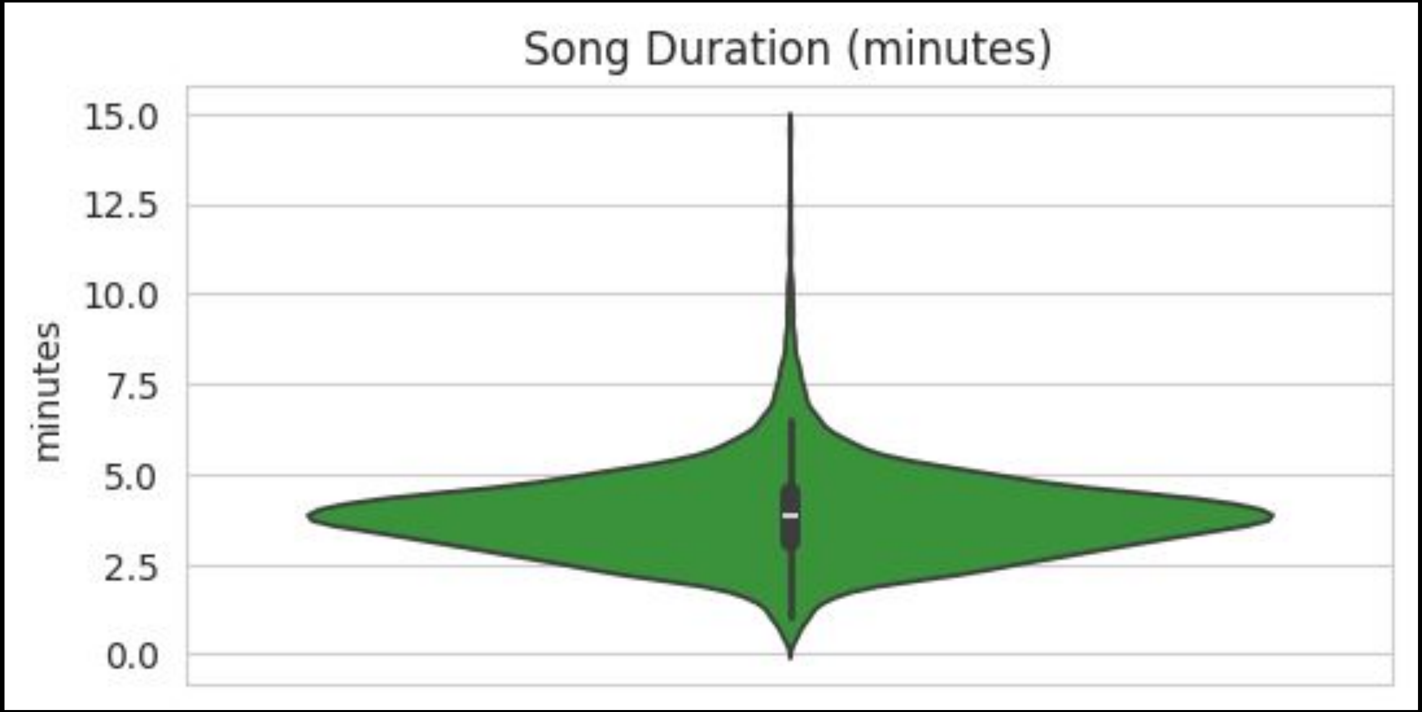
The Lakh MIDI dataset (Clean MIDI subset)

- MIDI file: “Musical Instrument Digital Interface” file, a more efficient way of storing sound data than traditional audio files
- Lakh MIDI: A collection of ~176k unique MIDI files
- Clean MIDI subset: A subset of MIDI files that include artist names and titles and uses ~17k fully uncorrupted MIDI files

Sources: https://umatechnology.org/what-is-midi-and-what-are-midi-files/
https://colinraffel.com/projects/lmd/

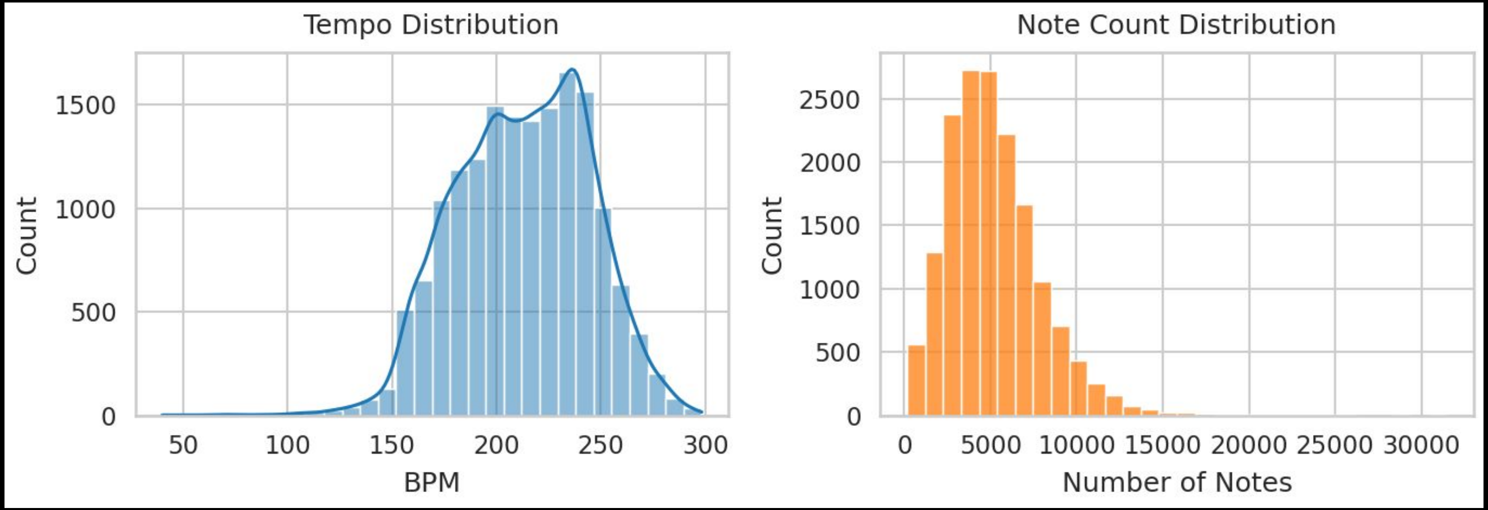
Song Duration

There's some rather pronounced right-skew in song duration.



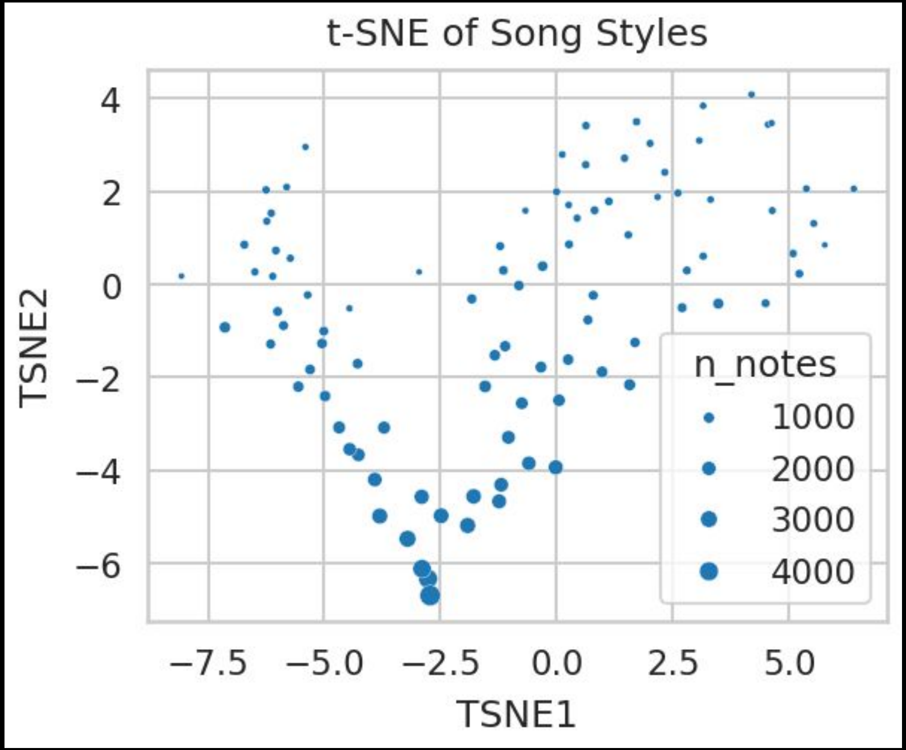
Notes and Tempo

There's a similar right skew in note-count, and there are one (or maybe two?) visual peaks in tempo. Perhaps we can include a tempo token in the future?



t-SNE of 100 Songs

Possible clusters corresponding to genres? Maybe we can use these clusters to improve coherence in the future.



03

Modeling/ Training Details

Evaluation Metrics

- Beyond accuracy in predicting the next pitch, we also calculated some basic evaluation metrics.
- **Consonance Score**: Percentage of the intervals between notes that are consonant (e.g., perfect fifths, major thirds)
- **Chord Recognition Rate**: Percentage of groups of simultaneously played notes that match standard chords
- **Pitch Class Entropy**: How spread out is the usage of pitch classes in a piece of music? Low pitch class entropy tends to be associated with high tonality, and high pitch class entropy tends to be associated with atonality.

Baseline Transformer

Beyond the initial input and position-encoding layers, we sandwiched a multi-head attention layer with LayerNorm, two Dense layers, and another LayerNorm layer. Repeating this sandwich twice then produced the general transformer.

Performance metrics:

Consonance score	0.0
Chord recognition rate	1.0
Pitch class entropy	2.075

output

Baseline Transformer Training Details

Data	MIDI files parsed into pitch sequences of length 128, 10043 total windows
Epochs	2
Training Time	8 minutes
Hyperparameters	sparse_categorical_crossentropy
Other details	Positional encoding via learned Embedding(SEQ_LEN, EMBED_DIM) Metric is accuracy

Advanced Transformer

- Similar architecture to the baseline model, but added pitch duration and instrument type (in addition to pitch) as inputs and outputs.

Metric	Score
val_pitch_accuracy	0.1981
val_duration_loss	0.2317
val_program_loss	1.8700
consonance_score	0.449
chord_recognition_rate	1.0
pitch_class_entropy	1.833

output

Advanced Transformer Training Details

Data	Truncated: 1000 songs. Converted to TF dataset and split into train/val sets.
Epochs	10
Training Time	~5 mins
Hyperparameters	Batch size = 128, learning rate = 0.001 (Adam default)
Other details	Loss function: validation pitch accuracy Used early stopping - but stopped at epoch 10

References:

- Dong et al., 2023: https://arxiv.org/pdf/2207.06983
- Hsiao et al., 2021: https://github.com/YatingMusic/compound-word-transformer?tab=readme-ov-file#readme
- MIDI-GPT: https://www.metacreation.net/projects/midi-gpt

04

Results/ Conclusions

Baseline Transformer

Metric	Score
consonance_score	0.0
chord_recognition_rate	1.0
pitch_class_entropy	2.75

Advanced Transformer

Metric	Score
consonance_score	0.449
chord_recognition_rate	1.0
pitch_class_entropy	1.833

"I'm Not in Love" by 10cc

Metric	Score
consonance_score	0.382
chord_recognition_rate	1.0
pitch_class_entropy	3.315

Conclusions:

- Adding duration and instrument variation makes a large impact.
- Although the AT didn't achieve a very high pitch accuracy, its consonance score vastly improved from the baseline model and was higher than the real song.
- Not much data was needed (1000 songs) to get a real-sounding output.

05

Future Work/ Improvements

Transfer Learning – Hugging Face MusicGen

- Keeping it short- It failed :(
- Preparing dataset – Midi -> Wav files, with empty text entry
- Fine Tuning with a dataset of 1000 sampled
  - 42000 seq len at 32 hz
- Error:
  - `Starting training…`
  - `Error during training: CUDA out of memory. Tried to allocate 143.05 GiB. GPU 0 has a total capacity of 21.96 GiB of which 18.44 GiB is free.`
- Incomplete audio
  - Prompt: "A gentle piano melody with soft background strings"

Output

Future Work and Improvements

Harmonic & Polyphonic Generation	Expressive Performance Modeling	Genre–Aware Style Control & Evaluation
Feed chord progressions into the model for a clear harmonic roadmap Expand from single–note to true polyphonic prediction (multiple voices/instruments)	Include note durations and velocities so the model learns expressive timing and loudness Apply advanced positional encodings to capture motifs and thematic arcs	Label each track by its t–SNE – derived genre cluster and fine–tune per style Run listening tests to ensure our objective metrics align with human perception

Thank you

Models

Baseline Transformer

After turning our data into time-ordered pitch sequences, we trained a standard Transformer to predict the next pitch.

Advanced Transformer

We modified this standard Transformer so that it predicts duration and instrument type in addition to pitch.