Faster Music Generation with Dual Pattern Speculative Decoding

Dongkyun Kim

Carnegie Mellon University
School of Computer Science

Motivation

Challenges of Autoregressive Music Generation

- Autoregressive transformer-based decoders paired with quantized audio tokenizers are the dominating method for music generation models due to their high fidelity reconstruction capabilities.
- The autoregressive decoding process, however, leads to the inference speed of these models scaling to the number of codebooks.

Gaps in Existing Solutions

- Methods utilizing parallel codebook interleaving patterns have been presented
- Their implementation is impeded by the challenges of considering the conditional dependence between different codebook levels.

Overview

Our approach consists of a novel inference strategy for incorporating dual pattern speculative decoding.

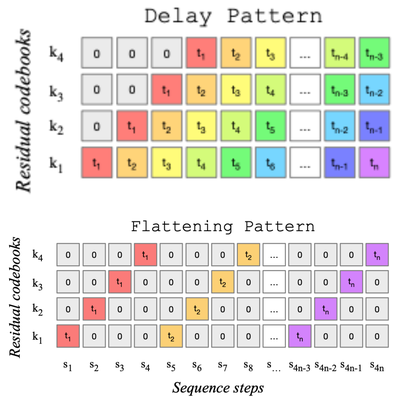
Speculative decoding is a technique widely used in large language models, where a smaller draft model predicts potential future sequences and the original model validates it, thus taking advantage of parallelism.

Dual pattern speculative decoding extends this concept by utilizing the delayed codebook pattern model as the draft model while utilizing the flatten codebook pattern model as the validator. Thus, the method effectively leverages parallel codebook interleaving without sacrificing the high generation capabilities of sequential methods.

Methods

1. Generation Model Pre-training

Objectives: Train two autoregressive transformer models with different codebook interleaving patterns.



2. Speculative Decoding Inference

Objectives: Utilize the delay pattern model as a draft model and the flattening pattern model as a validator.

Key Innovations: The use of dual pattern model allows 2-3x inference speed from parallel codebook interleaving while maintain ing similar generation quality of flattened patterns.

3. Codebook Level Specific Speculation

Objectives: We observe a higher level of disagreement in the higher codebook levels, due to the uncertainty. To account for this fact, we apply an increasing temperature scaling from 0.3 to 0.7 for the draft model probability in levels k2 to k4.

Key Innovations: The use of codebook-specific temperature accounts for the increasing level of uncertainty in higher levels, leading to faster inference speed.

Experiments

Draft Model	Candidate Model	Inference Speed
N/A	musicgen-large-flatten	55.2 seconds
N/A	musicgen-large-delay	16.2 seconds
musicgen-large-delay	musicgen-large-flatten	28.4 seconds
musicgen-small-delay	musicgen-large-delay	13.3 seconds

Experiments done on A4000 GPU

Dual pattern speculative decoding leads to a 2x speedup compared to a flattened model. We also conducted experiments with two delay pattern model of different sizes, which leads to a 20% speed up compared to a single model.

Future Directions

The generation quality of the speculative decoding music generation still requires testing. Conducting a more thorough evaluation of the music generation quality will allow me to ensure that the speed improvements have not compromised the quality of the generated music, allowing a potential publication on the topic.

Conclusion

- Our work offers new pathways for improving the performance of music generation systems.
- We conduct extensive experiments to verify the advantages of various data balancing techniques and self-training.
- The solution allows real-time, high-fidelity music generation, opening up new possibilities for interactive music applications, live performance aids, and more efficient music composition tools.