2024/10/08

- Yenan

Long-form music generation with latent diffusion

[https://arxiv.org/html/2404.10301v1]

Overview:
・previous work: generate long duration was a difficult task (10-30s, up to 90s)
→ contribution: generate long-form music(up to 4m45s) with good consistency
・base line: **MusicGen-large-stereo** → bad consistency
・ours also can generate Short-form music(or sound)

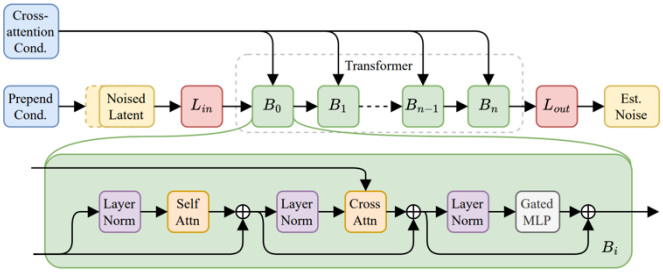

Figure 2: Architecture of the diffusion-transformer (DiT). Cross-attention includes timing and text conditioning. Prepend conditioning includes timing conditioning and also the signal conditioning on the current timestep of the diffusion process.

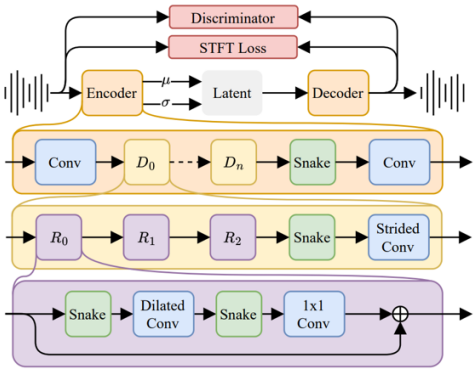

Figure 3: Architecture of the autoencoder.

Text encoder + Auto encoder + DiT(diffusion-transformer)

⭐Thanks to auto encoder, latent rate can be smaller to generate long-form music.

2 training stages (to generate a 4m45s music):

①pretrain (3m10s) ②fine-tune (4m45a)

2024/10/08

	sampling rate	STFT distance ↓	MEL distance ↓	SI-SDR ↑	latent rate	latent (channels)
DAC [23]	44.1kHz	0.96	0.52	10.83	86 Hz	discrete
AudioGen [24]	48kHz	1.17	0.64	9.27	50 Hz	discrete
Encodec [8, 22]	32kHz	1.82	1.12	5.33	50 Hz	discrete
AudioGen [24]	48kHz	1.10	0.64	8.82	100 Hz	continuous (32)
Stable Audio [14]	44.1kHz	1.19	0.67	8.62	43 Hz	continuous (64)
Ours	44.1kHz	1.19	0.71	7.14	21.5 Hz	continuous (64)

Evaluation metrics of quality of long-form music: FD, KL

	channels/sr	output length	FD_{openl3} ↓	KL_{passt} ↓	$CLAP_{score}$ ↑	inference time
MusicGen-large-stereo [8]	2/32kHz	2m	204.03	0.49	0.28	6m 38s
Ours (fully-trained)	2/44.1kHz	2m†	79.09	0.35	0.40	13s
MusicGen-large-stereo [8]	2/32kHz	4m 45s	218.02	0.50	0.27	12m 53s
Ours (fully-trained)	2/44.1kHz	4m 45s	81.96	0.34	0.39	13s

- Minghao

Integrating Text-to-Music Models with Language Models: Composing Long Structured Music Pieces

[https://arxiv.org/html/2410.00344v2]
[https://www.researchgate.net/publication/384563720_Integrating_Text-to-Music_Models_with_Language_Models_Composing_Long_Structured_Music_Pieces]

・contribution: generate up to 2m30s length music by integrating LLM with text-to-music model

2 challenges of this method:

① how to align the LLM with the text-to-music model.

Fine-tune or in-context learning are generally considered

→ using in-context learning to instruct ChatGPT to generate prompts for MusicGen, with providing 50 song descriptions from Pond5

2024/10/08

② how to introduce ChatGPT because ChatGPT tends to mix multiple genres in the prompts for a single piece, resulting in structures that are not particularly well-designed.

→ using frameworks like the ITPRA theory and using a chain of thought approach

⭐Difference from previous works = prompt to instruct MugicGen

Evaluation metrics:

Self-Similarity Matrices ↓ & subjective comparison by musician and non-musician