Generating Japanese Pop Music by Building On MusicGen

Dihong Huang, Akaash Dash, Saikanam Siam

CS 4644 / 7643 Deep Learning Fall 2023 – Professor Danfei Xu

INTRODUCTION

We fine-tuned MusicGen, a state-of-the-art AI model for music generation, using a custom J-pop dataset. Transfer learning helps retain existing knowledge while adding new insights tailored to this genre. We also tried to apply Low-Rank Optimization techniques to improve memory efficiency but failed due to difficulty in coding. Our goal is to test how MusicGen can be tailored to specific genres and generate more personalized music on users' interests. Although the results are not satisfying, we still made significant progress.

DATA

To finetune the model, we curated a collection of music from 29 J-pop singers, groups, and bands, some of which were top-ranked in 2023, alongside personal favorites. The dataset consists of two part: audio clips and corresponding metadata in JSONL format.

Metadata	Value
key	F#
artist	Sawano Hiroyuki
sample rate	44100
file extension	wav
description	Electronic, Industrial, ... Synth-pop song
duration	30.0 s
bpm	188
genre	Electronic, ... Synth-pop
title	Sawano_Hiroyuki_Sheut...158.wav
instrument	synthesizer, drums, ... drummachine
moods	epic, energetic, ... love
path	/content/drive/MyDrive/...chunk150.wav

Table 1. Example format of a music clip

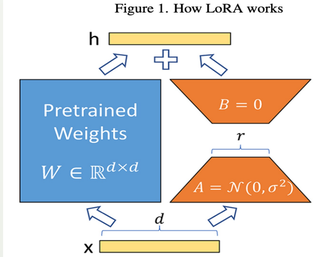
Data collection and processing:

- Download songs from YouTube playlists with yt_dlp, over 700 songs
- Audio splitting, train-test split, and autolabeling with essentia
- Meta data adjustment if necessary

TECHNICAL APPROACH

Our method is fine tuning MusicGen on a specific dataset to achieve a more tailored music generation. This is based on transfer learning, by initializing the training on a pretrained checkpoint we add task-specific knowledge to the model with less training data. In addition, we propose to apply Low-Rank Optimization for Approximate (LoRA) on this task to further improve memory efficiency. The implementation of our approach is in two stages: full fine tuning and LoRA fine tuning. LoRA freezes the pretrained weights of specific layer(s) and trains on a pair of low-rank adaption matrices instead. These adaptors will then multiply together to approximate the weight matrix. The baseline we use to compare our finetuned results will be the pretrained model, in our case the MusicGen 32khz small model.

Figure 1. How LoRA works



Evaluation metrics:

- Human ear test
- Cross entropy loss
- Perplexity

Parameter	Value
dataset.num_workers	2
dataset.valid.num_samples	1
dataset.batch_size	2
schedule.cosine.warmup	8
optim.optimizer	adamw
optim.lr	1e-4
optim.epochs	3
optim.updates_per_epoch	2000
optim.adam.weight_decay	0.01

Table 4. Some Hyperparameters

EXPERIMENTS AND RESULTS

We have set up two sets of experiments: one with a small dataset for preliminary results and one with full data for final report. With the configs demonstrated above, we conducted multiple training run and the best metrics are reported here:

Data	Time	LR	Grad Norm	Grad Scale	Cross-Entropy (ce)	PPL
200/2000	0.71 it/sec	2.28E-05	2.471E+00	19807.522	3.565	40.718
400/2000	0.74 it/sec	1.86E-05	2.436E+00	32768.000	3.449	37.391
600/2000	0.75 it/sec	1.47E-05	2.503E+00	32768.000	3.508	38.015
800/2000	0.76 it/sec	1.12E-05	2.498E+00	32768.000	3.562	39.101
1000/2000	0.76 it/sec	8.10E-06	2.423E+00	32768.000	3.364	34.233
1200/2000	0.76 it/sec	5.48E-06	2.454E+00	32768.000	3.408	34.957
1400/2000	0.77 it/sec	3.35E-06	2.465E+00	32768.000	3.419	34.767
1600/2000	0.77 it/sec	1.73E-06	2.452E+00	32768.000	3.364	32.880
1800/2000	0.77 it/sec	6.38E-07	2.476E+00	32768.000	3.446	35.394

Table 2. Training progress table for smaller dataset, Epoch 3

Data	Time	LR	Grad Norm	Grad Scale	Cross-Entropy (ce)	PPL
200/2000	0.83 it/sec	2.28E-05	2.106E+00	262144.000	4.302	83.217
400/2000	0.83 it/sec	1.86E-05	INF	207093.760	4.287	81.442
600/2000	0.83 it/sec	1.47E-05	2.024E+00	131072.000	4.267	79.955
800/2000	0.83 it/sec	1.12E-05	2.012E+00	131072.000	4.355	84.485
1000/2000	0.83 it/sec	8.10E-06	1.969E+00	131072.000	4.244	77.961
1200/2000	0.83 it/sec	5.48E-06	1.977E+00	131072.000	4.209	76.317
1400/2000	0.83 it/sec	3.35E-06	1.952E+00	131072.000	4.197	76.176
1600/2000	0.83 it/sec	1.73E-06	1.962E+00	131072.000	4.224	75.668
1800/2000	0.83 it/sec	6.38E-07	1.958E+00	131072.000	4.306	82.588

Table 4. Training progress table for larger dataset, Epoch 9

The small run was trained in 3 epochs and the full run in 9 epochs splitted into 3 runs. Unsurprisingly, the small run reported lower CE and PPL than the full run due to overfitting. According to ear test, the small model captured the style and even vocals of training data quite well but unable to generate anything else. In contrast, the full model seems to be undertrained, failing to generate passable audios with test labels provided. We think the full model is in the stage of early training

LIMITATION AND FUTURE IMPROVEMENT

Our model was trained with limited data and epochs, and the audio quality is not the best, therefore the poor results.

Hopefully we recognized the issues and here we propose some potential improvement:

- Better GPUs that allow faster and longer training
- More music data of higher quality, perhaps separated into components by instruments and with vocals removed
- More professional and detailed labels

CONCLUSIONS

In this project's we applied transfer learning to finetune a pretrained model MusicGen. With the curated J-pop dataset, we implement a training pipeline, but challenges in using LoRA and resource constraints led to a unsatisfying performance in genre-specific music generation. Nonetheless, we are on the right path and our implementation will facilitate future exploration.

REFERENCES

[1] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation, 2023. 1

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 1

CONTACT

Dihong Huang – dhuang329@gatech.edu
Akaash Dash – adash37@gatech.edu
Saikanam Siam – siam@gatech.edu
Professor Danfei Xu – danfei@gatech.edu