دانشکده مهندسی کامپیوتر

درس مبانی پردازش زبان و گفتار- شعبه ۰۱ – نیم‌سال دوم ۰۴-۱۴۰۳

استاد محترم: دکتر برادران

تاریخ تدوین : ۱۴۰۴/۰۴/۱۰

تولید موسیقی مبتنی بر پردازش زبان طبیعی

محمد امین کیانی ۴۰۰۳۶۱۳۰۵۲

فهرست مطالب

• هدف تحقیق:

تولید موسیقی از متن، مسئله‌ای چندرشته‌ای و نوین در مرزهای بین پردازش زبان طبیعی و یادگیری عمیق است که امکان تعامل خلاقانه انسان و ماشین را فراهم می‌سازد.

• اهمیت موضوع:

– رشد چشمگیر کاربردهای هوش مصنوعی در هنرهای خلاقانه
– نیاز به سیستم‌هایی برای تولید موسیقی بر اساس احساسات یا سناریوهای توصیفی
– تسهیل ساخت موسیقی برای فیلم، بازی، متاورس و تولیدات بدون نیاز به آهنگ‌سازی دستی

• دلایل انتخاب موضوع:

– نبود راهکارهای دقیق برای ارزیابی موسیقی‌های تولیدشده توسط مدل‌های متنی
– چالش‌های ترکیب درک زبانی با ساختار زمانی و صوتی موسیقی
– رشد سریع مدل‌های ژنراتیو مانند MusicGen، MusicLM، JEN-1 و نیاز به مقایسه مؤثر آن‌ها

رویکردهای ابتدایی:

سیستم‌های اولیه مانند TransProse (2014) با تکیه بر قواعد معنایی ساده و واژگان احساسی، نت‌های موسیقی را مستقیماً از تحلیل احساسی متن استخراج می‌کردند.

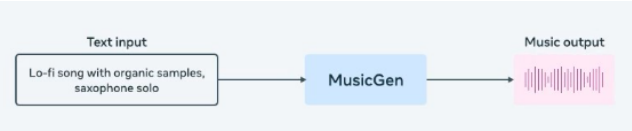
مدل‌های کلاسیک یادگیری ماشین:

استفاده از مدل‌های RNN و LSTM برای پردازش توالی‌های متنی و تولید نت، با کنترل ویژگی‌هایی مثل شدت و زمان‌بندی صداها.

محدودیت‌ها:

- وابستگی به قواعد دست‌نویس
- کیفیت پایین موسیقی تولیدی
- عدم تنوع در خروجی
- عدم توانایی در فهم عمیق مفاهیم متن

- با پیشرفت یادگیری عمیق، نسل جدیدی از مدل‌ها با معماری‌های پیچیده ظاهر شد که می‌توانستند روابط پنهان بین زبان و موسیقی را به‌صورت داده‌محور استخراج کنند.



مدل‌های شاخص:

- **Riffusion (2022) :** استفاده از مدل دیفیوژن برای تولید طیف‌نگاره‌ها (Spectrogram) از متن.
- **MusicLM (Google, 2023) :** مبتنی بر مدل دو مرحله‌ای رمزگذار–رمزگشا با کدک صوتی و ترنسفورمر.
- **MusicGen (Meta, 2023) :** تولید مستقیم waveform با مدل ترنسفورمری آموزش‌دیده روی داده‌های موسیقی–متن.
- **JEN-1 (2024) :** مدل چندوجهی با توانایی تولید صدای خواننده با کلمات مشخص‌شده.

- این مدل‌ها با استفاده از داده‌های عظیم، از ویژگی‌های معنایی و ساختاری متن، موسیقی‌هایی با کیفیت واقعی تولید می‌کنند.

• ارزیابی کیفیت موسیقی تولیدشده، فقط با گوش دادن کافی نیست بلکه به معیارهای کمی و قابل‌مقایسه نیاز داریم.

• استفاده ترکیبی از این معیارها، تصویری جامع از وفاداری معنایی، کیفیت شنیداری و شباهت آماری می‌دهد.

معیارهای اصلی:

1. **MOS (Mean Opinion Score)** : میانگین نمره شنوندگان (۵ نمره)، معیاری سنتی ولی ذهنی.

2. **FAD (Fréchet Audio Distance)** : فاصله توزیع ویژگی‌های صدای تولیدی با صدای واقعی که مشابه FID در تصویر است.

3. **CLAP Score (Contrastive Language-Audio Pretraining)** : شباهت معنایی بین متن ورودی و موسیقی خروجی با مدل CLAP.

4. **KAD (Kernel Audio Distance)** : معیار جدید و سبک‌وزن برای اندازه‌گیری اختلاف برداری صوت‌ها.

ساختار کلی MusicGen

- **نوع مدل:** ترنسفورمر
- **ورودی:** توصیف متنی آهنگ مورد نظر
- **خروجی:** دنباله‌ای از بردارهای صوتی که پس از پردازش تبدیل به موج صوتی WAV می‌شوند.

اجزای اصلی:

1.متن‌پرداز (Text Encoder) :

– برای تبدیل متن به Embedding معنایی.
– این بردار معنایی راهنمای تولید موسیقی است.

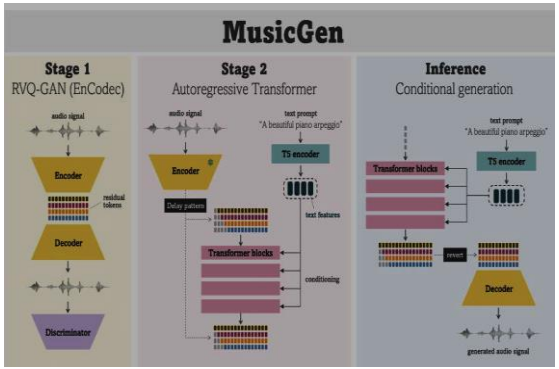
2.کدک صوتی (Audio Codec) :

– استفاده از EnCodec (مدل کدکننده صوتی از Meta) برای فشرده‌سازی و رمزگشایی صدا.
– صوت نهایی از دنباله‌ی کدهای توصیف‌شده بازسازی می‌شود.

3.ترنسفورمر مولد (Music Transformer Decoder) :

– دنباله‌ی کدهای صوتی را از متن رمزگذاری‌شده تولید می‌کند.
– از مکانیزم attention برای هماهنگی بین متن و ساختار زمانی صدا بهره می‌برد.

نسخه	اندازه پارامتر	کیفیت
small	~300M	سریع و سبک
medium	~1.5B	بالانس دقت و سرعت
large	~3.3B	کیفیت بالا
melody	+ ورودی ملودی	قابلیت کنترل دقیق‌تر




```
# نصب کتابخانه‌های لازم
!pip install transformers datasets scipy

from transformers import AutoProcessor, MusicgenForConditionalGeneration
from IPython.display import Audio

# بارگذاری مدل و پردازشگر MusicGen (نسخه کوچک)
processor = AutoProcessor.from_pretrained("facebook/musicgen-small")
model = MusicgenForConditionalGeneration.from_pretrained("facebook/musicgen-small")

# تعریف پرامپت متنی (می‌تواند فارسی یا انگلیسی باشد)
text_prompt = "A calm Persian classical music with santur and a slow rhythm"

# آماده‌سازی ورودی مدل
inputs = processor(text=[text_prompt], return_tensors="pt")

# تولید توکن‌های صوتی به صورت خودبازگشتی
audio_tokens = model.generate(**inputs, do_sample=True, guidance_scale=3.0,
max_new_tokens=256)

# به موج صوتی و پخش آن (تبدیل خروجی مدل (tokens
audio_array = audio_tokens[0].numpy()  # استخراج آرایه‌ی صوتی از خروجی مدل
sampling_rate = model.config.audio_encoder.sampling_rate  # نرخ نمونه‌برداری (مثلاً 32000 هرتز)
Audio(audio_array, rate=sampling_rate)
```

❗ چالش‌های مفهومی:

- **ابهام در زبان طبیعی:**
کلمات می‌توانند بار احساسی یا سبک موسیقایی مبهم یا چندپهلو داشته باشند.

- **نگاشت پیچیده متن به ویژگی‌های صوتی:**
احساسات یا سبک‌ها در متن الزاماً معادل دقیق و مستقیمی در ویژگی‌های آکوستیکی ندارند.

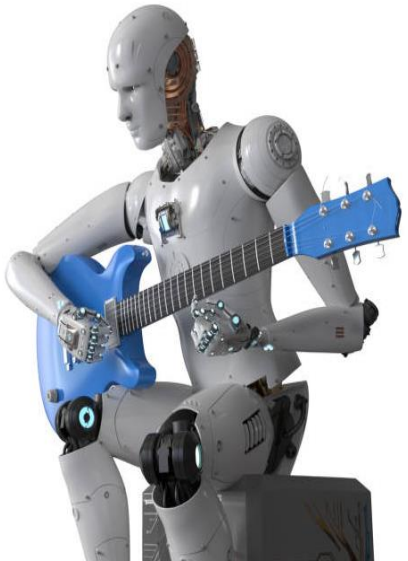
- **نبود معیار ارزیابی انسانی:**
بیشتر معیارها آماری و غیرشنیداری‌اند. ارزیابی واقعی کیفیت موسیقی تولیدی نیازمند گوش انسان است.

⚙ چالش‌های پیاده‌سازی:

- **مصرف منابع بالا:**
مدل‌های حجیمی مانند MusicLM یا CLAP نیاز به VRAM و RAM بالا دارند. اجرا روی سیستم‌های ضعیف یا Google Colab بسیار دشوار است.

- **محدودیت در دیتاست‌ها:**
منابع داده‌ای موسیقی همراه با توضیح متنی کم هستند یا حق کپی‌رایت دارند.

- **سازگاری نمونه‌برداری (Sampling Rate) :**
مدل‌های مختلف (تولید و ارزیابی) ممکن است نرخ‌های متفاوتی داشته باشند.



📌 جمع‌بندی:

• رویکردهای پردازش زبان طبیعی برای تولید موسیقی پیشرفت زیادی کرده‌اند؛ از روش‌های ساده مبتنی بر قواعد تا مدل‌های پیچیده ژنراتور مانند MusicGen و MusicLM.

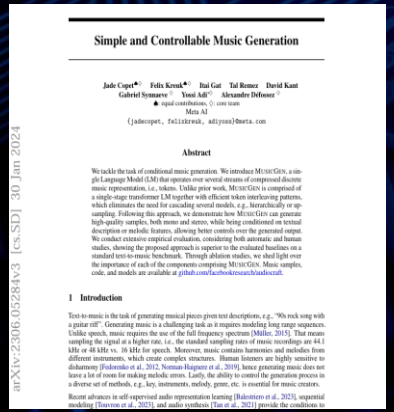
• ارزیابی موسیقی تولیدی نیازمند هم معیارهای آماری و هم شنیداری انسانی است و پیاده‌سازی موفق نیازمند درک دقیق از ارتباط میان زبان، احساس و ساختار موسیقی است.

💡 پیشنهادات آینده:

1. ترکیب ارزیابی انسانی و ماشینی برای سنجش بهتر خروجی‌ها.
2. ساخت دیتاست‌های چندزبانه و سبک‌محور برای آموزش مدل‌های تطبیق‌پذیرتر.
3. بهینه‌سازی مدل‌ها برای محیط‌های کم‌منبع.
4. افزایش کنترل‌پذیری خروجی موسیقی با افزودن پارامترهای سبک، ریتم، ساز و ... در پرامپت.
5. استفاده از موسیقی‌درمانی یا آموزش موسیقی به‌عنوان کاربردهای عملی مدل‌های متن به موسیقی.
6. ...؟

منابع

1. Yoonjin Chung, Pilsun Eu, et 21 Feb 2025: “KAD: No More FAD!”
2. copet, J., et al. (2023). “Simple and Controllable Music Generation.” (MusicGen model card).
3. https://huggingface.co/facebook/musicgen-small
.4 ...



Simple and Controllable Music Generation

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, Alexandre Défossez

Meta AI

Abstract

We tackle the task of conditional music generation. We introduce MusicGen, a single Language Model (LM) that operates over several streams of compressed discrete music representation, i.e., tokens. Unlike prior work, MusicGen is comprised of a single-stage transformer LM together with efficient token interleaving patterns, which eliminates the need for cascading several models, e.g., hierarchically or upsampling. Following this approach, we demonstrate how MusicGen can generate high-quality samples, both mono and stereo, while being conditioned on textual description or melodic features, allowing better controls over the generated output. We conduct extensive empirical evaluation, considering both automatic and human studies, showing the proposed approach is superior to the evaluated baselines on a standard text-to-music benchmark. Through ablation studies, we shed light over the importance of each of the components comprising MusicGen. Music samples, code, and models are available at github.com/facebookresearch/audiocraft.

1 Introduction

Text-to-music is the task of generating musical pieces given text descriptions, e.g., “90s rock song with a guitar riff”. Generating music is a challenging task as it requires modeling long range sequences. Unlike speech, music requires the use of the full frequency-spectrum [Müller, 2015]. That means sampling the signal at a higher rate, i.e., the standard sampling rates of music recordings are 44.1 kHz or 48 kHz vs. 16 kHz for speech. Moreover, music contains harmonies and melodies from different instruments, which create complex structures. Human listeners are highly sensitive to disharmony [Fedorenko et al., 2012, Norman-Haignere et al., 2019], hence generating music does not leave a lot of room for making melodic errors. Lastly, the ability to control the generation process in a diverse set of methods, e.g., key, instruments, melody, genre, etc. is essential for music creators.

Recent advances in self-supervised audio representation learning [Baevski et al., 2020], sequential modeling [Tsauvan et al., 2021], and audio synthesis [Tan et al., 2021] provide the conditions to

arXiv:2306.05284v3 [cs.SD] 30 Jan 2024

با تشکر از توجه شما

Aminkianiworkeng@gmail.com
https://github.com/M-Amin-Kiani/NLP-Proj