



fit@hcmus
VNUHCM - UNIVERSITY OF SCIENCE
FACULTY OF INFORMATION TECHNOLOGY

Khoa Luận Tốt Nghiệp

Hội đồng Kỹ thuật Phần mềm

Hệ thống AI hỗ trợ sáng tác nhạc

(AI-powered Support System for Music Composition)

Giảng viên hướng dẫn: TS. Trần Duy Hoàng
Giảng viên phản biện: TS. Lê Khánh Duy

Số thứ tự: 7
20120406 Phạm Quốc Vương
20120486 Ngô Phi Hùng

Nội dung

1. Giới thiệu đề tài
2. Các công trình liên quan
3. Phương pháp đề xuất
4. Kết quả thí nghiệm
5. Kết luận

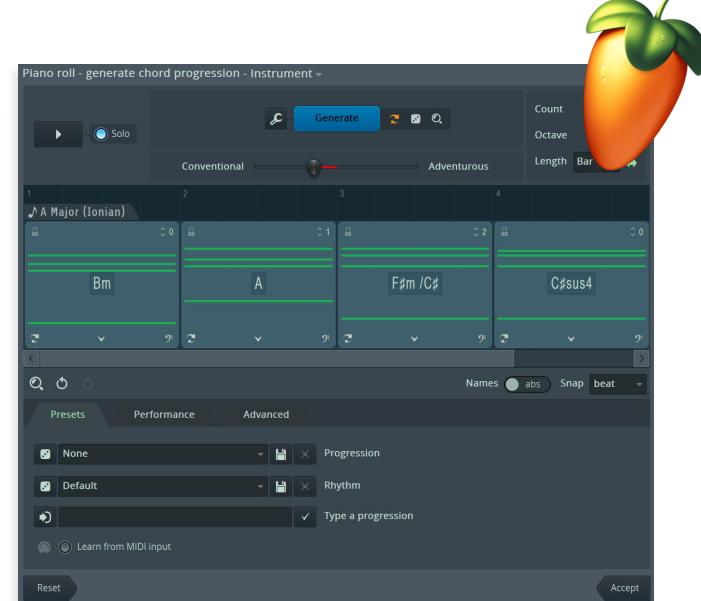
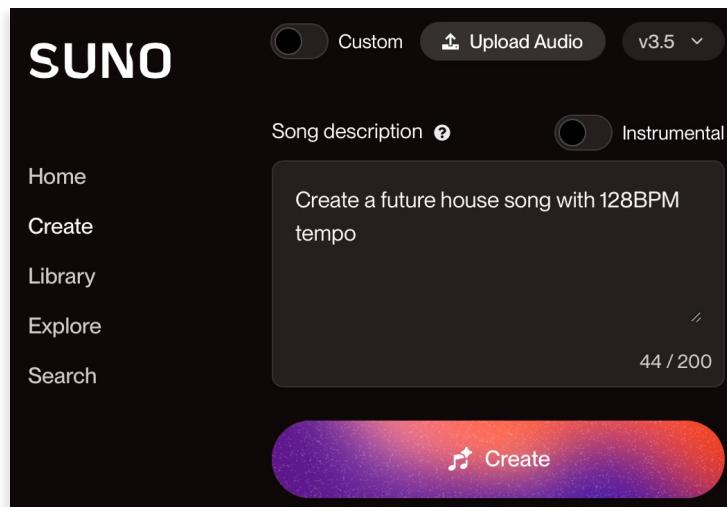
1. Giới thiệu đề tài

1.1. Đặt vấn đề

- Bài toán sinh nhạc: **text-to-audio** và **text-to-midi**.
- Bài toán trích xuất và chuyển đổi **mô tả cảm tính** thành **đặc tính kỹ thuật**.
- Bài toán chuyển **đặc tính kỹ thuật** thành **âm nhạc**.

1.2. Mục tiêu

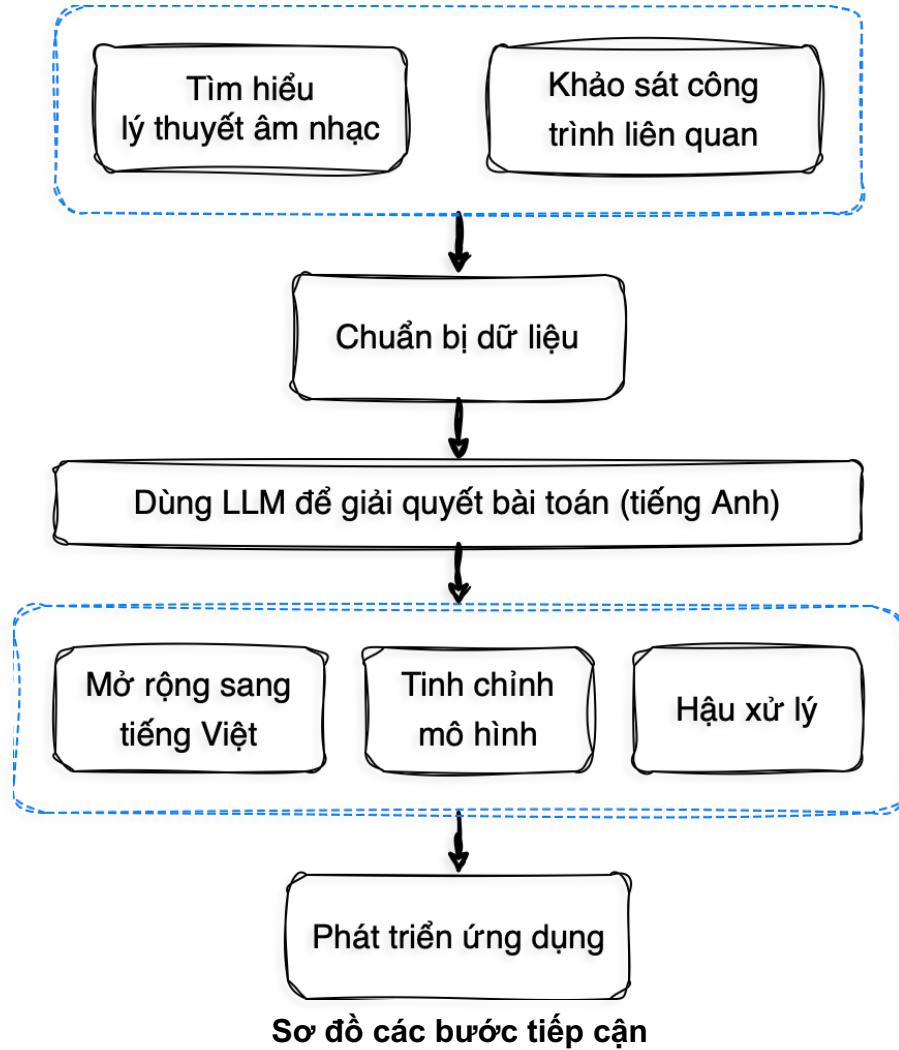
- Mô hình sinh nhạc dạng **text-to-midi** tiếng Anh
- Mô hình sinh nhạc dạng **text-to-midi** tiếng Việt
- Phần mềm sinh nhạc theo yêu cầu với đối tượng ưu tiên là người làm nhạc.



Một số giải pháp phần mềm³ cho bài toán sinh nhạc

1. Giới thiệu đề tài

1.3. Các bước tiếp cận



1.4. Đóng góp

- Mô hình sinh nhạc với đầu vào là mô tả tiếng Anh hoặc tiếng Việt.
- Phần mềm sinh nhạc theo yêu cầu.
- Giải pháp hậu xử lý dữ liệu âm nhạc (dạng MIDI).
- Kỹ thuật cào dữ liệu.
- Bài báo khoa học (chuẩn bị nộp).

2. Công trình liên quan

2.1. Các nghiên cứu

- **MuseCoco**^[1]: text-to-midi, văn bản mô tả mang tính kỹ thuật.
- **MusicLM**^[2]: text-to-audio, văn bản mô tả mang tính kỹ thuật lẫn ngũ cảnh.
- **MUGEN**^[3]: text-and-video-to-audio, văn bản mô tả ngũ cảnh của đoạn video về game.

2.2. Dữ liệu

- **MuseCoco**:
 - Các cặp “text – encoded text”;
 - Các bản nhạc ở định dạng MuseCoco.
- **ChatGPT**:
 - Hỗ trợ tạo các mẫu câu mô tả âm nhạc.
- **Hooktheory**: Dữ liệu âm nhạc có thể chuyển về dạng MIDI.

2. Công trình liên quan

2.3. Cơ sở lý thuyết

- Nhạc lý
- Kỹ thuật huấn luyện:
 - Masked Language Modelling
 - Casual Language Modelling
 - LoRA
 - Fairseq

2.4. Mô hình ngôn ngữ lớn

- Transformer
- BERT
- GPT2

3. Phương pháp đề xuất

3.1. Vấn đề chi tiết

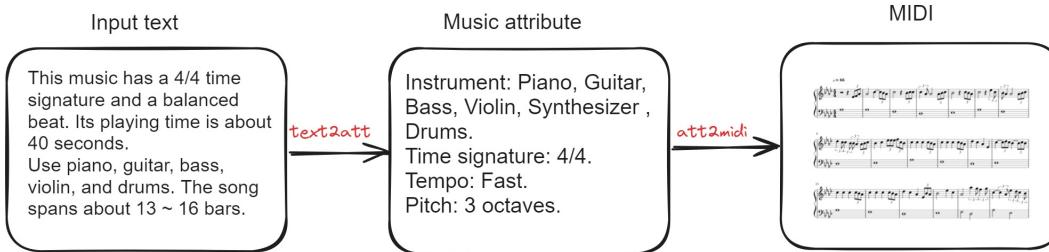
Mô hình sinh nhạc dạng **text-to-midi** phải đảm bảo:

- Có nguồn dữ liệu chất lượng cao để huấn luyện.
- Phù hợp với nguồn tài nguyên hiện có.
- Đầu ra đảm bảo tính đúng và chất lượng.

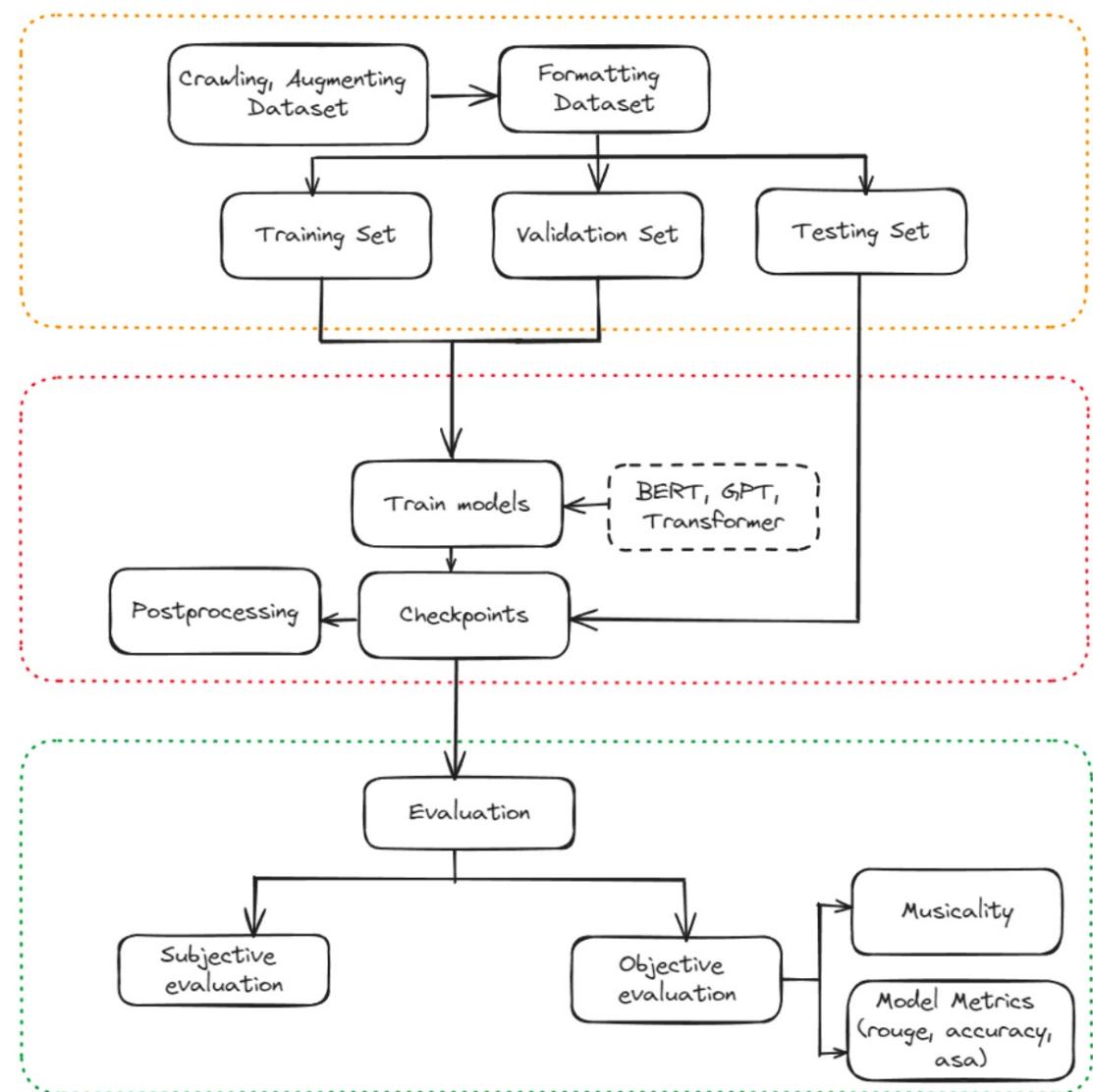
3. Phương pháp đề xuất

3.2. Tổng quan về phương pháp

- Kiến trúc tổng quát:
 - ❑ Mô hình **text-to-attribute**;
 - ❑ Mô hình **attribute-to-music**;
 - ❑ LoRA.
 - ❑ Mỗi mô hình mô hình áp dụng **quy trình thực nghiệm** như mô tả ở hình bên phải.



Sơ đồ kiến trúc tổng quát



Quy trình thực nghiệm

3. Phương pháp đề xuất

3.3. Chuẩn bị dữ liệu

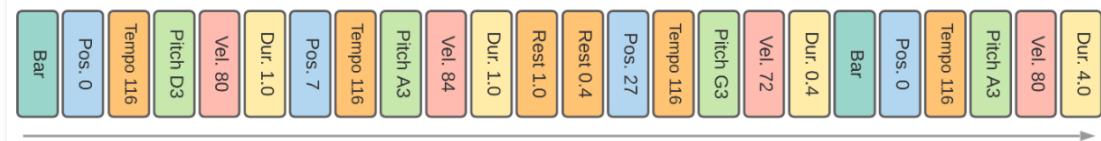
- Định dạng dữ liệu cho mô hình **text2att**:
 - ❑ Các **mẫu câu (template)** mô tả âm nhạc (tiếng Anh và tiếng Việt)
 - ❑ **Ví dụ:** “A track written in [KEY] key.
- Định dạng dữ liệu cho mô hình **text2midi**:
 - ❑ Các cặp “**command – music**” (dữ liệu dạng MuseCoco)
 - **Command:** Metadata của đoạn nhạc (tempo, nhịp, nhạc cụ, v.v).
 - **Music:** Đoạn nhạc được token hoá.
 - ❑ **Ví dụ:** Xem các hình bên dưới.

Ký hiệu	Mô tả
I1s2	Instrument
I4	Main Instrument
R3	Rhythm Intensity
B1s1	Bar
TS1s1	Time Signature
K1	Key
T1s1	Tempo
P4	Pitch Range
TM1	Time

Bảng 3.2: Mô tả các khoá trong command

I1s2 : (11, 4)
I4 : (11, False)
R3 : 1
B1s1 : (16, 3)
TS1s1 : (4, 4)
K1 : major
T1s1 : (114.03503592196344, 1)
P4 : 3
TM1 : (33.45463058047076, 2)

Minh họa một command



Minh họa phương pháp token hoá âm nhạc REMI

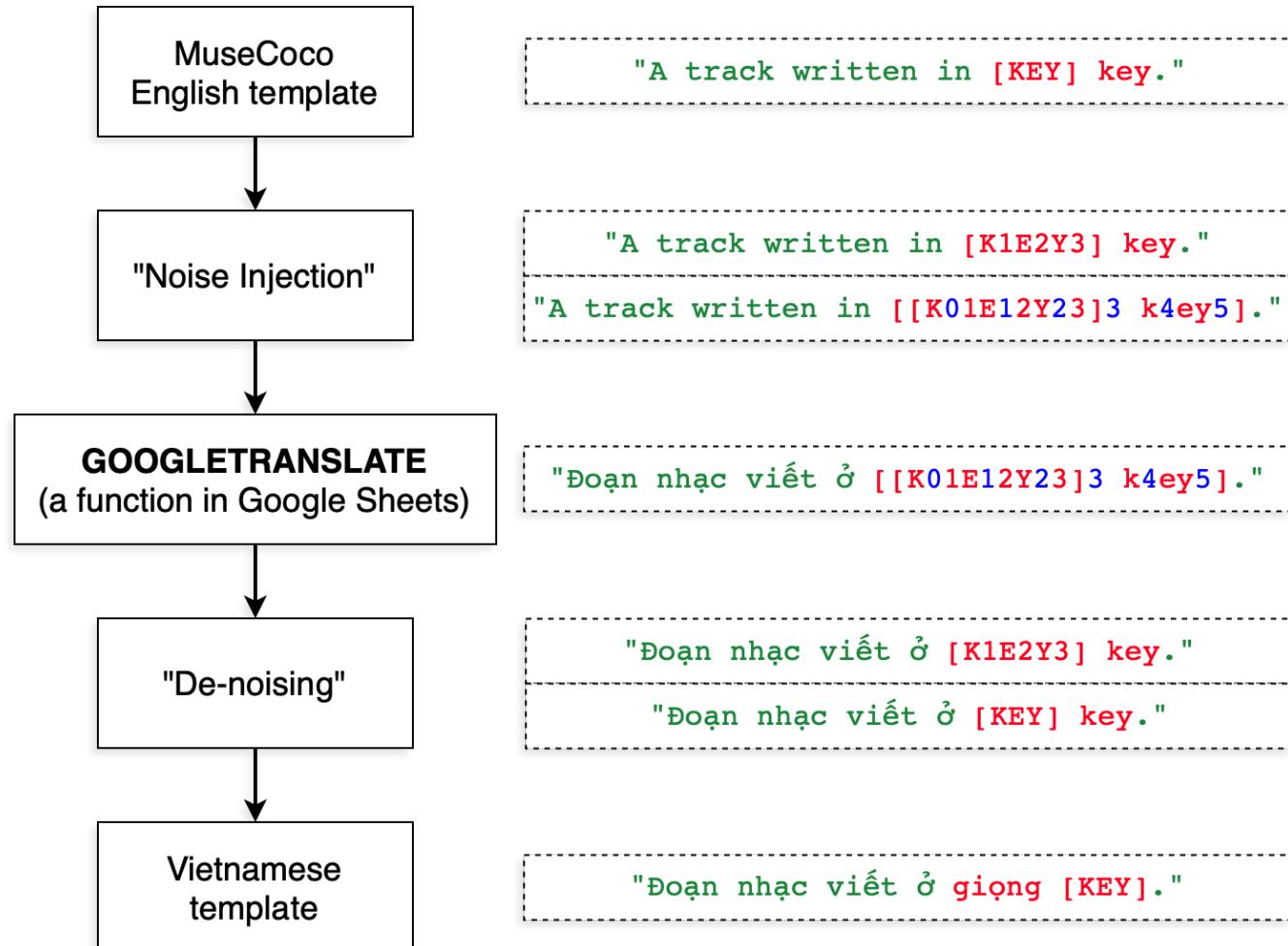
(Nền tảng cho phương pháp của MuseCoco)

3. Phương pháp đề xuất

3.3. Chuẩn bị dữ liệu

➤ Dữ liệu cho mô hình **text2att**:

- ❑ **Nguồn dữ liệu:** Các template mô tả âm nhạc tiếng Anh từ MuseCoco; Tạo thêm bằng **prompt engineering** với ChatGPT.
- ❑ **Vấn đề:** Dịch thuật số lượng lớn (để có thêm dữ liệu tiếng Việt); Dịch thuật ngoài ý muốn.



Quy trình dịch thuật

3. Phương pháp đề xuất

3.3. Chuẩn bị dữ liệu

- Dữ liệu cho mô hình **text2att**:

Kết quả chuẩn bị:

- Các nhãn sử dụng được liệt kê ở bảng bên phải
- Số mẫu dữ liệu tiếng Anh: 14900
- Số mẫu dữ liệu tiếng Việt: 14900

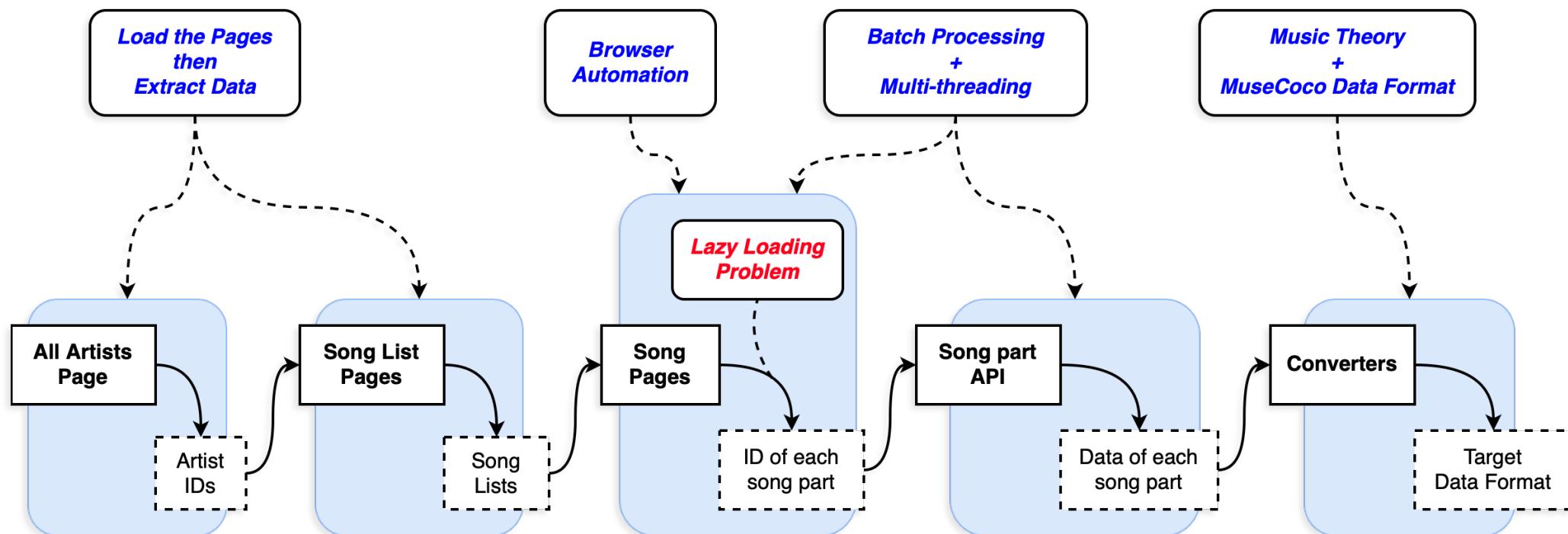
Tên nhãn	Giá trị
Instrument	28 instruments: piano, keyboard, percussion, organ, guitar, bass, violin, viola, cello, harp, strings, voice, trumpet, trombone, tuba, horn, brass, sax, oboe, bassoon, clarinet, piccolo, flute, pipe, synthesizer, ethnic instrument, sound effect, drum. Mỗi nhạc cụ: 0: Được chơi, 1: Không được chơi, 2: NA
Pitch	Range: 0-11: octaves, 12: NA.
Rhythm	0: danceable, 1: not danceable, 2: NA.
Danceability	
Bar	0: 1-4 bars, 1: 5-8 bars, 2: 9-12 bars, 3: 13-16 bars, 4: NA.
Time	0: 4/4, 1: 2/4, 2: 3/4, 3: 1/4, 4: 6/8, 5: 3/8, 6: các nhịp khác, 7: NA.
Signature	
Key	0: major, 1: minor, 2: NA.
Tempo	0: chậm (<=76 BPM), 1: trung bình (76-120 BPM), 2: nhanh (>=120 BPM), 3: NA.
Time	0: 0-15s, 1: 15-30s, 2: 30-45s, 3: 45-60s, 4: >60s, 5: NA.

**Tên và giá trị tương ứng của các nhãn trong mỗi câu template
từ bài báo MuseCoco (đã được rút gọn)**

3. Phương pháp đề xuất

3.3. Chuẩn bị dữ liệu

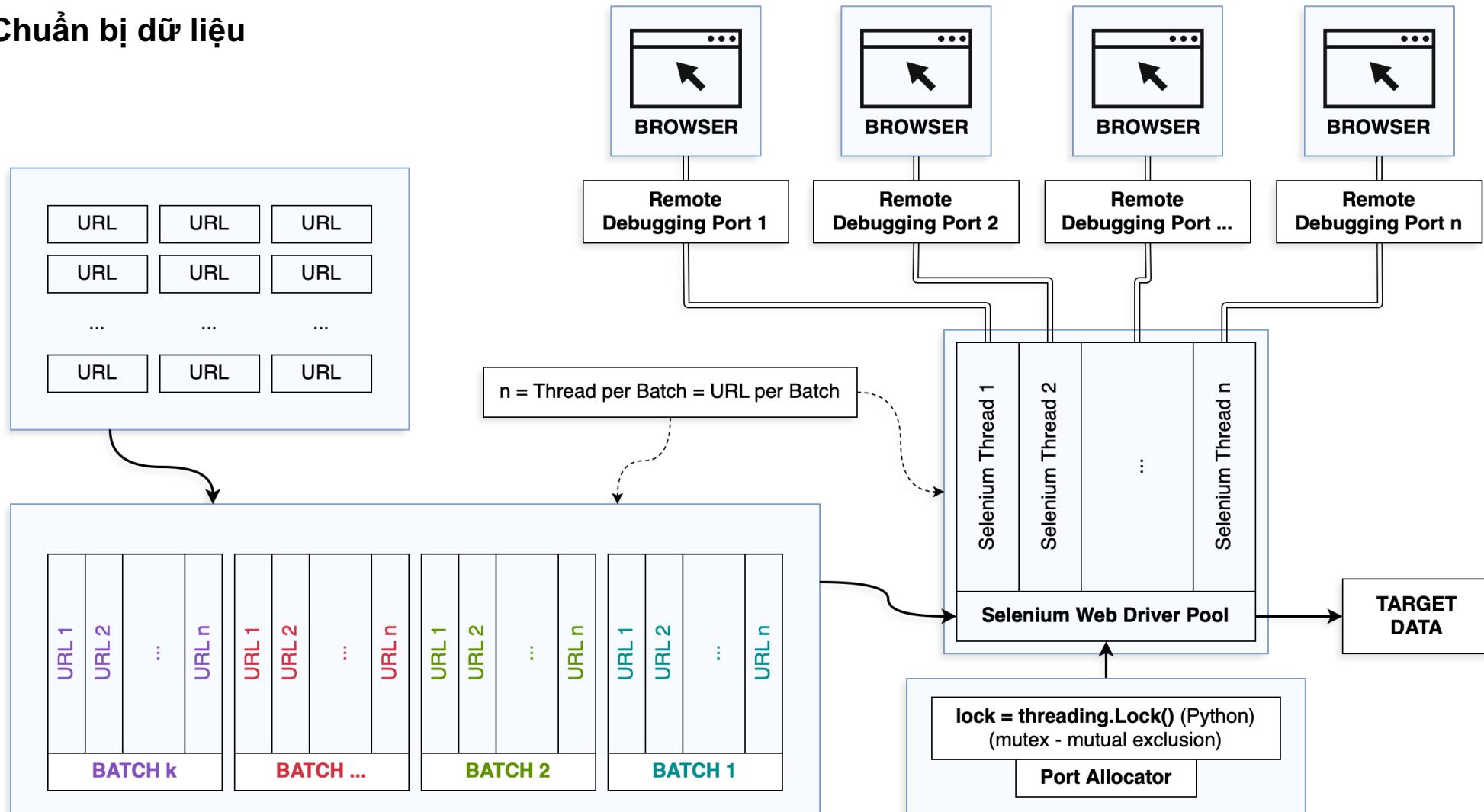
- Dữ liệu cho mô hình att2midi:
 - ❑ Nguồn dữ liệu: Dữ liệu âm nhạc của MuseCoco và Hooktheory.
 - ❑ Vấn đề: lazy loading; số lượng dữ liệu cần thu thập lớn.



Quy trình thu thập dữ liệu âm nhạc¹²

3. Phương pháp đề xuất

3.3. Chuẩn bị dữ liệu



Kiến trúc cào dữ liệu áp dụng 3 kỹ thuật: **Batch Processing, Multi-threading và Browser Automation**

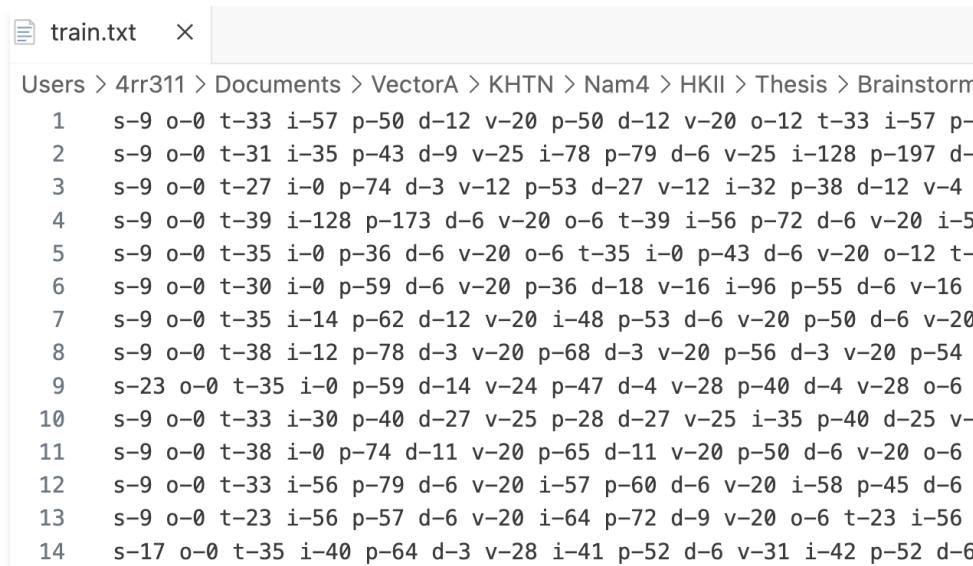
3. Phương pháp đề xuất

3.3. Chuẩn bị dữ liệu

- Dữ liệu cho mô hình **att2midi**:

- Kết quả chuẩn bị:**

- 300 cặp “**command – music**” công khai của MuseCoco.
 - 29000 cặp thu thập từ Hooktheory và xử lý.



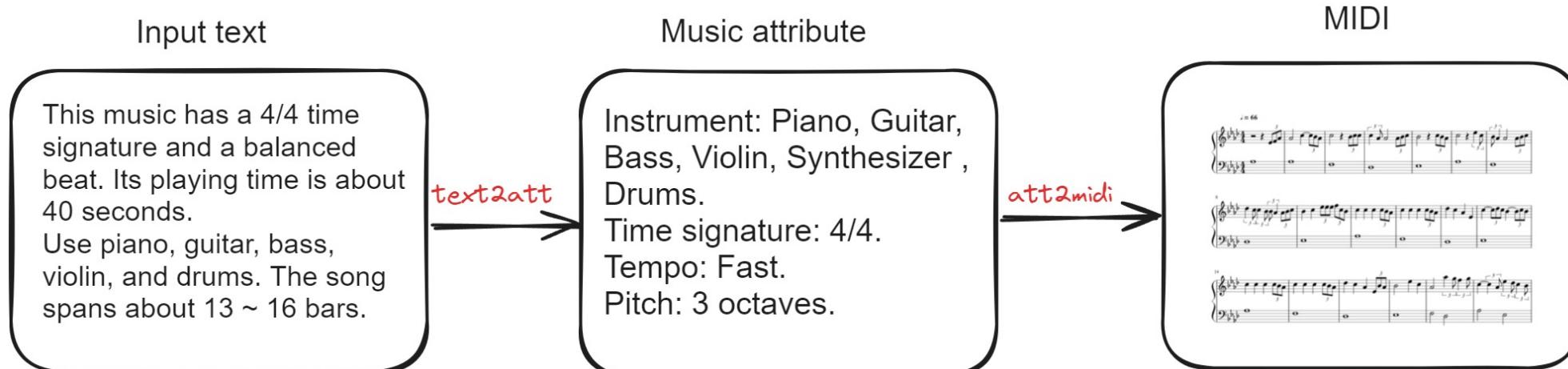
The screenshot shows a text editor window titled "train.txt". The file contains a list of 14 numbered lines, each representing a sequence of musical commands. The commands are separated by spaces and include letters such as s, o, t, p, d, v, i, and numbers ranging from 0 to 79. The file path is visible at the top of the editor window.

```
train.txt  X  
Users > 4rr311 > Documents > VectorA > KHTN > Nam4 > HKII > Thesis > Brainstorming  
1   s-9 o-0 t-33 i-57 p-50 d-12 v-20 p-50 d-12 v-20 o-12 t-33 i-57 p-5  
2   s-9 o-0 t-31 i-35 p-43 d-9 v-25 i-78 p-79 d-6 v-25 i-128 p-197 d-3  
3   s-9 o-0 t-27 i-0 p-74 d-3 v-12 p-53 d-27 v-12 i-32 p-38 d-12 v-4 i  
4   s-9 o-0 t-39 i-128 p-173 d-6 v-20 o-6 t-39 i-56 p-72 d-6 v-20 i-57  
5   s-9 o-0 t-35 i-0 p-36 d-6 v-20 o-6 t-35 i-0 p-43 d-6 v-20 o-12 t-3  
6   s-9 o-0 t-30 i-0 p-59 d-6 v-20 p-36 d-18 v-16 i-96 p-55 d-6 v-16 i  
7   s-9 o-0 t-35 i-14 p-62 d-12 v-20 i-48 p-53 d-6 v-20 p-50 d-6 v-20  
8   s-9 o-0 t-38 i-12 p-78 d-3 v-20 p-68 d-3 v-20 p-56 d-3 v-20 p-54 d  
9   s-23 o-0 t-35 i-0 p-59 d-14 v-24 p-47 d-4 v-28 p-40 d-4 v-28 o-6 t  
10  s-9 o-0 t-33 i-30 p-40 d-27 v-25 p-28 d-27 v-25 i-35 p-40 d-25 v-2  
11  s-9 o-0 t-38 i-0 p-74 d-11 v-20 p-65 d-11 v-20 p-50 d-6 v-20 o-6 t  
12  s-9 o-0 t-33 i-56 p-79 d-6 v-20 i-57 p-60 d-6 v-20 i-58 p-45 d-6 v  
13  s-9 o-0 t-23 i-56 p-57 d-6 v-20 i-64 p-72 d-9 v-20 o-6 t-23 i-56 p  
14  s-17 o-0 t-35 i-40 p-64 d-3 v-28 i-41 p-52 d-6 v-31 i-42 p-52 d-6
```

Dữ liệu âm nhạc sau khi encode

3. Phương pháp đề xuất

3.4. Kiến trúc mô hình



- Mô hình gồm 2 stage: text2att và att2midi.
- Lựa chọn kiến trúc trên phù hợp vì:
 - ❑ Khối lượng tài nguyên tính toán;
 - ❑ Khối lượng dữ liệu để huấn luyện;
 - ❑ Lượng từ vựng cần xử lý;
 - ❑ Tách biệt khâu xử lý văn bản và âm nhạc.

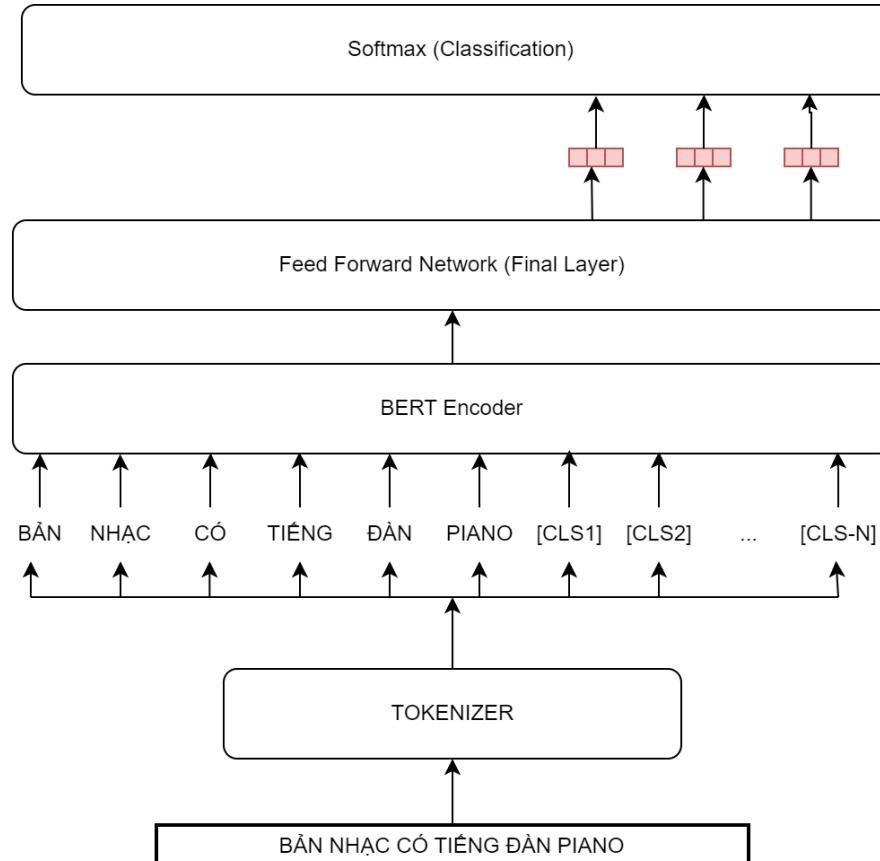
3. Phương pháp đề xuất

3.5. text2att

- Sử dụng bert-uncased-multilangage làm bộ checkpoint gốc.
- Thêm nhiều [CLS] đại diện cho mỗi nhãn để mô hình có thể học được bối cảnh trong câu.
 - Ví dụ: [CLS] BERT là một mô hình ngôn ngữ mạnh mẽ. [SEP].
 - Sau khi qua các lớp của BERT, nhận được một ma trận biểu diễn cho mỗi token trong chuỗi, bao gồm cả token [CLS]. Vector biểu diễn cho token [CLS] sẽ chứa thông tin tóm tắt của toàn bộ câu và được sử dụng cho tầng phân loại tiếp theo để đưa ra dự đoán.

3. Phương pháp đề xuất

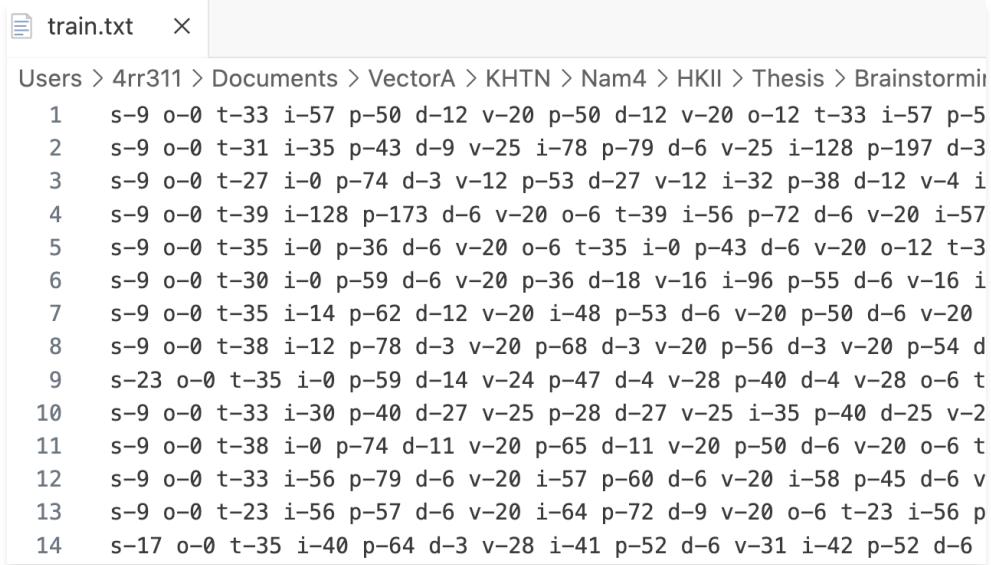
3.5. text2att



Mô tả flow text2att

3. Phương pháp đề xuất

3.6. Chuẩn bị bộ tokenizer cho mô hình sinh nhạc



The screenshot shows a text editor window titled "train.txt". The file path is displayed at the top: "Users > 4rr311 > Documents > VectorA > KHTN > Nam4 > HKII > Thesis > Brainstorming". The content of the file is a sequence of 14 numbered lines (1 to 14) representing tokens. Each line consists of a number followed by a space and a string of lowercase letters separated by hyphens. The strings represent sequences of phonemes (p), vowels (o), and consonants (t, d, s, v, i, n, l, r, m, f, z, ñ, ññ, ñññ, ññññ).

```
1 s-9 o-0 t-33 i-57 p-50 d-12 v-20 p-50 d-12 v-20 o-12 t-33 i-57 p-5  
2 s-9 o-0 t-31 i-35 p-43 d-9 v-25 i-78 p-79 d-6 v-25 i-128 p-197 d-3  
3 s-9 o-0 t-27 i-0 p-74 d-3 v-12 p-53 d-27 v-12 i-32 p-38 d-12 v-4 i  
4 s-9 o-0 t-39 i-128 p-173 d-6 v-20 o-6 t-39 i-56 p-72 d-6 v-20 i-57  
5 s-9 o-0 t-35 i-0 p-36 d-6 v-20 o-6 t-35 i-0 p-43 d-6 v-20 o-12 t-3  
6 s-9 o-0 t-30 i-0 p-59 d-6 v-20 p-36 d-18 v-16 i-96 p-55 d-6 v-16 i  
7 s-9 o-0 t-35 i-14 p-62 d-12 v-20 i-48 p-53 d-6 v-20 p-50 d-6 v-20  
8 s-9 o-0 t-38 i-12 p-78 d-3 v-20 p-68 d-3 v-20 p-56 d-3 v-20 p-54 d  
9 s-23 o-0 t-35 i-0 p-59 d-14 v-24 p-47 d-4 v-28 p-40 d-4 v-28 o-6 t  
10 s-9 o-0 t-33 i-30 p-40 d-27 v-25 p-28 d-27 v-25 i-35 p-40 d-25 v-2  
11 s-9 o-0 t-38 i-0 p-74 d-11 v-20 p-65 d-11 v-20 p-50 d-6 v-20 o-6 t  
12 s-9 o-0 t-33 i-56 p-79 d-6 v-20 i-57 p-60 d-6 v-20 i-58 p-45 d-6 v  
13 s-9 o-0 t-23 i-56 p-57 d-6 v-20 i-64 p-72 d-9 v-20 o-6 t-23 i-56 p  
14 s-17 o-0 t-35 i-40 p-64 d-3 v-28 i-41 p-52 d-6 v-31 i-42 p-52 d-6
```

Ví dụ về dữ liệu âm nhạc

➤ Bộ từ vựng 1253 từ bao gồm:

- Gồm từ vựng âm nhạc (REMI)
- Từ vựng command (metadata)
- Từ vựng khác (CLS, UNK,...)

3. Phương pháp đề xuất

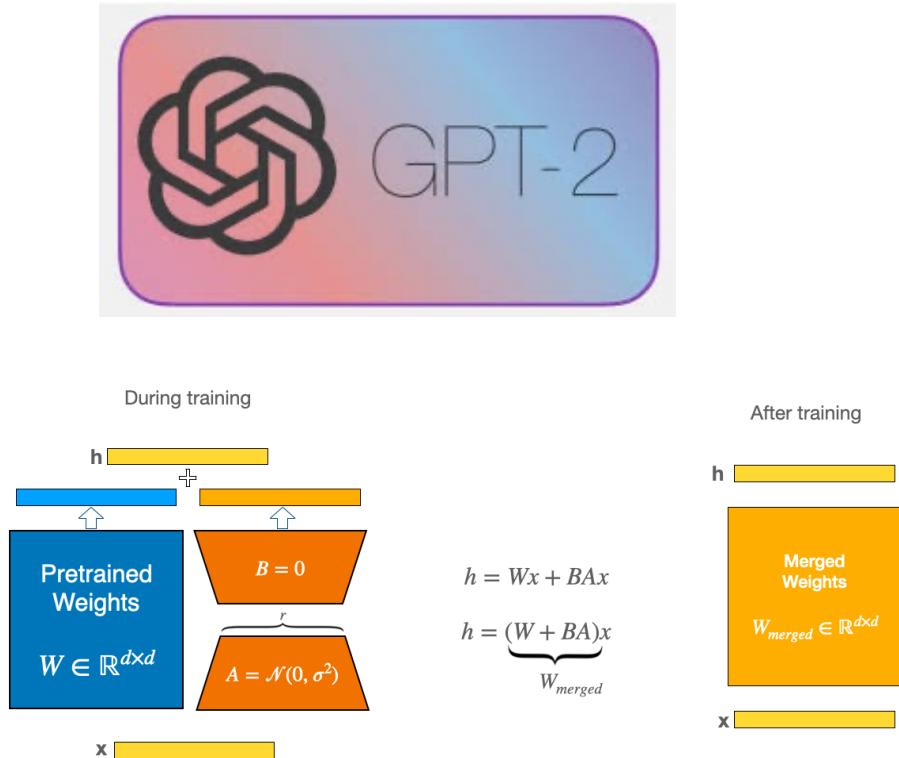
3.6. Chuẩn bị bộ tokenizer cho mô hình sinh nhạc

Siêu tham số	Giá trị	Ý nghĩa
epoch	100	Số lượng epoch xác định số lần mô hình sẽ được huấn luyện trên toàn bộ dữ liệu.
batch size	32	Kích thước của mỗi batch dữ liệu đưa vào mô hình trong mỗi bước huấn luyện.
optimizer	AdamW	Bộ tối ưu hóa sử dụng để cập nhật các tham số của mô hình trong quá trình huấn luyện.
learning rate	2e-5	Tốc độ học xác định mức độ điều chỉnh các trọng số của mô hình sau mỗi bước huấn luyện.
max sequence length	128	Chiều dài tối đa của chuỗi đầu vào, các chuỗi dài hơn sẽ bị cắt ngắn.
warmup steps	500	Số bước đầu tiên trong đó tốc độ học tăng dần đến giá trị tối đa đã định.
dropout rate	0.1	Tỷ lệ dropout sử dụng trong quá trình huấn luyện để tránh overfitting.
gradient accumulation steps	2	Số bước gradient accumulation trước khi cập nhật trọng số mô hình.
weight decay	0.01	Hệ số điều chỉnh tỷ lệ suy giảm trọng số, giúp tránh overfitting.

Bảng 3.3: Các hyperparameter được sử dụng trong quá trình huấn luyện mô hình BERT

3. Phương pháp đề xuất

3.6. att2midi - GPT2 LoRA



- 200 Triệu tham số.
- Lựa chọn các tham số tuyến tính để huấn luyện.
- 3.5 Triệu tham số được huấn luyện.

3. Phương pháp đề xuất

3.6. att2midi - GPT2 LoRA



Mô tả dạng huấn luyện command-music, att2midi

3. Phương pháp đề xuất

3.6. att2midi - GPT2 LoRA

- Thông tin tham số cấu hình:

Tham số	Giá trị	Ý nghĩa
n_layer	20	Số lượng lớp Transformer
n_head	16	Số lượng head của multi-head attention
n_emb	1024	Kích thước của vector embedding
vocab_size	1253	Kích thước từ vựng, tổng số lượng token
n_positions	2048	Số lượng vị trí tối đa trong một chuỗi

Bảng 3.5: Bảng tham số cấu hình cho mô hình GPT-2

Tham số	Giá trị	Ý nghĩa
r	16	Hệ số giảm chiều của các lớp được điều chỉnh
lora_alpha	12	Hệ số mở rộng (scaling factor) cho Lora
lora_dropout	0.1	Xác suất dropout áp dụng cho các lớp Lora
target_modules	[c_proj, c_attn, wte, lm_head]	Danh sách các mô-đun đích để áp dụng LoRA
task_type	CAUSAL_LM	Loại tác vụ mà mô hình được áp dụng

Bảng 3.6: Bảng tham số cấu hình cho LoraConfig.

3. Phương pháp đề xuất

3.7. att2midi - Fairseq

- 200 Triệu tham số.
- Cùng một số hyperparameter với mô hình GPT2 - LoRA.

facebookresearch/
fairseq

Facebook AI Research Sequence-to-Sequence
Toolkit written in Python.

310
Contributors

3k
Used by

30k
Stars 6k
Forks



Công cụ fairseq do Facebook phát triển

Tham số	Giá trị	Mô tả
n_layer	20	Số lớp trong Transformer.
n_head	16	Số lượng head attention.
n_emb	1024	Kích thước vector embeddings.
FFN size	2048	Kích thước của Feed Forward Network.
dropout ratio	0.1	Tỷ lệ dropout để tránh overfitting.
optimizer	AdamW	Optimizer sử dụng để huấn luyện mô hình.
β_1	0.9	Tham số β_1 của Adam optimizer.
β_2	0.98	Tham số β_2 của Adam optimizer.
ϵ	10^{-9}	Tham số ϵ của Adam optimizer.
learning rate	2×10^{-4}	Tốc độ học của mô hình.
warmup steps	2000	Số bước đầu tiên để tốc độ học tăng dần đến giá trị tối đa đã định.

**Các tham số của mô hình
Casual Linear Transformer**

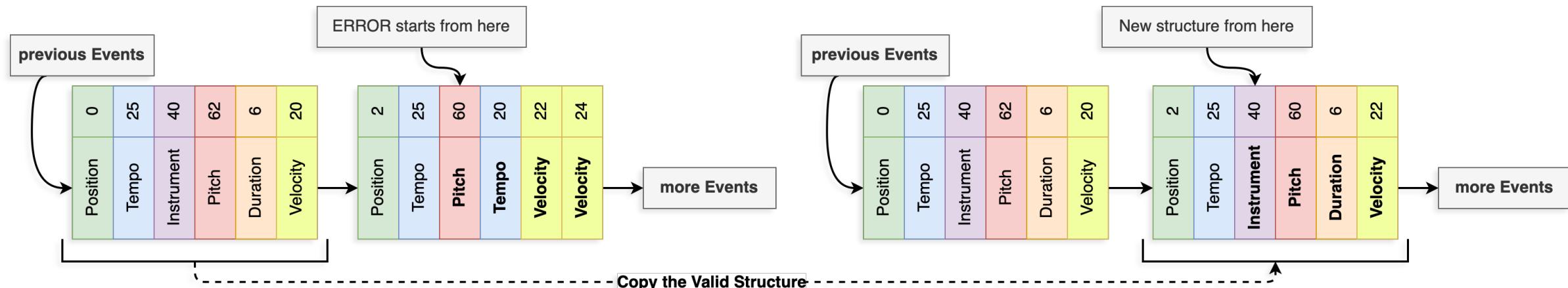
3. Phương pháp đề xuất

3.8. Hậu xử lý dữ liệu

- **Vấn đề:** Rủi ro dữ liệu sinh ra sai cấu trúc.
- **Giải pháp:**
 - ❑ Dữ liệu = time series = {Các sự kiện MIDI};
 - ❑ Tìm vị trí lỗi dựa trên đặc tả.
 - ❑ Khôi phục cấu trúc tại điểm lỗi;
 - ❑ Điền giá trị cho các thành phần cấu trúc.

Thuộc tính	Tên đầy đủ	Mô tả	Thuộc tính liền sau
s	Time Signature	Nhịp của bản nhạc	b, o.
o	Position	Thời điểm xuất hiện của sự kiện	t.
t	Tempo	Tốc độ bản nhạc	i.
i	Instrument	Nhạc cụ	p.
p	Pitch	Cao độ nốt nhạc	d.
d	Duration	Độ dài nốt nhạc	v.
v	Velocity	Tốc độ nhấn/Lực nhấn phím đàn	i, b, p, o.
b	Bar	Vạch nhịp	s.

Bảng 2.2: Đặc tả dữ liệu âm nhạc dạng văn bản của MuseCoco



Ví dụ minh họa quá trình hậu xử lý

3. Phương pháp đề xuất

3.9. Đánh giá mô hình

➤ Objective Evaluation:

- Micro: Tính toán các số liệu accuracy cho từng nhãn riêng lẻ, rồi lấy trung bình. Ưu tiên khi muốn đảm bảo mỗi nhãn đều được đánh giá công bằng, không bị các nhãn lớn lấn át.
- Macro: Tính toán các số liệu accuracy trên toàn bộ các dự đoán, coi tất cả nhãn như một tập hợp lớn. Ưu tiên khi quan tâm đến hiệu suất tổng thể.

$$accuracy = \frac{\text{số dự đoán đúng}}{\text{tổng số dự đoán}}$$

- Đánh giá dựa trên các đặc tính kỹ thuật.

➤ Subjective Evaluation:

- Đánh giá dựa trên cảm nhận.

4. Kết quả thí nghiệm

4.1. Kết quả huấn luyện

Instrument	Accuracy (ENG)	Accuracy (VIE)
accordion	0.94	0.93
brass	0.98	0.96
celesta	0.91	0.92
choir	0.95	0.97
guitar	0.99	0.93
harmonica	0.97	0.94
organ	0.90	0.91
piano	0.96	0.95
synth	0.92	0.94
viola	0.91	0.90
violin	0.93	0.92
voice	0.95	0.96

Bảng 4.1: Instrument Accuracy

Category	Accuracy (ENG)	Accuracy (VIE)
Rhythm Danceability	0.85	0.87
Bar	0.91	0.89
Time Signature	0.94	0.92
Key	0.88	0.90
Tempo	0.93	0.85
Pitch Range	0.90	0.94

Bảng 4.2: Categories Accuracy

Độ đo	LoRA GPT2	Fairseq Transformer seq2seq
ASA	0.64	0.57
Instrument	0.67	0.6
Pitch Range	0.68	0.52
Rhythm Danceability	0.96	0.89
Bar	0.51	0.42
Time Signature	0.37	0.43
Key	0.57	0.51
Tempo	0.69	0.61

Bảng 4.4: So sánh kết quả giữa hai mô hình.

4. Kết quả thí nghiệm

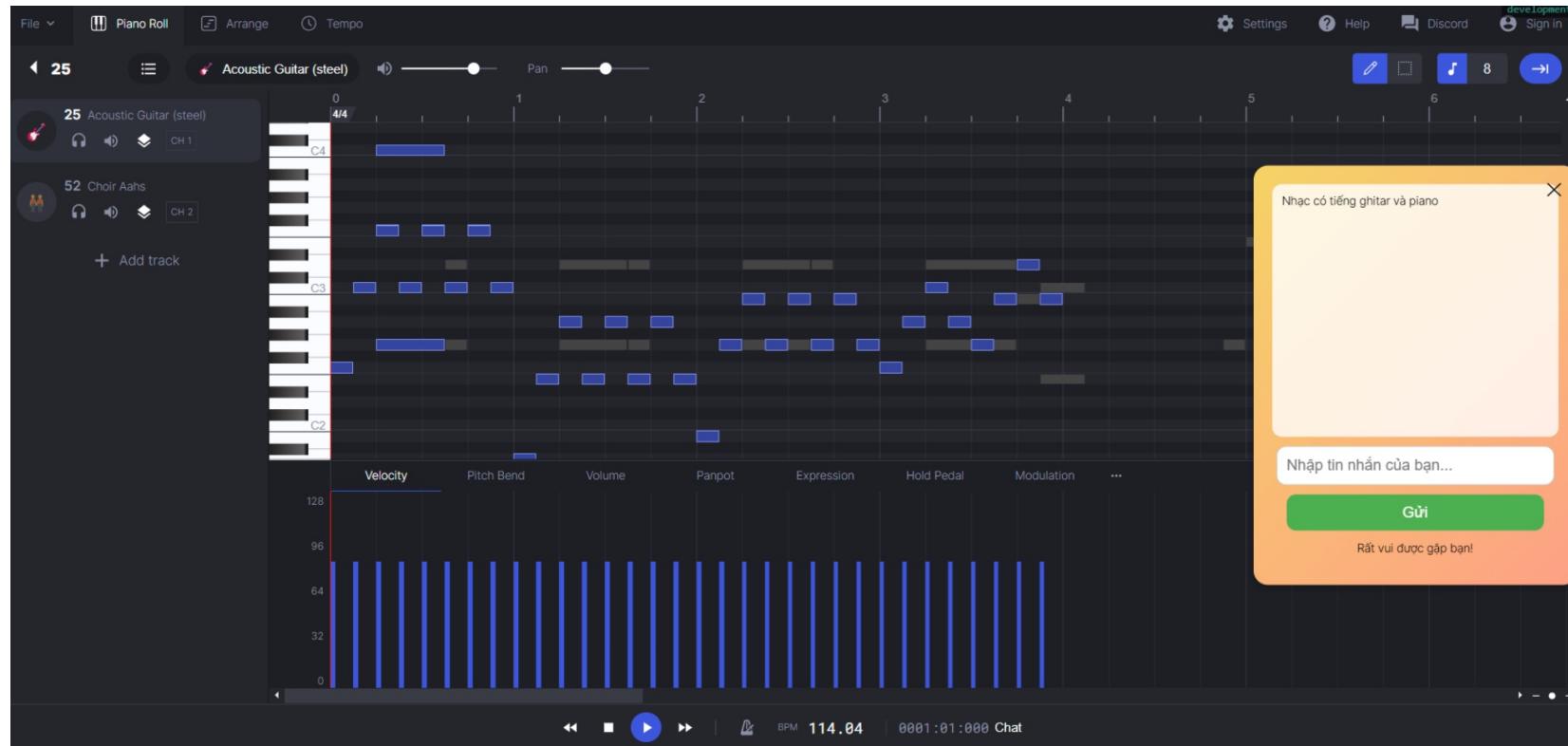
4.2. Nhận xét

Tiêu Chí	GPT2 LoRA	Fairseq
Subjective Evaluation	Nghe hay hơn, chính xác hơn Fairseq	Nghe bị giật, không mượt.
Kích thước	900MB, nhỏ hơn Fairseq	2.8GB, lớn hơn GPT2 LoRA
Thời gian inference	Nhanh	Chậm
Cài đặt môi trường	Đơn giản	Nhiều thư viện phức tạp.

Đánh giá hai mô hình dựa trên các tiêu chí khác nhau

4. Kết quả thí nghiệm

4.3. Phần mềm demo



Giao diện phần mềm demo

5. Kết luận

- Áp dụng được kiến trúc hai lớp.
- Làm giàu dữ liệu bằng GPT.
- Áp dụng LoRA tối ưu chi phí.
- Mang lại nhiều lợi ích cho người làm nhạc.

TÀI LIỆU THAM KHẢO

1. Lu, Peiling et al. MuseCoco: Generating Symbolic Music from Text. 2023. arXiv: 2306.00110 [cs.SD].
2. Agostinelli, Andrea et al. MusicLM: Generating Music From Text. 2023. arXiv: 2301.11325 [cs.SD].
3. Hayes, Thomas et al. MUGEN: A Playground for Video-Audio-Text Multimodal Understanding and GENeration. 2022. arXiv: 2204.08058 [cs.CV].

CẢM ƠN QUÝ THẦY CÔ ĐÃ LẮNG NGHE