DJ_LMM

Richard Goudelin, Pierre Pauchet

Summary

- 1. Context, Problem
- 2. Methods :
  - The MIDI protocol
  - Tokenizer BPE
  - Encoder-Decoder Models
- 3. Experiments
  - REMI-BPE
  - Training BART
- 4. Results and demos
  - What made it work
- Conclusion and perspectives

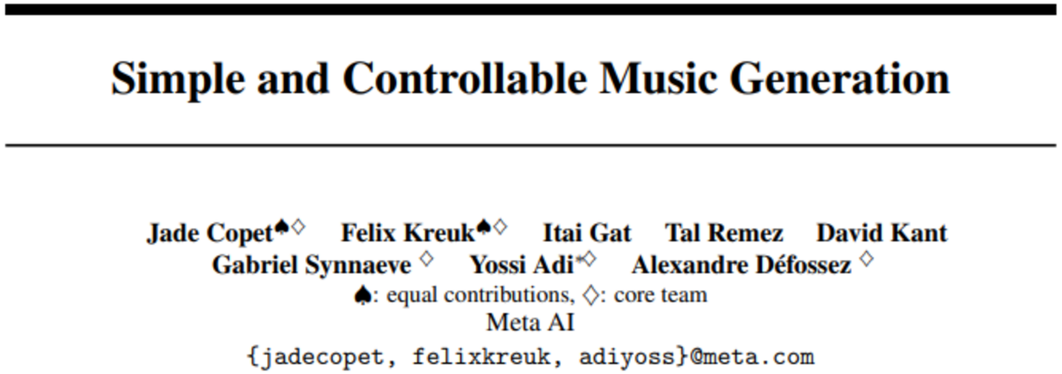
Fig. 1 : an image generated from stablediffusionweb.com. The text prompt was “An abstract representation of a DJ LLM.”

1. Context - Transitions between music

- Streaming platforms : fade out/in
- Demands manual care

- LLMs for Music : already explored
- Mutltiple conditionning schemes

- **Our Idea** : LLM for Conditional Music Generation
  - Focus on transitions
  - Single-channel
  - Lightweight



Simple and Controllable Music Generation

Jade Copet[♠◇] **Felix Kreuk**[♠◇] **Itai Gat** **Tal Remez** **David Kant**
Gabriel Synnaeve [◇] **Yossi Adi**[◇] **Alexandre Défossez** [◇]
♠: equal contributions, ◇: core team
Meta AI
{jadecopet, felixkreuk, adiyoss}@meta.com

Fig 2 : Title card of the paper “Simple and controlable music Generation from Copet et al. (24)

2. Methods - The MIDI Protocol

- Music formats : symbolic vs audio
- **MIDI Protocol** : comprised of Events over Time
  - Note On: a note is being played, specifies its pitch and velocity;
  - Note Off: a note is released, specifies the note (by its pitch) to stop and the velocity;
  - Time Signature Change, Tempo changes...
- Time resolution : ticks per quarter (tps)

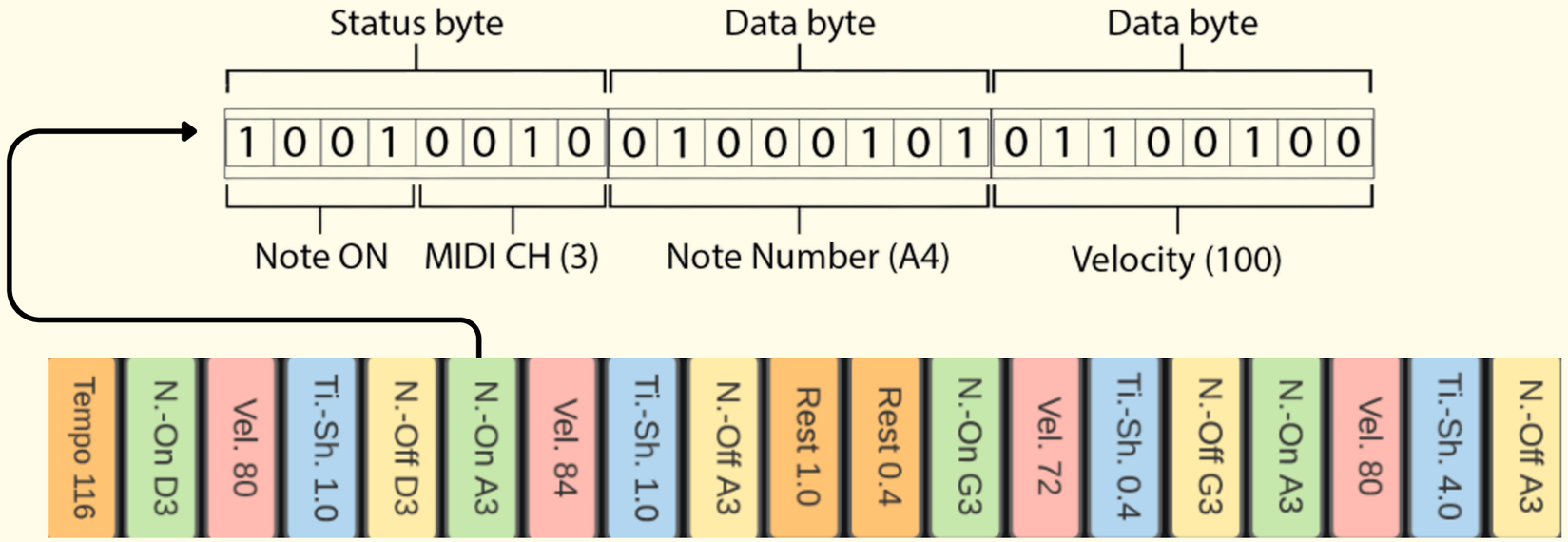


Fig 3. Example of sequence of MIDI events, with a focus of the binary encoding of one Note On Event

2. Methods - Tokenizing music

- How to discretize music for LLM purposes ?
- First approach : **MIDI-Like** (cf last slide) [1]
- Improves: **REMI** (Revamped MIDI) [2]

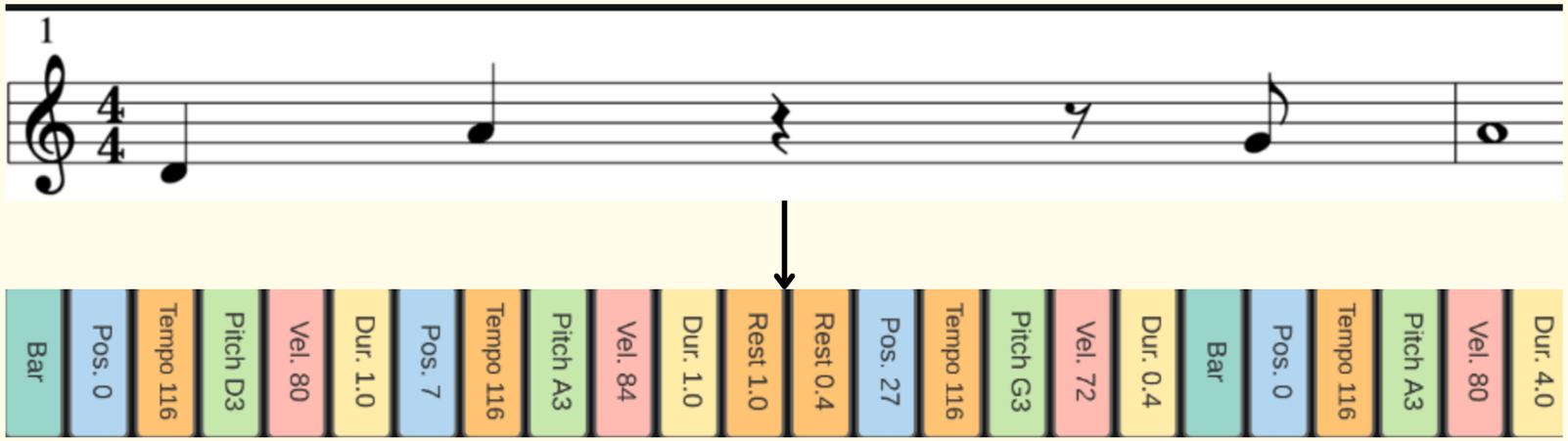


Fig 4. Translation of structural music into REMI tokenization

- Other approaches : more info/token

[1] : Oore, S., et al (2018). This Time with Feeling : Learning Expressive Musical Performance (arXiv:1808.03715). arXiv. https://doi.org/10.48550/arXiv.1808.03715

[2]: Huang, (2020). Pop Music Transformer : Beat-based Modeling and Generation of Expressive Pop Piano Compositions. https://doi.org/10.1145/3394171.3413671

2. Methods - BPE Enhancing

- Vocab base size : 408
  - Needs enhancing, better performance
- **Solution : BPE** (Byte Pair Encoding)
  - 408 20k vocab size
  - Better results for music generation [3]
- Overall a 32% compression ratio

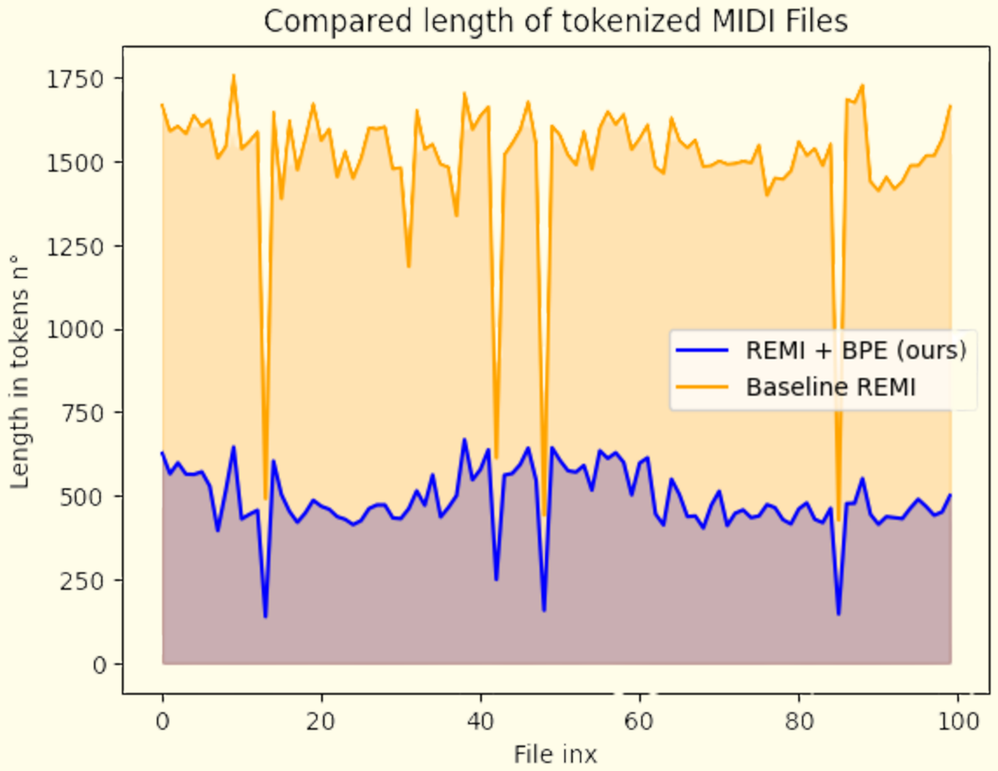


Fig. 5 : Comparison of sequence length

[3]: Chou, Y.-H., et al. (2024). BERT-like Pre-training for Symbolic Piano Music Classification Tasks (arXiv:2107.05223). arXiv. https://doi.org/10.48550/arXiv.2107.05223
https://leria.univ-angers.fr/wp-content/uploads/2023/09/Prez-Seminaire-LERIA-2023.pdf

2. Models - Encoder/Decoder

- **What model to chose ?**

- **BERT** :
  - + : Bidirectional context understanding, good for downstream task (eg. classification), handles long-term dependencies
  - - : Somewhat computationally expensive, long inference, encoder-only

- **GPT** :
  - + : Fast generation, few-shot learning, relatively easy inference
  - - : Computational costs for good perfomance

2. Models - BART

- **BART** : a solution that combines both approaches [4]

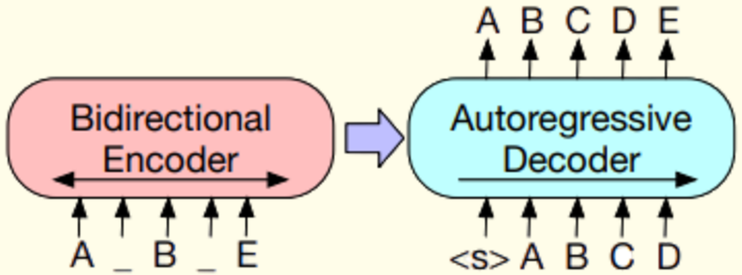


Fig. 6 : Strucural concept behind BART

- **Model pre-training :**

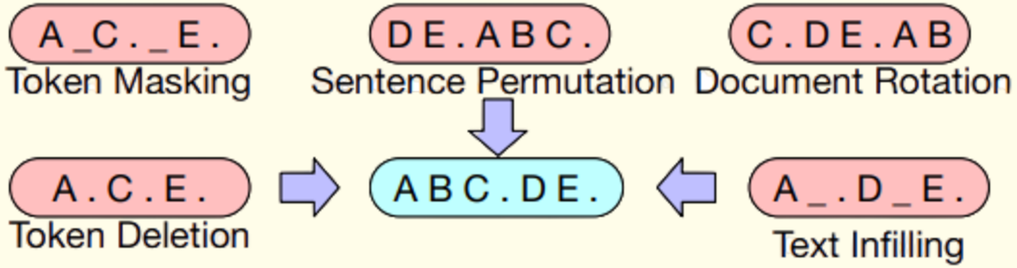


Fig. 7 : BART Tokenizers pre-training

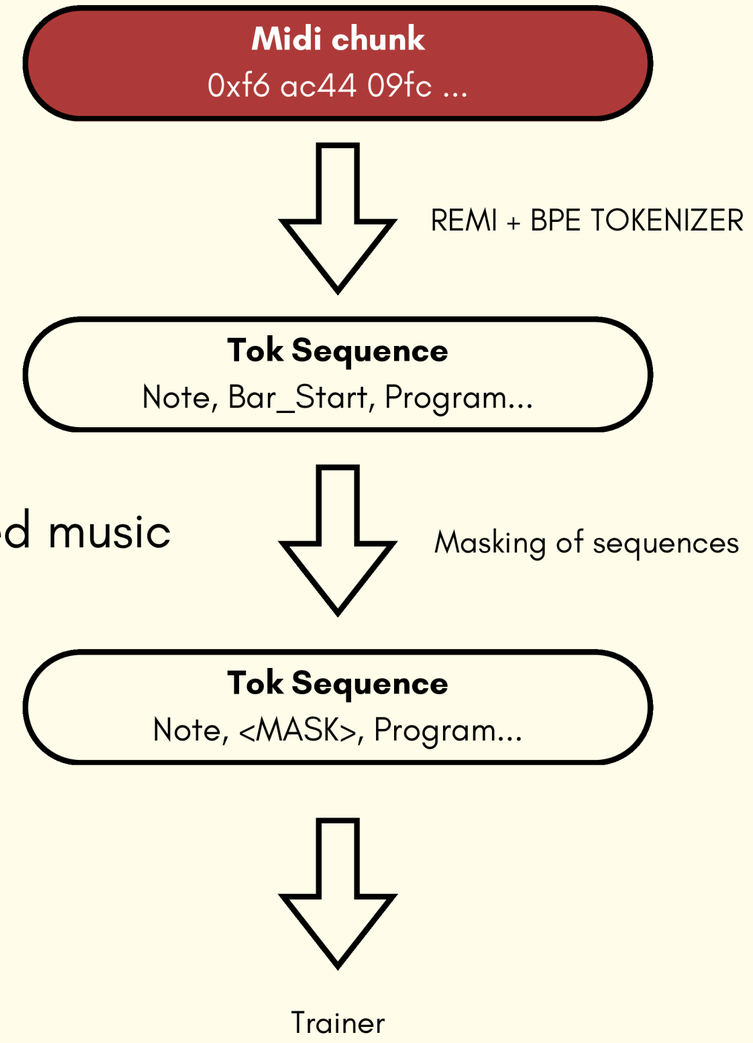
[4]: Lewis, et al (2019). BART : Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (arXiv:1910.13461). arXiv. Chou, Y.-H., et al. (2024). BERT-like Pre-training for Symbolic Piano Music Classification Tasks (arXiv:2107.05223). arXiv. https://doi.org/10.48550/arXiv.2107.05223

3. Experiments - training and issues

- **Items :**
  - Dataset : Giant Midi Piano, 10k files ;
  - Tokenizer : our REMI-BPE ;
  - Model : BartForConditionalGeneration;
  - Loss : BCE over entire sentence.

- **Objective :**
  - Train for reconstruction of masked sequences
  - Inference : transform masked transition into generated music

- **Data pretreatment :**
  - Split the MID files into chunks of fixed length (40s)
  - Tokenize using REMI_BPE
  - Add random masking, sequences of masking

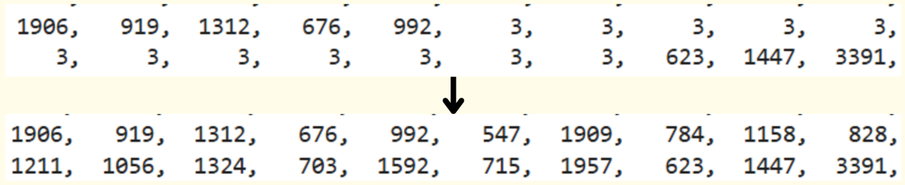


3: Experiments - Masking Protocol

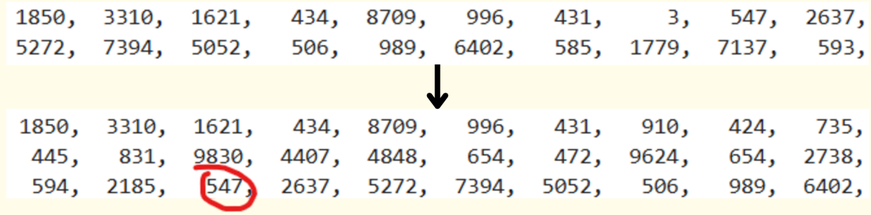
- **Masking protocol :**
  - Length : Poisson's Law $\frac{\lambda^k e^{-\lambda}}{k!}.$
  - Ignores special token
  - Attention mask

- **High training instability**

- **Training method 1:** Sequence masking
  - replace 15 tokens by **15 mask tokens**
  - loss after training **0.213**



- **Training method 2:** Implicit sequence masking
  - replace 15 tokens by **1 mask token**
  - loss after training **0.432**



4. Analysis - Reconstruction

- **Does the model reconstruct masked sequences well ? :**
  - 4% common tokens in generated parts
  - 69% whole data for 2nd model

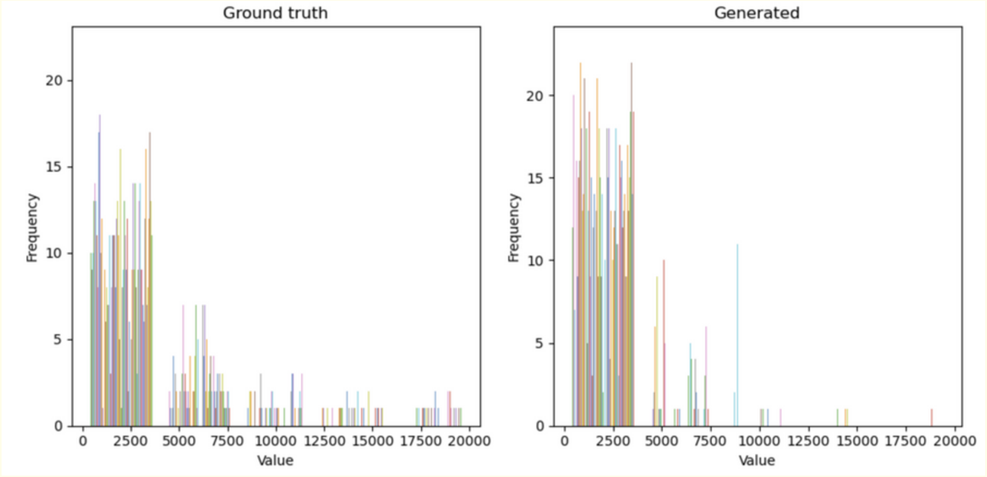


Fig. 8 : Generation **diversity** between approaches, calculated on the test dataset

First approach

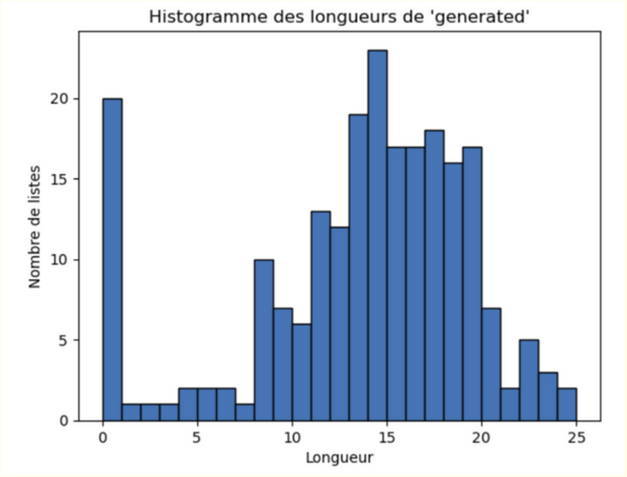
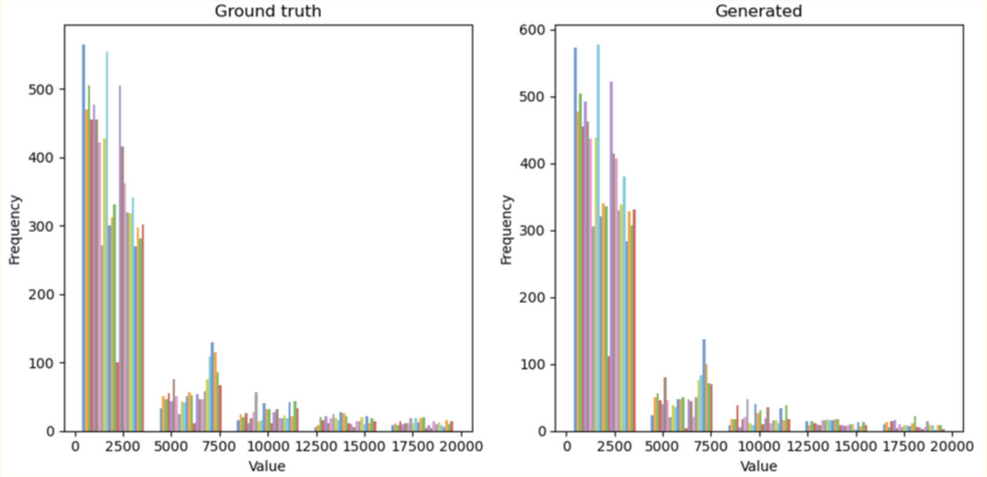
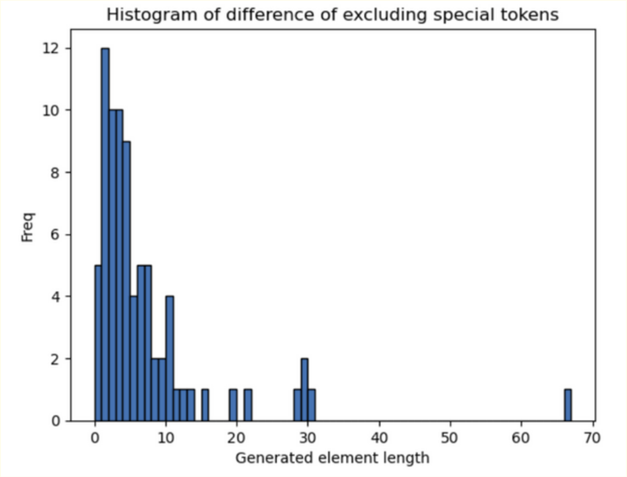


Fig. 9 : Generation **length** between approaches



Second apprach



4. Demos

- **Reconstruction and generation**
  - Example 1 :

Orginal sample

Reconstructed sample

[2] Reference

4. Demos

- **Reconstruction and generation**
  - Example 1 :

Orginal sample

Reconstructed sample

[2] Reference

4. Demos

- **Reconstruction and generation**
  - Example 1 :

Orginal sample

Reconstructed sample

[2] Reference

4. Demos

- **Reconstruction and generation**
  - Example 1 :

Orginal sample

Reconstructed sample

  - Example 2 :

Orginal sample

Reconstructed sample

[2] Reference

4. Demos

- **Reconstruction and generation**
  - Example 1 :

Orginal sample

Reconstructed sample

  - Example 2 :

Orginal sample

Reconstructed sample

[2] Reference

4. Demos

- **Transition generation**

Orginal sample

- **Ending of a piece**

[2] Reference

4. Demos

- **Transition generation**

Orginal sample

- **What about random noise ?**

- **Ending of a piece**

Orginal sample

[2] Reference

RANDOM

- **Reconstruction and generation**

Orginal sample

- **Ending of a piece**

Orginal sample

- **What about random noise ?**
  - Not quite good !...

[2] Reference

5. Conclusion

- LLM are an interesting paradigm for conditional music generation. Using custom REMI scheme tokenizers and BPE, we could encode MIDI files.

- Our BART model was able to **reconstruct** masked inputs and generate **de novo transitions**.

- Next improvements could be on more acurate, less redundant tokenizers, and better parameters optimization.

Thank you for your attention

BART Model architecture