Multimodal Self-Supervised Approach to Text to Music Generation

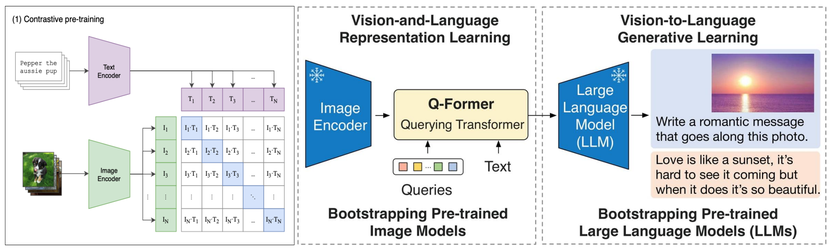
Nick Nguyen, Computer Systems Engineering
Mentor: 'YZ' Yezhou Yang, Associate Professor
Ira. A Fulton School of Engineering

Objective & Research Question:

This research focuses on developing conditioning discrete music ability by proposing a transformer model that merges an encoder text language model and a decoder music language model, with the purpose of capturing the text-music relationship to learn text-music generation.

Background:

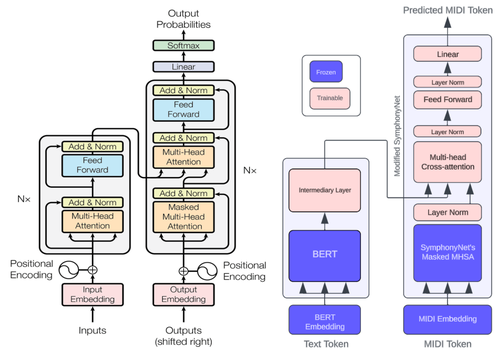
- Many discrete music decoders are capable of music generation but they are unconditional and difficult to control. (give name)
- There exists a variety of models that have rich text knowledge and can represent unstructured text documents to make them become learnable vectors.
- Recently, Text-conditional Image Generation tasks have been attempted successfully by adding and utilizing the cross-modalities learning concept (CLIP, Perceiver). Many successful text-to-image models such as GPT4, Kosmos, BLIP.
- A good starting approach to text conditional music generation is merging the text vector representations with the discrete music vector representations prior to the generation stage.



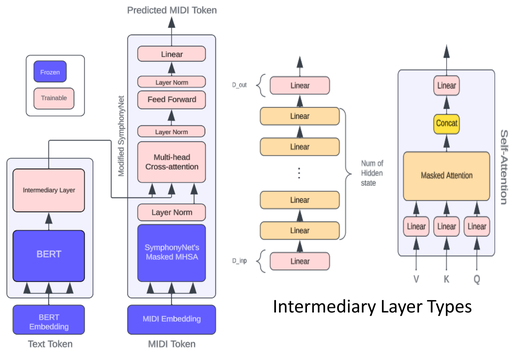
CLIP BLIP

Methodology:

- Two approaches:
  - Inspired by the original Transformer architecture. Apply cross-attention as a middle block between the encoder and the decoder.
  - Utilize the Autoencoder and create a linear text-to-music decoder model
- Intermediary Layer is added to enhance the cross-learning capabilities. It is either a block of sequential linear layers or a multi-head self-attention block.



Transformer Architecture



Proposed Architecture

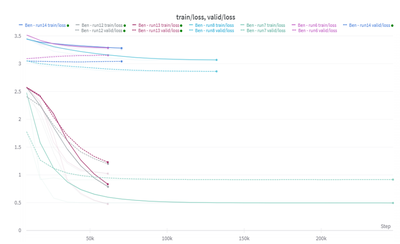
Intermediary Layer Types

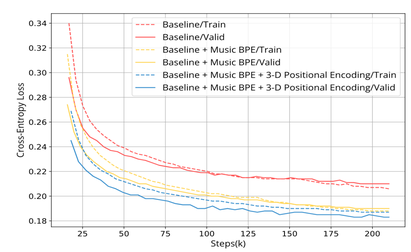
Execution

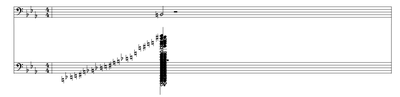
- BERT and SymphonyNet models are chosen for this research.
- Cross-attention along with the entire end-to-end model is built with Fairseq, FastTransformers, & HuggingFace libraries.
- Training is conducted on both scenarios: fine-tuning (frozen) and transfer-learning (unfrozen) two pre-trained models.
- Trained on over 40,000 samples of pairs of English text poems and MIDI scraped from multiple websites.

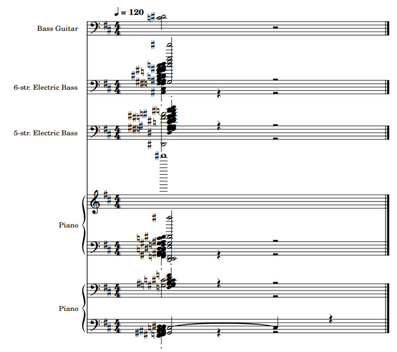
Results

- There is no sign of underfitting which is a good indicator that the proposed end-to-end model is able to learn.
- Our best performance loss is close to the baseline loss.
- However, the model still struggles to generate musical pieces from text and position some musical pieces.
- Future work will increase the complexity of the cross-modalities learning component and involve Contrastive Learning concept implementation









Credits: Thanks Benjamin Herrera, Minglei Kang, and Zeel Shah for support and contribution to this project

FURI

Ira A. Fulton Schools of Engineering
Arizona State University