VIVYNet: A New Approach to Creating AI-Generated Music

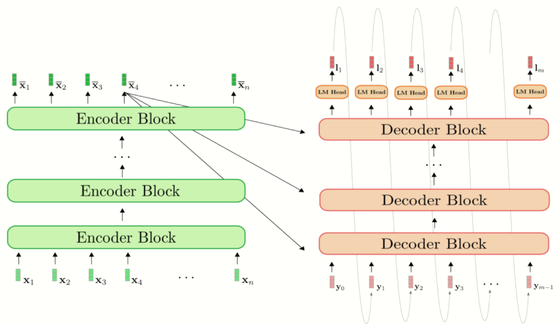
Benjamin Joseph L. Herrera, Computer Science
Mentor: 'YZ' Yezhou Yang, Ph.D.
School of Computing and Augmented Intelligence

Abstract

In today's current AI research scene, language models can map word embeddings to multiple different other types of other domains and generate content or policies from this. This includes videos, images, music, and action spaces. However, many projects have not efficiently procured ways to generate such content, especially in the text to music generation. In addition to this dilemma, very few projects have tackled the issue where mappings between two different spaces are not too coherent with one another (e.g., text to music). We propose an efficient solution that not only aims to improve generation quality from mappings between two different incoherent spaces, but also improve its training efficiency. We present our ideas through text-to-music content generation.

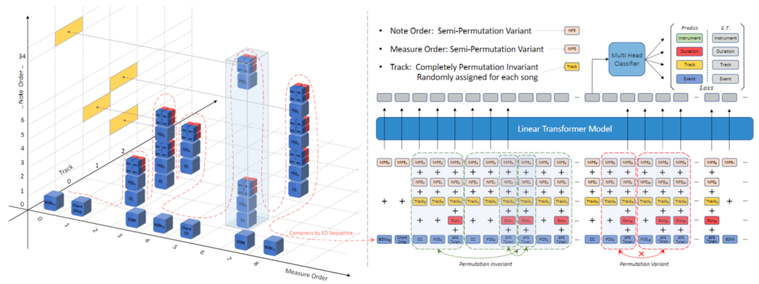
Related Work

Leveraging Pre-trained Checkpoints for Sequence Generation Tasks (Rothe et al., 2020)



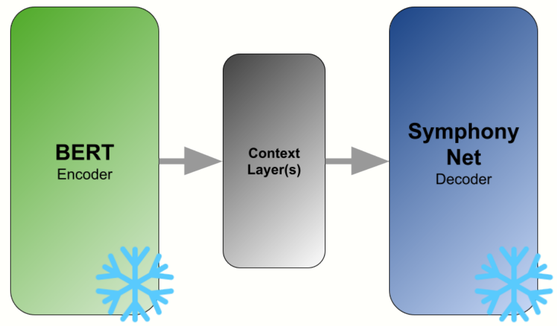
- Utilizes pretrained encoders and decoders to achieve better performances for NLP tasks
- Pretraining can be skipped altogether, saving time and money for fine-tuning

Symphony Generation with Permutation Invariant Language Model (Liu et al., 2022)



- Generates a full music sequence from a couple of measures worth of notes from varying instruments
- Decoder model that uses additional tokens to represent one note (one to many)

Research Work



- Applying the method from Rothe et al. (2020), we can utilize a BERT (Devlin et al., 2019) multilingual model as an encoder and SympnonyNet as a decoder.
- In the middle of the autoencoder is a context layer(s)
- Freezing the encoder and decoder models in this autoencoder, ensures that mapping between encoder to decoder is easier.
- The only weights being updated during the training process should only be the context layer(s).
- This layer generally holds decoder transformers because of its generative properties.
- This ensures that the context layer(s) learns how to map the vocabulary from the encoder into the vocabulary of the decoder

Progress and Final Remarks

This research is still ongoing to find the best approach and configuration with the context layer(s). Experiments so far have shown potential mappings between text and music spaces, but the current mappings, don't yield quality results. A reason for this is because since the encoder outputs one embedding for a single character group, while the decoder intakes multiple embeddings for one single note. This means that the context layer must map one token from the encoder to multiple different tokens for the decoder. In the future, we look forward to using the techniques discovered in this work in other applications and projects.

References

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Rothe, Sascha, Shashi Narayan, and Aliaksei Severyn. "Leveraging pre-trained checkpoints for sequence generation tasks." Transactions of the Association for Computational Linguistics 8 (2020): 264-280.

Liu, Jiafeng, et al. "Symphony Generation with Permutation Invariant Language Model." arXiv preprint arXiv:2205.05448 (2022).

FURI

Ira A. Fulton Schools of Engineering
Arizona State University