International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

LLM based Music Captioning

Ms. Yogita Chavan
Department of Computer Engineering
New Horizon Institute of Technology and Management
India
e-mail: yogitachavan@nhitm.ac.in

Yash Palkar
Department of Computer Engineering
New Horizon Institute of Technology and Management Thane, Thane, India
e-mail: yashpalkar202@nhitm.ac.in

Ajay Gore
Department of Computer Engineering
New Horizon Institute of Technology and Management
India
e-mail:ajaygore202@nhitm.ac.in

Apurv More
Department of Computer Engineering
New Horizon Institute of Technology and Management Thane, Thane, India
e-mail: apurvmore202@nhitm.ac.in

***Abstract*—Automatic music captioning, a process that generates natural language descriptions for music tracks, is pivotal in enhancing the comprehension and organization of vast musical datasets. However, researchers encounter challenges due to the scarcity and labor-intensive nature of collecting existing music-language datasets, often limited in size. To address this data scarcity issue, this project proposes leveraging large language models (LLMs) to artificially generate description sentences from extensive tag datasets. This initiative leads to the creation of a novel Large Language Model based Pseudo music caption dataset. The project systematically evaluates this large-scale music captioning dataset using various quantitative evaluation metrics from the natural language processing domain, supplemented by human evaluation. Additionally, the project trains a transformer-based music captioning model and assess its performance under both zero-shot and transfer-learning settings. The findings indicate that proposed approach surpasses the performance of a supervised baseline model, underscoring the efficacy of employing LLMs for music captioning tasks.**

***Keywords*—Automatic music captioning, natural language descriptions, large language models (LLMs), pseudo music caption dataset, MusicCaps, dataset generation, quantitative evaluation metrics, human evaluation, transformer-based model, zero-shot learning, transfer learning, music understanding, data scarcity, music-language datasets.**

I. INTRODUCTION

Music captioning is a music information retrieval (MIR) task of generating natural language descriptions of given music tracks. The text descriptions are usually sentences, distinguishing the task from other music semantic understanding tasks such as music tagging. Recently, there has been some progress in music captioning including track-level captioning and playlist-level captioning. These approaches usually utilize a deep encoder- decoder framework which was originally developed for neural machine translation using a pre-trained music tagging model as a music encoder and an RNN layer initialized with pre-trained word embeddings for text generation. introduced a temporal Attention mechanism for alignment between audio and text by pairing a pre-trained harmonic CNN encoder with an LSTM layer. generated playlist titles and descriptions using pre- trained GPT-2 Currently, the primary challenge of track- level music captioning is the scarcity of large-scale public datasets. used private production music datasets. also used a private dataset with 44M music-text pairs on YouTube, to address this data issue, a community-driven data collection initiative has been proposed. As of now, the only publicly available dataset for track-level music captioning is Music- Caps.

II.PROPOSED SYSTEM

Large Language Models (LLMs) are a class of neural network- based models known for their immense size and deep learning capabilities, primarily leveraging the Transformer architecture. These models, like the GPT series by OpenAI, are trained on vast datasets of text from the internet, enabling them to understand and generate human- like text. The second model is composed of five main building blocks: a text embedding module, an audio feature extractor, a multimodal encoder, an attention mechanism and a natural language decoder.

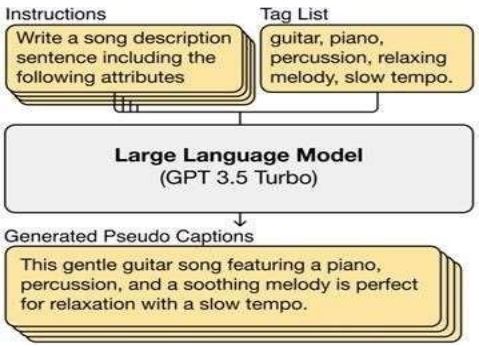

Figure 1: LLM based Model

This model is composed of five main building blocks: a text embedding module, an audio feature extractor, a multimodal encoder, an attention mechanism and a natural language decoder. An overview of the architecture is presented in figure below.

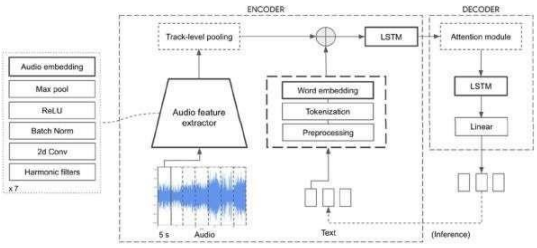

Figure 2: Encoder Decoder Model

III.IMPLEMENTATION PLAN

1. Initiation: Clearly state the goals and desired outcomes of the Music Captioning project. Identify team members and their respective roles in the project. a high-level schedule for the project and allocate the necessary budget and resources.

2. Requirements: Specify the scope of the project, such as the types of music content to be captioned and the desired captioning objectives (e.g., lyrical analysis). Gather a diverse dataset of music content and add labels or annotations to thisdata to train and evaluate the model. Divide the dataset into segments for model training and evaluation purposes.

3. Model Selection: Select an appropriate pre-trained Language Model (LLM) that will serve as the foundation for your captioning system. Adapt the LLM to the specific task of music captioning by training it on the annotated music dataset.

4. System Development: Create the software infrastructure that will utilize the fine-tuned model to generate captions for music content. Develop tools for processing and analyzing music data, as well as a user-friendly interface for interactingwith the system.

5. Evaluation: Establish criteria for assessing the quality and accuracy of the captions generated by the system. Test the system with validation and testing datasets to ensure it meetsthe defined criteria.

6. User Testing: Involve potential end-users to provide feedback on the system's usability, accuracy, and overall experience. Make necessary improvements and adjustments to the system based on the feedback received from users.

7. Documentation: Develop guides and materials to help end- users and developers understand how to use and work with the system. Ensure that the project team is well-trained and knowledgeable about the system's operation and maintenance.

8. Deployment: Make the system accessible to the target audience, whether through a web service, application, or API. Set up tools to continuously monitor the system's performance and user engagement.

9. Maintenance: Continuously improve the language model with more data and better training techniques. Provide ongoing maintenance and support to keep the system functioning smoothly.

10. Launch: Launch the system.

11. Post-launch: Keep an eye on user feedback and system performance in a real-world environment. Based on feedback and real-world usage, continue to make improvements to the system.

12. Scalability and Future Development: Consider future enhancements and expansion.

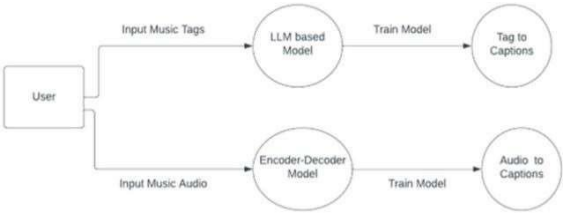

Figure 3: Project Workflow

IV. ALGORITHM

Algorithm for LLM based Model

1. Data Preprocessing: Load the dataset containing music tags and caption pairs. Tokenize the captions and tags. Preprocess the data by removing special characters, lowercasing, and stemming/lemmatizing.

2. Train LLM: Fine-tune a pre-trained language model (such as GPT) on the music caption dataset. Fine-tuning involves updating the parameters of the pre-trained model using the music caption dataset.

3. Generate Music Descriptions: Input a music tag or caption to the fine-tuned LLM. Use the LLM to generate music descriptions based on the input. Optionally, you can specify a maximum length for the generated descriptions to prevent overly long outputs.

4. Evaluation: Evaluate the quality of the generated descriptions using metrics like BLEU score, ROUGE score, or human evaluation. Adjust the fine-tuning process or model architecture based on evaluation results to improve the quality of generated descriptions.

Algorithm for Encoder Decoder Model

1. Preprocessing: Convert audio inputs into spectrograms or other suitable representations. Tokenize text captions to prepare them for input into the model.

2. Model Architecture: Utilize a Cross-Modal Encoder- Decoder Transformer architecture. The encoder processes audio inputs to extract meaningful representations. The decoder generates captions based on the encoded audio representation.

3. Initialization: Initialize the encoder and decoder components of the CMED model. Use pre-trained models for

audio encoding (e.g., VGGish) and text decoding (e.g., BERT), or train them jointly from scratch.

4. Training: Define a suitable loss function (e.g., cross- entropy) to measure the dissimilarity between generated captions and ground truth captions. Train the CMED model using paired audio-caption data, optimizing the loss function through techniques like back propagation and gradient descent. Fine-tune the model as necessary to enhance performance.

5. Inference: Given a new audio input, pass it through the encoder to obtain its representation. Use the encoded representation as the initial state for the decoder. Generate captions iteratively by predicting the next word until an end-of- sequence token is generated or a maximum caption length is reached. Post-process the generated captions to remove special tokens and convert them into readable text.

6. Evaluation: Evaluate the model's performance using metrics such as BLEU, METEOR, ROUGE, or CIDEr, comparing generated captions with ground truth captions. Iterate on the model architecture and training process to improve performance if necessary.

7. Deployment: Deploy the trained CMED model in a production environment, allowing it to accept audio inputs and generate captions either in real-time or batch mode.

8. Monitoring and Maintenance: Continuously monitor the deployed model's performance and user feedback. Update the model with new data or retrain it as needed to maintain or enhance its performance over time.

V. RESULTS

Fig 4: Generation of Music Caption from existing tags

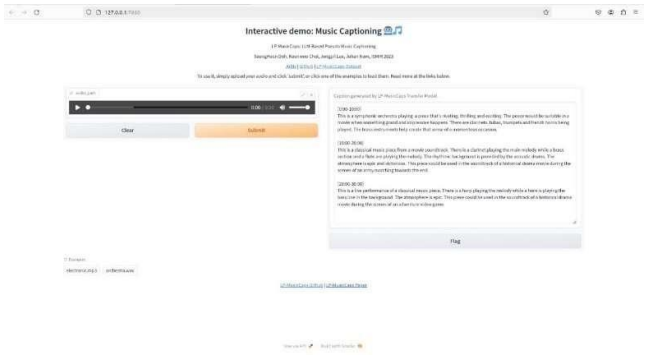

Fig 5: Caption generation from Encoder-Decoder Model

VI. CONCLUSION

A tag-to-pseudo caption generation approach is proposed with large language models to address the data scarcity issue in automatic music captioning. We conducted a systemic evaluation of the LLM-based augmentation, resulting in the creation of the LP-MusicCaps dataset, a largescale pseudo- music caption dataset. The project also trained a music captioning model with MusicCaps and showed improved generalization. The proposed approach has the potential to significantly reduce the cost and time required for music- language dataset collection and facilitate further research in the field of connecting music and language, including representation learning, captioning, and generation. However, further collaboration with the community and human evaluation is essential to enhance the quality and accuracy of the generated captions. Additionally, it is believed that exploring the use of LLMs for other topics under music information retrieval and music recommendation could lead to novel and exciting applications.

ACKNOWLEDGMENT

In this survey paper, we extend our heartfelt gratitude to Ms. Yogita Chavan for her invaluable contributions. Ms. Yogita Chavan's guidance has been instrumental in shaping the innovative solutions proposed in this project.We are deeply grateful for her mentorship, support, and dedication throughout the duration of this project. Her expertise has been a driving force behind the success of our endeavor, and we sincerely appreciate her contributions.

REFERENCES

[1] SeungHeon Doh, Keunwoo Choi, Jongpil Lee, Juhan Nam, "LP-MusicCaps: LLM-Based Pseudo Music Captioning" in the 24th International Society for Music Information Retrieval Conference (ISMIR 2023).

[2] H. Kim, S. Doh, J. Lee, and J. Nam, "Music playlist title generation using artist information," in Proceedings of the AAAI-23 Workshop on Creative AI Across Modalities, 2023.

[3] S. Doh, M. Won, K. Choi, and J. Nam, "Toward universal text-to-music retrieval," in IEEE InternationalConference on Acoustics, Speech and Signal Processing(ICASSP), 2023.

[4] G. Gabbolini, R. Hennequin, and E. Epure, "Dataefficient playlist captioning with musical and linguistic knowledge," in Proceedings of the 2022 Conference onEmpirical Methods in Natural Language Processing (EMNLP), 2022.

[5] T. Chen, Y. Xie, S. Zhang, S. Huang, H. Zhou, and J. Li, "Learning music sequence representation from text supervision," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.

[6] S. Doh, J. Lee, and J. Nam, "Music playlist title generation: A machine- translation approach," in Proceedings of the 2nd Workshop on NLP for Music and Spoken Audio (NLP4MuSA), 2021.

[7] Manco, E. Benetos, E. Quinton, and G. Fazekas, "Muscaps: Generating captions for music audio," in International Joint Conference on Neural Networks (IJCNN). IEEE, 2021.

[8] T. Cai, M. I. Mandel, and D. He, "Music auto tagging captioning," in Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA), 2020.

[9] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in International Conference on Learning Representations (ICLR), 2020.

[10] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in Proceedings of the Advances in neural information processing systems (NeurIPS), 2017.