MSc Project - Reflective Essay

Project Title:	Bridging Paintings and Music – Exploring Emotion based Music Generation through Paintings
Student Name:	**Tanisha Hisariya**
Student Number:	**220929356**
Supervisor Name:	Jinhua Liang
Programme of Study:	MSc Artificial Intelligence

This research paper outlines the innovative intersection of visual art, i.e. transiting from early to modern paintings and music, with the goal of preserving and translating their emotional content and creative essence. It also aspires to create a unique auditory experience by improving the accessibility for visually impaired persons, providing them with a completely new and integral way, inspired by a modern text-to-speech system. The core methodology of this project is carried out by combining two deep learning models- the first for text translation from image and the other for text-to-music generation. Besides, the study explores about how different textual prompts affect the quality and resonance of music. It also addresses the critical gaps in the existing landscape of Generative AI, especially the discrepancy between the required prompts of models for optimal performance and varied prompts provided by users. In this reflective essay, we will analyse the strengths and weaknesses of the project, further highlighting the possibility of future work. We would also critically analyse the relationship between theory and practical work and the problems faced during implementation. After that, we will highlight the project's ethical, social and legal implications, preceding with a conclusion.

Strengths and Weaknesses of the Project

The project shows an important impact of AI based music generation system, leveraging the advance transfer learning technique and integration with different models. The basic strength lies in the systematic and fine-tuning approach of optimising, state of the art MusicGen (Copet et al., 2023) model. During the exploration and experiments, by introducing a better optimised way during the MusicGen training process, which is performed with the help of certain modifications, the model is able to achieve better results successfully. The modifications are further supported by evaluation techniques and theoretical understandings, resulting in a more efficient and robust architecture. This not only reduces the computational complexity but also increases the scalability, exposing unexplored area and possibility of further work in this domain.

Another novel contribution of this project is rigorous exploration of the influence of textual prompts on the generated music. By finetuning the model with a combination of text input, which are mainly based on emotions, it explores the behaviour of different prompts and their impact on the generated music by inspecting its structural characteristics. The analysis and discovery were further supported by a series of experiments, where we evaluated the output on a variety of objective metrics, which tells about the resemblance of music, its divergence capacity, and the amount of noise it holds during the generation process. The findings highlight the critical gap in these text-to-music generation models in terms of the prompt's effect on the generational process, giving a critical challenge to the models and their use case exploration.

Besides, the project examines a variety of combinations with different models for the required generation. The two noteworthy approaches given to this are- the use of the pre-trained ResNet50 model modified with the additional GRU and attention layer, making the accuracy improved by 60%, which was the best achievement so far with the number of complexities held by our painting dataset. We are using prompt engineering techniques by using Falcon 1B model to enrich the contextual meaning of textual prompts. The integration of Falcon 1B for the generation of more refined and nuanced prompts, which in turn led to richer and appropriate musical outputs. This methodological improvement presents the possibility of combining advanced NLP models like BERT and GPT with generative AI systems.

However, this project has some underlying weaknesses, such as a restricted dataset, which could have been more comprehensive to understand the emotions better; a more extensive dataset can lead to a better modality approach. This experiment also lacks subjective evaluation, which may give us more insights into the model's performance including the resonance and quality of music. The third issue that always comes up while dealing with this large model is the cost implication with the regressive runs having more days of training with moderate progress. For the model to run with a greater variety of experiments and controlled parameters, it needs more computational power to find the best optimal group of weights for each feature.

Future Work

If I had more time for this project and as a scope of further work, I would first consider annotating a more extensive dataset based on emotions. The new dataset can be collected or created from existing datasets by applying advanced data augmentation techniques. This addition would have given better generalization abilities and improved the performance of the model, particularly in reflection with emotional resonance.

Another key enhancement I suggest is the implementation of parameter-efficient fine-tuning techniques, such as Low-Rank Adaptions (LoRA) (Xu et al, 2023), a technique to efficiently fine tune these large models reducing the computational overhead. By freezing the weights of the original model and introducing a trainable rank decomposition matrix to each transformer layer, LoRA reduces the number of parameters required. Even after this it does not compromise with the integral knowledge of the model's pre-trained weights, making it a cost-effective solution.

Furthermore, to address the problem of blank space and noise that sometimes appears in generated music, I would explore the incorporation of Long Short-Term Memory (LSTM) networks or similar Recurrent Neural Network (RNN) architectures in the post-processing stage. LSTM networks are well-suited for sequential data and could be utilized to smooth out inconsistencies and enhance the continuity and quality of the generated music. The final result would likely show greater coherence by refining the output through these techniques, thereby enhancing the overall experience.

Critical analysis of the relationship between theory and practical work produced

This project has been a real eye-opener for me, as I have learned the major difference between theoretical and practical work. I have always been fascinated with the Generative AI model and its use in creativity. Always working with this technology, I came up with the idea of using cross-modality approaches. Starting with my initial literature review, I began my search in the generative model on the innovative side, which ended up with idea of - generating sounds using images and videos. My journey started with the *im2wav* model (Roy and Yossi, 2022), which uses a two-layer transformer architecture. While going through the paper, I made several experiments based on it, but after so many unsuccessful attempts, I had to change my methodologies after realising their limitations. Furthermore, I studied more advanced models like AudioLDM (Liu et al, 2023) based on diffusion models and MelFusion (Chowdhury et al, 2024), another recently launched model for audio generation through images. However, after multiple trials and careful consideration, I decided that these models did not completely meet the requirements of my project. This directed me to the decision of using a two-way architecture model using a hybrid model approach.

The dataset preparation part was one of the most challenging aspects of this project. There was no existence of paired datasets with the painting and music sharing the same emotions. This led me to create a customised paired dataset using two different forms and joining them based on their shared emotional attributes. One of the most time-consuming parts of my project was to manually analyse both datasets, along with the augmentation of the Music Dataset to make them fit with the model's requirements.

Now, the time has come to implement a classification model for classifying emotion labels and generating a text prompt based on that. During my entire coursework, I have gained a solid understanding of the CNN models for the classification task which provides a baseline for this project work. While implementing these models for my dataset, I was not getting perfect results initially and faced several problems such as overfitting even after trying out with data augmentation and batch normalisation techniques. To solve this challenge, I have researched certain areas and models which ultimately led me to integrate the GRU and attention layer within existing model. This integration enhanced the model's performance, giving me better results and showing the need for adaptability.

While training my integrated model along with MusicGen, I was getting very poor results on my datasets on the first go. This was mainly because of the improper split of training and evaluation set within each class label. These steps made me realize that figuring out the data pre-processing techniques and the data split technique in this kind of model is one of the most important steps to get better results and to avoid repeating the same operations. The planning for subjective evaluation did not go as intended, so we had to include some more objective evaluation metrics to give a more transparent and detailed view of the results.

Lastly, the only thing that is important while performing this project is that you cannot plan all your experimental work while going through theory as sometimes nothing can go as planned and you should always have a wider knowledge on various topics to make some incorporations overcome the challenges.

Legal, ethical and social considerations

As an AI student, I am fully aware of the challenges, including both positive and negative implications of using Artificial Intelligence in the society. This project has the potential to make a strong impact on the artistic expressions for visually impaired persons. This enables a broader range of audiences to engage with it and access the creative realm of it. However, while dealing with such creative models in larger audiences with cultural divergence as the arts are deeply rooted within them, there is always a risk in the interpretation of embedded emotional aspects. To ensure the wider acceptance of the model and to not hurt the sentimental feelings of individuals, it is really important to consider their cultural sensitivity.

The ethical considerations presented by this project are related to biasness, transparency, and consent. The first problem arises due to the potential biasness of the training data used to train the model. If the training data is not diverse enough to handle all the emotional contents, then it will lead to failure in evaluating fairer experiences. The Asilomar principle (Future of Life Institute, 2017) also suggested reducing the biases and increasing the transparency of the system to benefit human values. Moreover, the most significant consideration is the potential misuse of Generative AI content in manipulative ways. It should be clearly mentioned that whether the content is AI-generated content or not. The issue of consent also needs to be taken care of if we are using the artistic creation in order to train or generate the results to meet ethical compliance and ensure the safe handling. In the article "The UK Governments Generative AI Framework" (Evans, 2024), discussed the need of alignment of these models with the values and ethics of humans and organisations.

The legal implications of this project combine IP (Intellectual Property) rights along with copyright law. For example, if the generated music closely matches with the existing work of some other artist, there may be legal challenges regarding its rights and ownership (Oliver Brown, 2024). This problem will lead to the need for legal frameworks to address AI-generated works, which are currently not covered by most of the existing laws and principles.

The sustainability concern for this project is mainly focused on the environmental impact for computational power by training and deploying such large models (Sfondrini et al., 2024). The computer has been under constant load for training and running several experiments, leading to higher energy consumption and increased carbon footprints. In the article "Gen AI's carbon footprint", the author has discussed its regular implications on the environment and also suggested a way to overcome this by integrating it with Quantum Computing.

Personal Development / Conclusion

In conclusion, I have successfully implemented a Generative model to trigger emotional music based on paintings. Through this project, I aimed to use the knowledge which I gained during my last two semesters at QMUL and wanted to implement something based on that. This is the major reason of choosing a multi-modality project. In addition, I have used my soft skills to effectively organise the project implementations according to the timeline and to communicate with my supervisor. This project has laid the groundwork for future innovations, and I am excited about further research and development. Overall, this project developed my professional and personal skills in order to solve the real world AI challenges.

REFERENCES

1. Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y. and Défossez, A. (2023). Simple and Controllable Music Generation.[online]arXiv.org.doi:https://doi.org/10.48550/arXiv.2306.05284.

2. Xu, L., Xie, H., Qin, S.-Z., Tao, X. and Wang, F. (n.d.). Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. [online] Available at: https://arxiv.org/pdf/2312.12148 [Accessed 15 August 2024].

3. Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W. and Plumbley, M.D. (2023). AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. arXiv:2301.12503 [cs, eess] [online] Available at: https://arxiv.org/abs/2301.12503.

4. Chowdhury, S., Nag, S., Joseph, Srinivasan, B. and Manocha, D. (n.d.). MELFUSION: Synthesizing Music from Image and Language Cues using Diffusion Models. [online] Available at: https://arxiv.org/pdf/2406.04673 [Accessed 18 Aug. 2024].

5. Future of Life Institute (2017). AI Principles. [online] Future of Life Institute. Available at: https://futureoflife.org/open-letter/ai-principles/ [accessed on 15 August 2024]

6. Bashir, N., Donti, P., Cuff, J., Sroka, S., Ilic, M., Sze, V., Delimitrou, C. and Olivetti, E. (2024). The Climate and Sustainability Implications of Generative AI. An MIT

Exploration of Generative AI. [online] Available at: https://mit-genai.pubpub.org/pub/8ulgrckc/release/2 [accessed 15 August 2024]

7. Evans, R. (2024). The UK Government's Generative AI Framework: How these principles can help you build ethical, collaborative & diverse AI teams - Digital Futures. [online] Digital Futures. Available at: https://digitalfutures.com/insights/news/gen-ai-how-these-principles-can-help-you-build-ethical-collaborative-diverse-ai-teams/ [accessed 15 August 2024]

8. Oliver B. (2024) “Generative AI: copyright, ethics, and the future of creativity” [online] Available at: https://www.businessthink.unsw.edu.au/articles/generative-AI-creative-industries-copyright-fair-use [accessed 15 August 2024]

9. Sfondrini, N. (2024). Council Post: GenAI's Carbon Footprint: A New Challenge For Corporations. Forbes. [online] 12 Aug. Available at: https://www.forbes.com/councils/forbestechcouncil/2024/03/28/genais-carbon-footprint-a-new-challenge-for-corporations [accessed 15 August 2024]