Generating Music from Paintings: A Multimodal AI Approach

Dina Ahmed El Sayed Fathy
Arab Academy for Science, Technology and Maritime Transport College of Computing & Information Technology

Supervised by Prof. Dr. Mohamed Abo Rezka
Arab Academy for Science, Technology and Maritime Transport College of Computing & Information Technology

***Abstract*—This paper presents a novel approach to generating music from paintings using deep learning models. The system integrates image captioning, language translation, and music synthesis to produce culturally relevant music compositions. We employ models such as BLIP, CLIP, T5, MusicGen, and MarianMT. The results demonstrate the effectiveness of AI-driven multimodal learning in bridging visual art and music.**

***Index Terms*—AI, deep learning, music generation, image captioning, multimodal learning, cultural preservation.**

I. Introduction

The disconnect between visual art and music limits creative expression, cultural preservation, and accessibility. Our system translates paintings into emotionally and culturally relevant music using AI models. The proposed framework enables real-time music generation and offers a user-friendly interface. We discuss the interdisciplinary approach combining computer vision, natural language processing, and music synthesis.

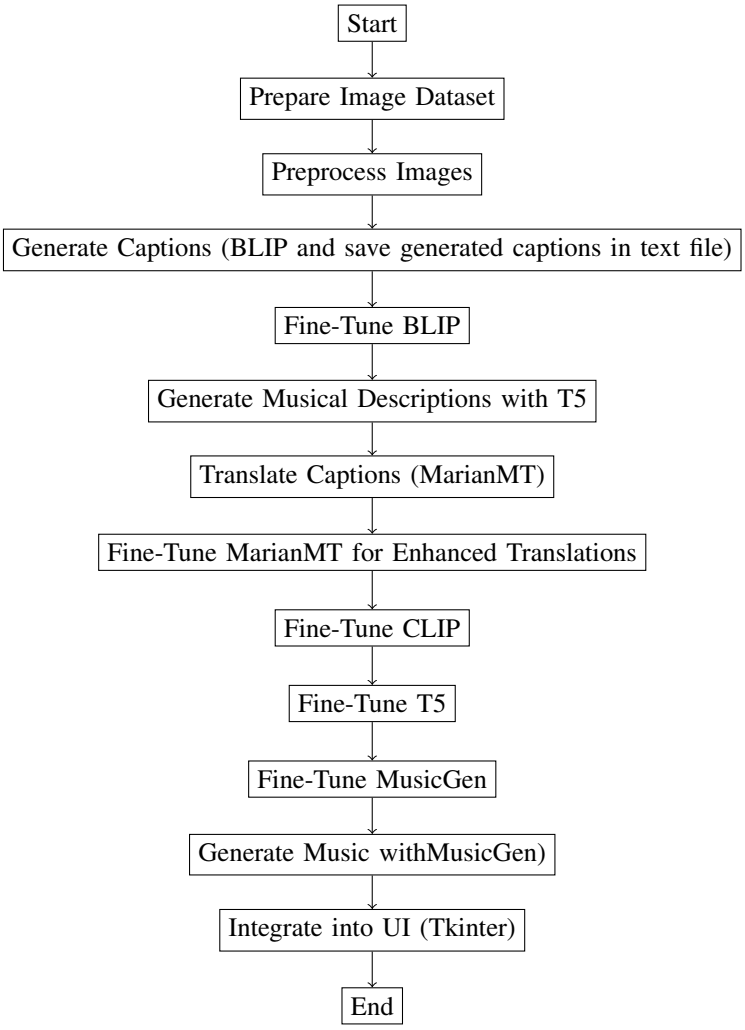
II. Workflow

The system's workflow is structured as follows:

1) Collect image datasets from museums.
2) Preprocess images using noise reduction and contrast enhancement techniques.
3) Generate captions using BLIP and refine them using T5.
4) Translate captions into Arabic using MarianMT if needed.
5) Predict appropriate musical elements using CLIP.
6) Generate music using MusicGen based on the refined textual descriptions.
7) Integrate all modules into a Tkinter-based user interface for real-time interaction.

A. Algorithm Workflow

The workflow of our system follows these steps:



III. AI Basics

A. Loss Function

Is a crucial component in machine learning that quantifies the difference between the predicted outputs of a machine learning algorithm and the actual target values.Also referred to as the error function

B. Cross Entropy Loss Function

- Cross-Entropy is a loss function used in classification tasks.
- It measures the difference between two probability distributions
- The true distribution (actual labels).
- The predicted distribution (model's predictions).
- It is commonly used in multi-class classification, where each instance belongs to one of multiple classes

C. Gradient

Is a vector that points in the direction of the steepest increase of a function.

D. Gradient Descent

Is an optimization algorithm used to minimize a loss function . To minimize a loss function we go in the opposite direction of the gradient.

E. Adam Optimizer

- Is one of the most widely used optimization algorithms in deep learning.
- It combines the best properties of Momentum-based Gradient Descent and RMSprop to adaptively adjust the learning rate for each parameter.

IV. Learning Paradigms

A. Self-Supervised Learning

Self-supervised learning enables models to learn from unlabeled data by creating pretext tasks. This is used in our model to refine captioning and audio synthesis tasks.

B. Transfer Learning and Fine-Tuning

Pretrained models such as ViT and T5 are fine-tuned on our dataset to improve performance in captioning and music generation

V. AI Model Explanation

A. Transformer Learning

The Transformer model, introduced by Vaswani et al., revolutionized deep learning by enabling parallelization and improving performance for sequence-to-sequence tasks. It uses an attention mechanism that enhances context understanding for both short- and long-range dependencies.

B. Vision Transformer (ViT)

ViT adapts the Transformer architecture for image processing by treating images as sequences of patches. It has shown superior performance over traditional CNNs when trained on large datasets. In our system, ViT extracts features from paintings to assist in generating captions and music.

C. BLIP

BLIP is an image captioning model that generates descriptions by encoding images into textual representations.

D. CLIP

CLIP is a contrastive learning-based model that aligns images and text to improve contextual understanding.

E. T5

T5 is a text-to-text transformer for text processing, particularly in refining descriptions for better musical context.

F. MusicGen

MusicGen is a transformer-based model trained on large-scale music datasets to generate harmonious compositions from text inputs.

G. MarianMT

MarianMT is a machine translation model for multilingual support, ensuring captions can be understood in multiple languages.

VI. Software Stack

A. Pytorch

- Is open-source deep learning available with Python and C++ interface
- Developed by "Facebook
-

B. Pytorch Key Features

- Data must be processed in form of Tensors
- Automatically computes gradients for backpropagation
- Simplifies building neural networks
- Dynamically computes computational graphs(more flexible for deep learning models)

C. Tensors

- Are specialized data structure that are very similar to arrays and matrices.
- Use them to encode the inputs and outputs of a model, as well as the model's parameters.
- Are like NumPy's arrays, except that tensors can run on GPUs or other hardware accelerators.

D. Hugging Face

Hugging Face is a platform that provides open-source machine learning tools, pre-trained models, and datasets, empowering developers to build and deploy AI solutions in natural language processing, computer vision, and audio.

E. OpenCv

Is huge open source library for computer vison, machine learning and image processing. Allows to perform various operations like image enhancement and object detection

F. Tkinter

Is a wrapper for the Tk GUI toolkit, making it easy to create windows, dialogs, buttons, text fields, and more

VII. Results and Analysis

- The BLEU score demonstrated high-quality caption generation, while MusicGen effectively produced harmonically rich compositions. The system successfully adapted to different artistic styles, preserving the cultural essence of the paintings.
- The generated music was evaluated by human listeners, confirming the emotional and cultural relevance of the compositions.
- Spectral Flatness: Quantifies how noise-like or tonal the audio signal is, with values closer to 1 indicating a noise-like signal and values closer to 0 suggesting a more tonal or harmonic structure.
- Root Mean Square (RMS) Energy: Represents the overall loudness of the audio signal by calculating the square root of its average power, giving an idea of the signal's intensity.
-

The following table presents a summary of the evaluation results:

Metric	Score
BLEU Score	0.75
Spectral Flatness	0.45
RMS Energy	0.62

TABLE I
SUMMARY OF EVALUATION RESULTS.

VIII. Conclusion and Future Work

This study presents a robust AI-driven system for generating music from paintings. The proposed approach integrates advanced deep learning models for multimodal understanding. Future enhancements include:

- Expanding the dataset
- Improving Arabic text refinement
- Adding an AI singer to perform
- Include function to merge tracks and create music
- Incorporating reinforcement learning to optimize music selection based on user feedback.
- Increasing GPU power will enable faster training and inference, allowing for more complex models and higher-quality music generation.
- Deploying the system as a mobile application and integrating advanced AI techniques will further enhance performance and accessibility.
- This study will be expanded in my master's research

This work will be further expanded as part of **my master's research.**

References

[1] Vaswani, A., et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems (NeurIPS).

[2] Dosovitskiy, A., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint.

[3] Hisariya, T., Zhang, H., Liang, J. (2023). Bridging Paintings and Music–Exploring Emotion-Based Music Generation through Paintings. Queen Mary University of London.

[4] Malchiodi, C. A. (2012). Handbook of Art Therapy. Guilford Press.

[5] Styliani, S., Fotis, L., Kostas, K., & Petros, P. (2009). Virtual museums: A survey. Journal of Cultural Heritage.

[6] Briot, J. P., Hadjeres, G., & Pachet, F. (2020). Deep Learning Techniques for Music Generation. Springer.

[7] Panda, R., Malheiro, R., Rocha, B., Oliveira, A., & Paiva, R. P. (2023). Multi-Modal Music Emotion Recognition: A New Dataset, Methodology, and Comparative Analysis. CISUC– Centre for Informatics and Systems of the University of Coimbra, Portugal.

[8] Tan, X., Antony, M. (2021). Automated Music Generation for Visual Art through Emotion. Independent Researcher CognAI, Hong Kong.

[9] Muhamed, A., et al. (2020). Symbolic Music Generation with Transformer-GANs. Amazon Web Services Carnegie Mellon University.