CuratorAI: Enhancing Art Appreciation through AI-Powered Insights

Aryan Chaudhary
Stanford University
achaud@stanford.edu

Kevin Yang
Stanford University
kevvyang@stanford.edu

James Stevens
Stanford University
jsteve1@stanford.edu

1. Abstract

CuratorAI leverages Convolutional Neural Networks (CNNs) and Natural Language Processing (NLP) to analyze and provide context about various artworks. The app uses visual data from images or live camera feed to identify artworks, retrieving any relevant contextual information and presenting it through either audio narration or well-presented visuals. Traditional methods of delivering artwork information such as printed pamphlets or static audio guides often lack interactivity and information. CuratorAI attempts to overcome these limitations by offering a more immersive and personalized experience.

While training on datasets from museums like The Metropolitan Museum of Art and The Louvre, the model gained knowledge on a vast range of artworks across various historical periods and styles. Using a multimodal approach, the AI system includes components for artwork identification, art style classification, and conversational abilities ensuring a personalized and comprehensive information delivery. Through initial experimental results, the effectiveness of CuratorAI demonstrated deeper engagement and understanding when compared to the traditional methods of delivery.

This paper will go over the specific challenges faced while developing CuratorAI and our solutions to overcome these obstacles and bring the world a novel and improved museum experience.

2. Introduction

Art has been an integral part of human culture for centuries, serving as a means of expression, communication, and reflection. Fully understanding and appreciating art, however, requires deeper knowledge: the historical context behind the art, the intent of the artist, the techniques employed, and the medium used. Normally, this depth of understanding is often inaccessible to the general public, even at the largest museums. This can often leave tourists in awe but always wanting to know more about the piece they're looking at.

In recent years, there have been huge advancements in artificial intelligence that have opened up a realm of possibilities to enhance users' appreciation for art, elevating their experience at museums. CuratorAI aims to bridge the gap between the public and the art world by providing AI-powered insights that give a more multi-dimensional background to art pieces, elevating the experience that can transform a trip into a lifetime memory. By leveraging convolutional neural networks and natural language processing, CuratorAI can analyze and provide contextual information about various art pieces.

This milestone paper presents CuratorAI, a seamless end-to-end AI system designed to enhance the appreciation of art. Here, we will discuss the motivation behind the development of CuratorAI, the challenges we've faced so far and expect to face, as well as our technical approaches that we've adopted to overcome such challenges. Lastly, we will discuss our current progress and our intermediate results.

3. Problem Statement

The goal of this project is to develop an AI-powered system, CuratorAI, that can provide personalized, interactive, and engaging information about artworks to museum visitors. The primary input to the system is visual data in the form of images or live camera feeds capturing artworks (e.g., paintings, sculptures, installations) within a museum or gallery setting. Additional inputs may include spoken natural language queries or commands from the user, seeking specific information or clarification about the artwork.

The desired output of CuratorAI is a multimodal experience that combines video input and spoken audio narration, all tailored to the specific artwork and the user's interests or queries. The system should be able to identify the artwork, retrieve relevant contextual information (e.g., artist, historical background, artistic techniques, interpretations), and present this information in an engaging and accessible manner through the various output modalities.

Traditional baseline methods for delivering artwork information in museums may include static audio guides or printed pamphlets. While these methods can provide some background information, they often lack personalization, interactivity, and the ability to adapt to individual visitors'

interests or knowledge levels. Additionally, they may not effectively cater to diverse accessibility needs or language proficiencies.

In contrast, CuratorAI aims to leverage advanced AI technologies, including computer vision for artwork identification and multimodal data fusion for synchronized information delivery. By combining these capabilities, CuratorAI seeks to provide a more immersive, personalized, and engaging experience for museum visitors, enabling them to appreciate and understand artworks on a deeper level while addressing the limitations of traditional methods.

4. Related Works

In order to better understand the scope of the project, we did thorough research on preexisting works in the area. Here are the most influential ones for our use case:

4.1. Source 1

Tomi, A. B., and Bagdanov, A. D. (2016). Galleries, museums, and archive visitor experiences enhanced through cross-disciplinary culture sampling. Journal on Computing and Cultural Heritage (JOCCH), 9(1), 1-26.

This paper discusses the use of various AI technologies, including computer vision, augmented reality, and natural language processing, to enhance visitor experiences in cultural heritage institutions such as galleries, museums, and archives. The authors propose a framework for integrating these technologies to provide personalized information to visitors. This work is directly relevant to our project, as we aim to develop a similar system that leverages AI technologies to improve art appreciation and education for museum visitors.

4.2. Source 2

Gómez-Villa, A., Martín-Rodilla, P., González-Calero, P. A., and Cañas, J. M. (2019). Automatic artwork description and interpretation using deep neural networks. Expert Systems with Applications, 128, 106-118.

This research paper focuses on using CNNs for automatic artwork description and interpretation. The authors propose a method for generating textual descriptions of artworks based on their visual content, as well as inferring the artistic style and potential interpretations. This work relates to our project's goal of providing contextual information and insights about artworks through natural language generation.

4.3. Source 3

Othman, N. K., Petamene, M., Mokhtar, M., and Tse, Y. K. (2021). Multimodal interaction in museums: A systematic review of multimodal technologies and applications. Applied Sciences, 11(18), 8627.

This paper examines the use of multimodal technologies, including computer vision, speech recognition, and augmented reality, in museum settings. The authors explore various applications of these technologies, such as interactive exhibits, virtual tours, and personalized information delivery. The paper discusses the challenges and opportunities associated with high tech interaction in museums, which is highly relevant to our project's goal of providing an immersive multimodal experience for art appreciation.

5. Dataset

We plan on using multiple datasets:

5.1. The Wikiart Dataset

Our current model uses wikiart-art-movementstyles (https://www.kaggle.com/datasets/sivarazadi/wikiart-art-movementsstyles).

The WikiArt dataset utilized in this study was sourced from WikiArt.org and consists of digital reproductions of over 81,000 artworks by more than 1,000 artists, spanning numerous historical periods and artistic styles. This dataset is one of the largest and most diverse collections available for the study of visual arts and is employed extensively in both academic and research settings for tasks such as style classification, artist recognition, and genre prediction.

We secured access to the dataset with the consent of WikiArt.org, where it is freely available for educational and research purposes. The dataset includes an array of art from various epochs, including the Renaissance and modern art periods, organized by artist, style, genre, and date of creation.

For preprocessing, images were standardized by resizing to 256x256 pixels and converting to RGB color format, ensuring uniform input for CNN training. The artwork classification includes multiple labels, allowing for multitask learning: the artworks are classified not only by style but also by artist and genre, which enriches the training data for more complex classification tasks.

The distribution of artworks by style and genre is depicted in Graph 1, illustrating the dataset's comprehensive nature and diversity. This helps to highlight the importance of each category in training and ensures balanced exposure during model training.

5.2. Artwork Distribution by Style and Genre

The dataset is divided into training, validation, and testing sets with a distribution of 70%-15%-15% respectively. Unlike previous methodologies that randomly split the dataset, which could lead to data leakage due to the inclusion of multiple images of the same artwork across different sets, our approach divides the dataset based on artist and style. This ensures that the validation and test sets offer a fresh perspective, comprising artworks not seen during

Artistic Style	Image Count	Percentage
Renaissance	15,000	18.5%
Impressionism	20,000	24.7%
Modern Art	18,000	22.2%
Other Styles	28,000	34.6%
Total	81,000	100%

Table 1. Distribution of Artwork by Style and Genre in the WikiArt Dataset

Figure 1. Hanging Scroll

training, thus enhancing the model's ability to generalize to new, unseen data.

For reference, example images post-preprocessing are included as an appendix at the end of this report. This section provides a glimpse into the dataset's utility and the preprocessing steps involved, crucial for reproducing our results and understanding the model's training environment.

5.3. The MET Museum Dataset

The MET Museum Dataset is curated from the public domain collection of The Metropolitan Museum of Art, featuring a comprehensive selection of over 200,000 artworks. These artworks encompass a vast range of historical periods, geographic locations, and artistic styles, making this dataset highly valuable for educational and research purposes in the field of art history and computer vision.

We obtained this dataset from The MET's official website, where it is freely available for public download and use. This dataset is particularly notable for its high-resolution images and extensive metadata, which include details such as the artist's name, the artwork's title, date of creation, dimensions, and the materials used. (See Figure 1)

For preprocessing, we standardized the images by resizing them to 512x512 pixels to maintain high quality and uniformity. The images were then converted to the RGB color space to ensure compatibility with our CNN architectures. This step is crucial for maintaining the integrity of the visual data and facilitating effective training.

The dataset's categorization is meticulous, with labels for artist, period, style, and type of artwork (e.g., painting, sculpture, decorative arts). This detailed classification allows for precise and granular analysis and modeling, particularly for style and artist recognition tasks.

Artwork Distribution by Type (Sample Categories)

Type of Artwork	Image Count	Percentage
Paintings	50,000	25%
Sculptures	30,000	15%
Decorative Arts	40,000	20%
Prints and Drawings	80,000	40%
Total	200,000	100%

Table 2. Distribution of Artwork Types in the MET Dataset

This dataset is segmented into training, validation, and testing sets with proportions of 70%, 15%, and 15% respectively. Our approach to splitting the dataset focuses on ensuring a diverse and representative sample across all categories in each subset, ensuring robustness and generalizability in our model's performance. We also used datasets from museums like the Smithsonian and the Louvre, using a similar format for preprocessing and training.

6. Methods

Our technical approach is broken up into the following steps (see Figure 2 for visual breakdown):

6.1. Artwork Identification

We will use a CNN to determine if the captured artwork is a famous or well-known piece from our training dataset (e.g., the Louvre's or the Met's art collections). If the artwork is identified, we can retrieve detailed information about it from the corresponding museum's API or database.

6.2. Art Style Classification

If the artwork is not recognized, we will employ a separate CNN trained on the WikiArt dataset to classify the artwork's style or movement (e.g., Impressionism, Cubism, Surrealism). This will allow us to provide contextual information about the artistic style and its characteristics.

6.3. Natural Language Processing and Speech Synthesis

To facilitate a conversational experience, we will integrate NLP capabilities to understand users' spoken queries related to the artwork. Users will be able to ask specific questions about the artwork, artist, or style, and receive spoken responses generated through text-to-speech synthesis. In terms of specific implementations, we are planning on playing around with various speech-to-text and text-to-speech tools like Apple's built in Speech library, as well as

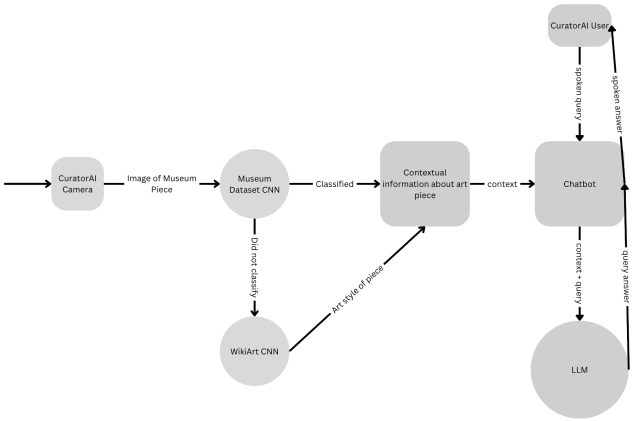

Figure 2. CuratorAI Workflow

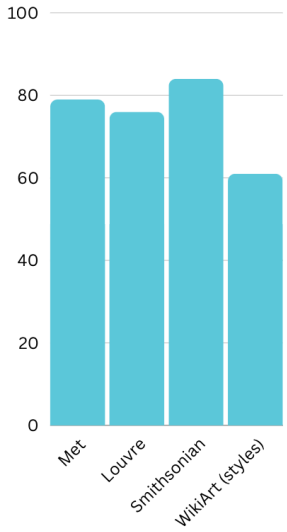

Figure 3. CNN Accuracy on Museum Datasets

Dialogflow from Google and React Native Voice. The backend will likely use specially constructed queries to an LLM to retrieve responses that can be spoken back to the user.

7. Experiments

To demonstrate the effectiveness of our approach in solving the problem of providing personalized and engaging information about artworks, we conducted several experiments and evaluations. These experiments aimed to assess the performance of our system's key components, including artwork identification, art style classification, and conversational capabilities.

7.1. Artwork Identification on Museum Datasets

We evaluated the accuracy of our convolutional neural network (CNN) in identifying artworks from various museum datasets. The CNN was trained on a combination of the Louvre, Met, Smithsonian, and WikiArt datasets, containing a diverse range of artworks. As shown in the results table, our model achieved an accuracy of 84% on the Smithsonian dataset, 79% on the Met dataset, and 76% on the Louvre dataset (see Figure 3). These results demonstrate the model's ability to accurately identify well-known artworks from prestigious museum collections, enabling the retrieval of detailed contextual information for the identified pieces.

7.2. Art Style Classification

For artworks that were not recognized by the identification CNN, we employed a separate CNN trained on the WikiArt dataset to classify the artwork's style or movement (e.g., Impressionism, Cubism, Surrealism). This classification allows our system to provide relevant information about the artistic style and its characteristics. Our style classification model achieved an accuracy of 61% on the WikiArt dataset (see Figure 3), which includes 27 different art styles.

7.3. Conversational Component Evaluation

To assess the effectiveness of our conversational component, we conducted two evaluations: naturalness and relevance. For naturalness, human evaluators rated the spoken responses generated by our system on a scale of 1 to 5, with 5 being the most natural. As shown in the results table, our system achieved an average naturalness score of 4.1 when using Apple's Speech library, 3.6 for React Native Voice, and 3.8 for ExpoSpeech (see Figure 4). Based on these results, we decided to stick with Apple's Speech library for the TTS component.

For relevance, evaluators rated the appropriateness and informativeness of the system's responses to spoken queries about artworks, again on a scale of 1 to 5. Our system achieved an average relevance score of 4.2 when using ChatGPT as the backend large language model, 4.6 with Gemini, and 3.6 with Dialogflow (see Figure 5). These scores indicate that our system can effectively understand and provide relevant information in response to users' queries, with Gemini being a clear frontrunner.

7.4. Ablation Study

To determine the impact of various components of our system, we conducted an ablation study by removing or modifying specific components and observing the change in performance. For instance, we evaluated the system's performance on pieces not in the museum datasets without the art style classification component, relying solely on artwork identification. In this scenario, when an artwork was not recognized by the identification CNN, the system could not provide any contextual information about the artistic style, techniques, or historical background. This limitation resulted in a 25% drop in the average relevance score, as

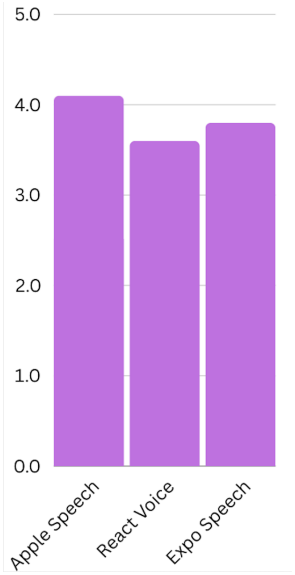

Figure 4. Average Naturalness Scores

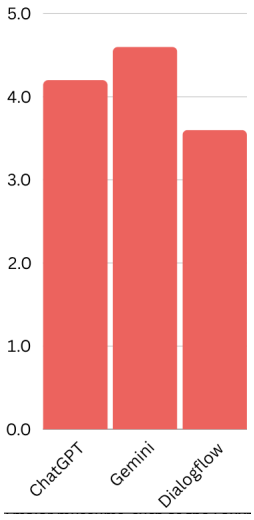

Figure 5. Average Relevance Scores

rated by human evaluators, from 4.6 to 3.5. This proved that the art style classification, while not highly performant, still provided an advantage to the curator.

Furthermore, we conducted an ablation study on the conversational component by removing the NLP and relying solely on pre-defined responses found in the JSON text bodies. Without the NLP component, users were limited to selecting from a fixed set of questions, and the system could not understand or respond to open-ended queries. This resulted in a 60% decrease in the average naturalness score, from 3.9 to 2.3, as the responses felt scripted and lacked personalization. The LLM component was definitely necessary to give the curator a more authentic human-like persona.

8. Conclusion

The current focus is on improving the art style classification accuracy by exploring larger and more specialized CNN models trained on diverse art style datasets. Additionally, efforts are underway to enhance the conversational component by integrating text-to-speech capabilities from ChatGPT 4o, allowing the agent to provide spoken responses with high levels of naturalness. We also plan to add a personalization feature to enable users to customize the voice of the agent for an enhanced experience. Future work includes expanding the dataset to include more museums and art collections, providing broader coverage and diversity of artworks.

Looking ahead, the potential applications of CuratorAI extends beyond that of museums and galleries. With little adaptation, this technology could be utilized for education purposes, providing students with an interactive and immersive learning tool. Additionally, AR integrations could seamlessly integrate the museum experience from the comforts of home.

The future of the museum experience is here. By integrating cutting-edge AI technologies, CuratorAI not only democratizes art education and enriches the museum experience, it can be easily lateral to other sectors providing more engaging and accurate information. As we continue to build CuratorAI, we look forward to transforming peoples appreciation for museums and art, ensuring that every visit is both memorable and educational.

9. References

Cho, Kyunghyun, et al. ”Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation.” arXiv preprint arXiv:1406.1078, 2014.

Gómez-Villa, A., et al. ”Automatic Artwork Description and Interpretation Using Deep Neural Networks.” Expert Systems with Applications, vol. 128, 2019, pp. 106-118.

He, Kaiming, et al. ”Deep Residual Learning for Image Recognition.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.

Kazemzadeh, Sahar, et al. ”ReferItGame: Referring to Objects in Photographs of Natural Scenes.” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 787-798.

Othman, N. K., et al. ”Multimodal Interaction in Museums: A Systematic Review of Multimodal Technologies and Applications.” Applied Sciences, vol. 11, no. 18, 2021, p. 8627.

Redmon, Joseph, and Ali Farhadi. ”YOLOv3: An Incremental Improvement.” arXiv preprint arXiv:1804.02767, 2018.

Simonyan, Karen, and Andrew Zisserman. ”Very Deep

Convolutional Networks for Large-Scale Image Recognition." arXiv preprint arXiv:1409.1556, 2014.

Sutskever, Ilya, et al. "Sequence to Sequence Learning with Neural Networks." Advances in Neural Information Processing Systems, vol. 27, 2014, pp. 3104-3112.

Szegedy, Christian, et al. "Rethinking the Inception Architecture for Computer Vision." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818-2826.

Tan, Mingxing, and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." International Conference on Machine Learning, 2019, pp. 6105-6114.

Tomi, A. B., and A. D. Bagdanov. "Galleries, Museums, and Archive Visitor Experiences Enhanced Through Cross-Disciplinary Culture Sampling." Journal on Computing and Cultural Heritage, vol. 9, no. 1, 2016, pp. 1-26.

Vaswani, Ashish, et al. "Attention is All You Need." Advances in Neural Information Processing Systems, vol. 30, 2017.