AI-Enhanced Art Appreciation: Generating Text from Artwork to Promote Inclusivity

Tanisha Shende

Oberlin College
tshende@oberlin.edu

Abstract

Visual art facilitates expression, communication, and connection, yet it remains inaccessible to those who are visually-impaired and those who lack the resources to understand the techniques and history of art. In this work, I propose the development of a generative AI model that generates a description and interpretation of a given artwork. Such research can make art more accessible, support art education, and improve the ability of AI to understand and translate between creative media. Development will begin with a formative study to assess the needs and preferences of blind and low vision people and art experts. Following the formative study, the basic approach is to train the model on a database of artworks and their accompanying descriptions, predict sentiments from extracted visual data, and generate a paragraph closely resembling training textual data and incorporating sentiment analysis. The model will then be evaluated quantitatively through metrics like METEOR and qualitatively through Turing tests in an iterative process.

Introduction

I am interested in generating a textual paragraph describing and interpreting artwork given an inputted image of the artwork. Generative artificial intelligence has risen in popularity and usage, and most applications are text-to-text, like ChatGPT, and text-to-image, like AI art generators. Image-to-text conversions have been studied to a lesser extent, but even these mostly deal with photographs. Artwork is inherently more subjective and open to interpretation due to its stylized forms, color palettes, and imagined compositions.

By studying the feasibility of feature extraction and semantic analysis of artworks, we can create more robust AI models that can still identify features regardless of how stylistic or distorted they are and can derive meaning from slight changes in appearance. This work can also help make art more accessible. It can be used in art education to complement human instructors and allow those unfamiliar with art to easily identify techniques, movements, etc. Additionally, the generated paragraphs can inform BLV users about the composition, texture, and colors of an artwork, so they are better equipped to visualize it and interpret it on their own.

Background

Prior research has looked into generating paragraphs from photographs (Krause et al. 2017; Liu et al. 2017; Huang et al. 2021). For example, Krause et al. decomposed images into regions and paragraphs into sentences in order to generate hierarchical semantic information for each region (2017). A limitation of this research is that it generates a series of facts about a realistic image; thus, there is limited opportunity for the subjective interpretations that are inherent to art.

There has since been research that seeks to introduce more variability and creativity to the multimodal exchange (Liu et al. 2018; Chen et al. 2023). Liu et al. generated poetry from images by extracting poetic clues associated with common sentiments from the images (2018). This work examines how sentiments associated with visual data can be extracted from images, and it experiments with creative output paired with straightforward visual data. As for converting creative visual input to creative textual output, Chen et al. captioned modern sentences from ancient Chinese paintings and used those to generate poetry (2023). My work will expand on this prior literature by using creative visual input rather than photographs and generating text that is both descriptive and interpretive.

Other related work is in the realm of creating assistive technologies for BLV people. Nair et al. developed a set of tools called ImageAssist that allows BLV users to explore images and their alt text given information on the art's important features (Nair, Zhu, and Smith 2023). Their rationale was that alt text is often low-quality and reduces disabled people's autonomy to interpret artwork themselves (Nair, Zhu, and Smith 2023). My work will expand on this by generating higher-quality paragraphs and ensuring BLV people retain their autonomy given their input.

Prior Work by the Applicant

My experience as a human-computer interaction researcher at Cornell Tech's Enhancing Ability Lab informed my approach to the inclusion of BLV participants and evaluators. I worked on a project that used haptic and audio feedback to make nonverbal cues, like smiling and nodding, more accessible to BLV users in social virtual reality, and this was an iterative process that required input from BLV participants throughout the study.

Approach

My goal is to develop a model that's accessible to BLV people while maintaining the quality recommended by art experts. Technology serving disabled people must be created with the input of the target audience to be effective and respectful. Additionally, given how subjective art can be, I will consult experts in art communication and art education to best understand the types of information to prioritize in the output.

To this end, I will first conduct a formative study with BLV people who regularly interact with art (and can thus better explain associated needs and challenges) as well as art experts. I will conduct semi-structured interviews to identify specific areas of need in art accessibility and understand how I can best target them without sacrificing autonomy or quality.

Given this information, I will train my model using images of artwork and their corresponding descriptive and analytical paragraphs from art history and museum databases. To respect the copyright of human artists, I will only consider artworks in the public domain. However, I will still try to use artwork from a variety of cultures, time periods, and styles to ensure the diversity and generalizability of the model. I will train a visual-semantic embedding (VSE) model, which maps visual data to textual data, on the artwork and their associated paragraphs in order to link the two modalities together.

For sentiment analysis, I will use a convolutional neural network (CNN), which processes and analyzes visual data to extract features and represent them hierarchically. I will train the CNN on a dataset of visual data with common accompanying sentiments, so it can predict the sentiments associated with visual data from a new dataset of input artworks.

For paragraph generation, I will use a generative adversarial network (GAN), which is good for creating text that closely resembles real art descriptions and analyses. I will need to ensure that the GAN incorporates sentiment into its output.

In the end, the model should accept an image of an artwork and output a paragraph of comprehensible language and high-quality art criticism.

Evaluation

I will evaluate my model using both quantitative and qualitative assessments. I will use metrics, like BLEU, ROUGE, METEOR, or SPICE, to assess the quality of the generated paragraphs in comparison to the existing descriptions written by art experts. My ideal metric would be one that correlates with human judgment and accounts for synonyms, acronyms, and varying sentence structures.

Qualitatively, I will reintroduce the BLV people and art experts from the formative study as co-designers and ask them to complete surveys and interviews to assess the model's ability to provide helpful and meaningful descriptions. This will be an iterative process. I will also conduct a Turing test with new participants, some of whom will be art experts, to see if they can differentiate between real paragraphs and generated paragraphs.

Discussion

Through the formative and evaluative qualitative studies, I hope to gain a better understanding of the needs of BLV people in art engagement and the recommendations of art experts in promoting art education. If my approach works, then BLV people will have greater access to artwork, and art as a whole will be easier to teach and categorize by theme. Regarding the field of AI, this work will improve multimodal learning, especially since artworks contain more ambiguous and stylized features than photographs. It will also demonstrate how disabled people respond to AI-generated content, which can impact the field of human-computer interaction and encourage the consideration of diverse needs in AI research.

Conclusion

Artwork-to-text conversions using generative AI have received limited attention but have the potential to strengthen AI models and their multimodal learning and to improve art accessibility. Based on input from BLV people and art experts, I will create a model that takes an image of an artwork, extracts and analyzes its features and sentiments, and generates a text paragraph describing and analyzing the artwork based on the model's findings. I will quantitatively evaluate the model through metrics like METEOR and SPICE and a Turing test, and I will qualitatively assess the model's usability and quality through feedback from BLV people and art experts.

Ethical Statement

At its best, this work can make art more accessible to a wider audience, such as BLV people and people lacking educational resources. However, the generated text may replicate existing social inequities. The inclusion of BLV people in the study is meant to create a product as suitable to their needs as possible, but it cannot encompass the full spectrum of the disabled experience. There may be problems that we did not fully anticipate, consider, or solve. Thus, it remains possible that the model may still limit or threaten the agency of BLV people, in which case it will be refined with more user input.

Furthermore, since the model is trained on data from museum curators and art historians, it may reproduce cultural biases from those fields. For example, it may echo implicit paternalism, orientalism, and racism held towards artworks from historically-marginalized cultures.

Finally, while the model can supplement art knowledge and demonstrate how to describe and interpret artworks, an overreliance on AI-generated interpretations may cause people to devalue critical thinking skills and alternative interpretations.

Acknowledgments

I thank Dr. Shiri Azenkot, Jazmin Collins, and Danielle Montour of The XR Access Initiative and Cornell Tech's Enhancing Ability Lab. Through unrelated research, they've taught me how to conduct studies and design accessible technologies for disabled people that incorporate their

voices, experiences, and preferences. Our collaboration was made possible by XR Access' Research Experiences for Undergraduates (REU) site, which was supported by the National Science Foundation under Grants 2051053 and 2051060.

References

Chen, J.; Huang, K.; Zhu, X.; Qiu, X.; Wang, H.; and Qin, X. 2023. Poetry4painting: Diversified poetry generation for large-size ancient paintings based on data augmentation. *Computers & Graphics*, 116: 206–215.

Huang, Y.; Liu, B.; Fu, J.; and Lu, Y. 2021. A Picture is Worth a Thousand Words: A Unified System for Diverse Captions and Rich Images Generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, 2792–2794. New York, NY, USA: Association for Computing Machinery. ISBN 9781450386517.

Krause, J.; Johnson, J.; Krishna, R.; and Fei-Fei, L. 2017. A Hierarchical Approach for Generating Descriptive Image Paragraphs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3337–3345.

Liu, B.; Fu, J.; Kato, M. P.; and Yoshikawa, M. 2018. Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, 783–791. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356657.

Liu, Y.; Fu, J.; Mei, T.; and Chen, C. W. 2017. Let Your Photos Talk: Generating Narrative Paragraph for Photo Stream via Bidirectional Attention Recurrent Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Nair, V.; Zhu, H. H.; and Smith, B. A. 2023. ImageAssist: Tools for Enhancing Touchscreen-Based Image Exploration Systems for Blind and Low Vision Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.