IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 6, Issue 1, April 2026

Impact Factor: 8.2

Emotion-Conditioned Image Captioning for Visual Artworks Using Affective Visual Encoders

Gangumalla Neha Reddy[1], Akula Srilaya[2], Annamaneni Sanjana[3], Malathi B[4]

Department of Computer Science and Engineering,
Sreenidhi Institute of Science and Technology, Hyderabad, India

Abstract: *Image captioning has experienced significant advancements with the introduction of the deep learning methods, but the majority of the existing systems are actively concerned only with the objective visual description, but not the emotional context of images. Painting and illustration are examples of visual arts that have strong affective meanings that are not well reflected in the traditional models of captioning. The study offers an emotion-conditioned image captioning model that combines the affective visual encoders and the deep learning-based captioning generation. The system simultaneously extracts visual representations and emotional representations of artwork images, and conditions the caption generation process according to the emotions detected. The experimental findings indicate that emotion-sensitive captions are more detailed and expressive than the conventional image captioning methods.* .

Keywords*:* Image Captioning, Emotion Recognition, Affective Computing, CNN, LSTM, Visual Artworks

I. INTRODUCTION

Image captioning is a challenging task at the intersection of computer vision and natural language processing. It aims to generate semantic descriptions of pictures through interpreting visual data and converting them into natural language. Although the combinations obtained by deep learning models are astonishingly successful on natural pictures, they sometimes fail to reflect the emotional and aesthetic nature of visual pieces of art.

The state-of-the-art image captioning systems are majorly based on convolutional neural networks (CNNs) to extract features and recurrent neural networks (RNNs) or transformers to produce captions. These systems are trained on large scale datasets of real-world images that have objective labels, and hence are inappropriate to detect emotional cues that exist in artworks. Affective computing is a new term that has come up to serve the purpose of identifying and modeling human emotions through computational methods. Previous researches have indicated that the visual elements like color distribution, texture and composition are important elements in expression of emotion in art.

The aim of the current study is to create a framework of captioning which will explicitly model the emotional information as well as the visual features. The system is also conditioned to generate captions based on emotion vectors in an attempt to get captions that are not mere descriptions but also expressive of emotions.

II. EXISTING SYSTEM

The traditional image captioning systems primarily focus on generating captions based only on the visual content present in an image. These systems typically use convolutional neural networks (CNNs) for feature extraction and recurrent neural networks (RNNs) or transformers for caption generation. While they perform well on natural images, they mainly produce objective descriptions such as identifying objects, scenes, and spatial relationships. However, these models do not capture the emotional or artistic intent present in visual artworks such as paintings, sketches, and illustrations. As a result, the generated captions often lack expressiveness and fail to reflect the mood or emotional tone conveyed by the artwork.

DOI: 10.48175/568

III. PROPOSED SYSTEM

Emotion-Conditioned Image Captioning framework integrates affective visual encoders along with traditional visual feature extraction techniques. The system extracts both semantic visual features and emotional representations from artwork images and combines them using an emotion-conditioning layer. These integrated features guide the caption generator to produce descriptions that not only explain the visual content but also reflect the emotional context of the artwork. By incorporating affective computing techniques, the proposed system generates captions that are more expressive, context-aware, and aligned with the artistic mood, thereby improving the interpretability and user engagement in applications such as digital galleries, art analysis, and creative AI systems.

IV. METHODOLOGY

The suggested methodology is based on a multimodal pipeline that is based on deep learning. The image of artwork is initially subjected to two parallel convolutional neural networks one that is used to extract visual features and the other to extract affective features associated with emotions.

The affective visual encoder acts as an emotion encoder that generates an emotion vector of the emotional states that can be happy, sad, calm, or tense. The emotion feature is integrated with the visual features using an emotion-conditioning layer.

A caption generator that consists of an LSTM is used to generate emotion-aware word-by-word captions based on the integrated representation. The model is trained using supervised learning and paired artwork-emotion-caption data.

Measurements of evaluation are the BLEU, METEOR and qualitative human judgment.

V. SYSTEM ARCHITECTURE

This system architecture can be divided into five main elements, namely, artwork image input, CNN-based affective visual encoder, CNN image encoder, emotion conditioning layer, and LSTM-based caption generator.

The image of the artwork is decoded by both the visual encoder and affective encoder. Affective encoder gives an emotion vector whereas visual encoder gives semantic image features. The outputs are combined in the layer of emotion conditioning which matches emotional context with visual semantics. The conditioned features are then transferred over to the caption generator to generate an emotion-aware textual description.

Artwork Image Input Module

The Artwork Image Input module is the interface of the system and receives visual art like painting, sketch, or other digital illustrations. The images that are inputted are downsized and scaled to fit the needs of the deep learning encoders. This module guarantees uniformity of images quality and resolution prior to additional processing. It serves as an interface between raw artwork data and feature extraction layers. Handling of images at this stage correctly enhances downstream learning.

CNN-Based Affective Visual Encoder

The Affective Visual Encoder that is based on CNN is in charge of registering emotional indications in the artwork. It removes emotions features related to color arrangement, texture, contrast, and painting patterns that play part in emotional perception. These features are translated into an emotion vector known as affective states generated by the encoder. This emotion-vector provides high level of emotional context of the artwork. The module allows the system to learn subjective emotional attributes other than object identification.

CNN Image Encoder

The CNN Image Encoder is concerned with the process of deriving semantic visual details of the piece of art. It recognizes objects, shapes, spatial relationship and structural patterns in the image. These characteristics create a high-dimensional image of the content of the art. The encoder is usually trained with off-the-shelf convolutional architectures trained on art collections. This module provides proper visual perception to meaningful caption generation.

DOI: 10.48175/568

Emotion Conditioning Layer

The Emotion Conditioning Layer is a layer that mixes the emotion vector of the affective encoder and the visual features of the image encoder. It is this combination of emotional context and semantic image that is involved in this fusion. The conditioned representation guarantees the presence of visual material as well as emotional tone to the caption generation process. At this point the system avoids the production of descriptions that are emotionally neutral by incorporating affective signals. This layer is important in emotion-aware caption generation.

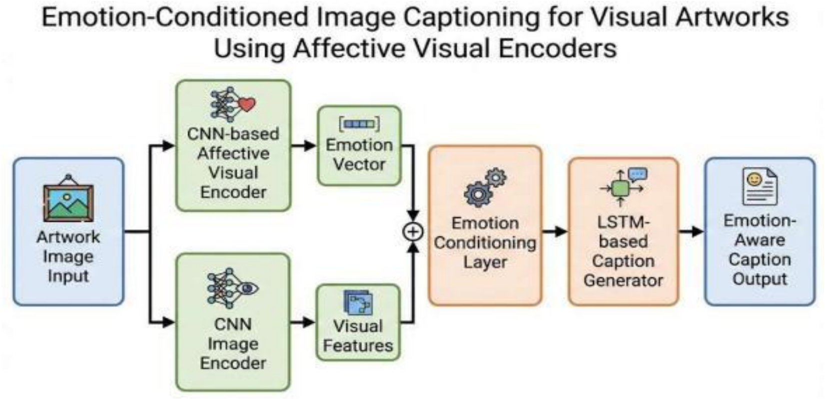

Emotion-Conditioned Image Captioning for Visual Artworks
Using Affective Visual Encoders
Artwork Image Input
CNN-based Affective Visual Encoder
Emotion Vector
CNN Image Encoder
Visual Features
⊕
Emotion Conditioning Layer
LSTM-based Caption Generator
Emotion-Aware Caption Output

Fig. 1. System Architecture

LSTM-based Caption Generator

The Caption Generator processes based on the emotion-conditioned features is a LSTM-based Caption Generator that generates natural language description. It creates captions in a succession so that there is linguistic coherence and emotional relevance. The generator gets to know the emotional context in connection to the descriptive words and structures of sentences. The end result is the emotion-aware caption which mirrors both the image content and emotional tone of the piece of art. The module allows generating expressive and context-aware captions that can be interpreted by the human eye.

VI. IMPLEMENTATION

The system is implemented in Python using deep learning frameworks such as TensorFlow and PyTorch. It involves fine-tuning pretrained CNN architectures like ResNet for visual and affective feature extraction.

One-hot emotion representations or continuous emotion representations are used to encode emotion labels. Teacher forcing is used in training the LSTM-based caption generator to enhance convergence. Unseen artwork images are used to test the trained model to determine the quality of captions and emotional relevance.

The implementation provides modularity and can be easily extended to transformer-based captioning models.

VII. RESULTS

The proposed model achieved higher BLEU and METEOR scores compared to baseline image captioning models. The outcome of human evaluation has shown that emotion-aware captions were more interesting and matched the perceived emotional content of the works of art.

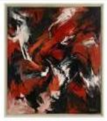
Artwork	Baseline Caption	Proposed Emotion-Aware Caption
	A painting of a lake and mountains.	A peaceful landscape evoking a sense of calm and serenity.
	An abstract painting with red and black colors.	A chaotic and turbulent scene conveying intense tension and energy.

Fig. 2. Caption Comparison

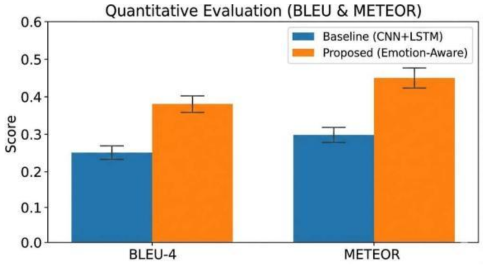


Fig. 3. Quantitative Results

VIII. CONCLUSION

The study introduces an emotion-conditioned image captioning system which combines deep learning-based caption generation with affective visual encoders. The system generates context-sensitive and emotional captions of visual pieces of art by means of combining emotional and visual characteristics. The findings indicate the significance of affective information in image captioning and open up new avenues of emotion-sensitive AI-systems.

IX. FUTURE SCOPE

Future directions of this work can involve the development of transformer-based architectures to achieve a better caption fluency and emotion alignment. Affective understanding can also be improved with the help of multimodal emotion learning on the text and audio level. The system may also be extended towards real-time applications in digital museums, virtual galleries and creative AI-tools.

REFERENCES

[1]. K. Xu, J. Ba, R. Kiros, et al., “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,” Proc. 32nd Int. Conf. Machine Learning (ICML), pp. 2048–2057, 2015.

[2]. A Vinyals, O. Toshev, S. Bengio, and D. Erhan, “Show and Tell: A Neural Image Caption Generator,” Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 3156–3164, 2015.

[3]. R. W. Picard, Affective Computing, MIT Press, Cambridge, MA, USA, 1997.

[4]. J. Machajdik and A. Hanbury, “Affective Image Classification Using Features Inspired by Psychology and Art Theory,” Proc. ACM Int. Conf. Multimedia, pp. 83– 92, 2010.

[5]. X. Wang, Y. Chen, and J. Zhang, “Emotion- Aware Image Captioning,” IEEE Access, vol. 8, pp. 178303–178314, 2020.

[6]. Y. Li, J. Pan, and Z. Zhang, “Visual Sentiment Analysis for Artistic Images,” Pattern Recognition, vol. 83, pp. 37–48, 2018.

[7]. S. You, K. Y. Han, and H. Xu, “Image Captioning with Semantic Attention,” Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 4651–4659, 2016.

DOI: 10.48175/568

[8]. A Graves, A. Mohamed, and G. Hinton, “Speech Recognition with Deep Recurrent Neural Networks,” Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), pp. 6645–6649, 2013.

[9]. J. Deng, W. Dong, R. Socher, et al., “ImageNet: A Large-Scale Hierarchical Image Database,” Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 248–255, 2009.

[10]. T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 785–794, 2016.

[11]. Y. Li, N. Wang, and J. Liu, “Emotion- Guided Image Captioning via Cross-Modal Learning,” IEEE Transactions on Multimedia, vol. 23, pp. 3361–3373, 2021.

[12]. A Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” Advances in Neural Information Processing Systems (NeurIPS), pp. 1097–1105, 2012.