IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

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

Impact Factor: 7.53

Volume 4, Issue 2, November 2024

Smart Caption: Intelligent Image Description Using Transformer Decoder

Ms. Deepali Govindrao Navalkar[1] and Prof. (Dr.) N. R. Wankhade[2]

Student, Computer Engineering, Late G. N.Sapkal College of Engineering, Nashik, India[1]

Head of Department, Computer Engineering, Late G. N.Sapkal College of Engineering, Nashik, India [2]

Abstract: *The need for automated image description systems has grown significantly with the rise of multimedia content across diverse domains such as social media, digital libraries, and accessibility technologies. Smart Caption presents a novel approach for generating intelligent and context-aware image descriptions by utilizing a combination of Vision Transformers (ViTs) and Natural Language Processing (NLP) Transformer decoders. Unlike traditional convolution-based methods, Vision Transformers treat images as sequences of patches, enabling them to capture global image features more effectively. These extracted visual features are then processed by a Transformer-based decoder to produce coherent and contextually appropriate captions, bridging the gap between image recognition and natural language generation. Our approach focuses on leveraging the attention mechanisms inherent in both Vision and NLP Transformers to enhance the quality of image descriptions. The ViT architecture is designed to focus on relevant regions within an image, while the NLP Transformer decoder is adept at generating fluent and detailed descriptions by attending to the most significant visual features. This two-stage process improves the precision of object detection, relationship understanding, and scene context in generated captions. Experimental results demonstrate that Smart Caption out performs traditional methods in terms of accuracy, contextual relevance, and fluency, marking a significant step forward in the field of automated image captioning. Through this research, we aim to provide a scalable, efficient, and intelligent solution for image description, with potential applications in areas such as accessibility tools for visually impaired users, digital content indexing, and automated content generation. Our findings highlight the strengths of transformer-based models in bridging vision and language tasks, offering promising directions for future research in multimodal AI..*

Keywords: Vision Transformers, NLP Transformers, Image Captioning, Transformer Decoder, Multimodal Learning, Deep Learning, Natural Language Processing, Visual Feature Extraction, Intelligent Image Description

DOI: 10.48175/IJARSCT-22130