Chapter 11
Image Captioning Made Easy:
Leveraging Vision Transformers and GPT-2 to Create Accurate and Coherent Descriptions From Images

Ayesha Taranum
http://orcid.org/0000-0002-3171-6656
Vidyavardhaka College of Engineering, India

Mohammed Ezhan
http://orcid.org/0009-0009-3194-4841
Northeastern University, USA

ABSTRACT

Image captioning, which is the generation of descriptive word text summaries from image content, has drawn considerable interest in computer vision and natural language processing (NLP). This research proposes a Python application that combines Vision Transformers (ViT) and GPT-2 for automatic image captioning. The system employs a pre-trained NLP connect/vit-gpt2-image-captioning model from Hugging Face, coupled with a graphical user interface (GUI) designed using Tkinter. The model efficiently extracts features from images and produces coherent, contextually appropriate captions, showing improvement over conventional Convolutional Neural Network-Long Short Term Memory(CNN-LSTM) based models. This study emphasises the architecture, methodology, and comparison of the system, highlighting

DOI: 10.4018/979-8-3373-4202-3.ch011

its applicability in real-world applications such as visually impaired accessibility, content management, and image retrieval. Performance measurement suggests the model's capacity to produce high-quality captions in an efficient manner.

INTRODUCTION

The capability to come up with relevant and contextually correct captions for images has also become a basic function in computer vision and natural language processing (NLP). The automatic creation of a descriptive textual summary of the content in an image is known as image captioning, and it has numerous real-world applications, such as accessibility for the blind, content management, social media platforms, and image search engines. The capacity to caption an image in natural language closes the gap between visual data and text information, facilitating easier search, classification, and comprehension of visual content. This project presents a Python application that is meant to incorporate image captioning functionality through a pre-trained Vision Encoder-Decoder model. Developed with the Tkinter library for the graphical user interface (GUI) and using the highly effective Vision Transformer (ViT) model coupled with GPT-2, this tool allows users to automatically create captions for uploaded pictures in a seamless and interactive process. The tool makes use of Hugging Face's pre-trained models, the `NLP connect/vit-gpt2-image-captioning`, to transform images and create precise, contextually correct captions from their content.

The system is built to be compatible with both CPU and GPU, so it is flexible to support varying configurations of hardware. It offers a simple GUI for ease of use, where one can easily upload an image, see its preview, and create a caption with minimal setup. The process is completely automated from loading the model to processing the image and creating the caption, while also offering real-time status updates to notify the user of the process in progress. Earlier methods of image captioning were based on predefined templates or object recognition in the explicit sense. But with the progress in deep learning, especially the introduction of encoder-decoder models, it became feasible to produce well-structured and contextually correct descriptions that are more than just recognising objects. The application utilises state-of-the-art deep learning models in the form of Vision Transformer (ViT) for understanding content from images and GPT-2 for generating natural language so that it can produce rich, detailed captions from visual input.

This project chronicles the creation of the image captioning system, outlining its major components, such as model integration, user interface design, and error handling mechanisms. The report also touches on the technology stack and how such an image captioning tool might have applications in real-world use. The technology