IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CAPTION GENERATION FROM IMAGES AND VIDEOS TO AID PATIENTS WITH VISUAL AGNOSIA

[1] Chandan Kumar S, [2]Ujwal T R, [3]Sudiptha J, [4]Dr. Leena Giri G

1, 2 & 3 Students, Dept of Computer Science and Engineering,
Dr Ambedkar Institute of Technology, Bangalore, Karnataka, India

4 Associate Professor, Dept of Computer Science and Engineering,
Dr Ambedkar Institute of Technology, Bangalore, Karnataka, India

Abstract: This project represents an initiative to enhance accessibility and inclusivity for individuals grappling with medical conditions like visual agnosia, a neurological condition characterized by difficulties in recognizing and interpreting visual information. The technical foundation is built upon a sophisticated two-stage architecture. Firstly, the Image Encoder leverages the CLIP encoder, to extract high-level features from images. These features serve as a rich representation of the visual content and are subsequently passed to RNN. The RNN employs LSTM network well-suited for sequential data processing. The LSTM is responsible for decoding the extracted image features into coherent and descriptive textual captions. Furthermore, an integral part of this project is the integration of the gTTS (Google Text-toSpeech) library, which introduces text-to-speech capabilities. This in addition lets the transformation of retrieved textual captions into spoken words, thereby creating a comprehensive and accessible experience for individuals with visual agnosia. The deployment of gTTS not only facilitates generating of audio descriptions but also enables users to customize speech speed, language preferences, and output formats. The system's overarching objective is to provide individuals with visual agnosia a robust and adaptable toolset for interpreting visual content. By combining image feature extraction with sequence generation and auditory synthesis, this project aims to bridge the gap in understanding visual stimuli, empowering users with detailed textual descriptions and spoken narratives. The intricate interplay of deeplearning method, neural network architectures, and library integrations underscores the project's technical complexity and potential impact on lives of individuals facing challenges in visual recognition.

Keywords **- CLIP, CNN,NLP**

I. INTRODUCTION

Visual agnosia is a neurological disorder that impairs an individual's ability to recognize and interpret visual stimuli despite having intact vision. This illness frequently arises from damage to the visual processing areas in the brain, leading to difficulties in identifying objects or even familiar scenes. One innovative approach to address visual agnosia involves analyzing visual content and generate descriptive captions, providing individuals with visual agnosia valuable contextual an explanation of the images and videos they perceive. This aids in enhancing their comprehension and recognition of visual stimuli, ultimately improving their overall visual understanding.

Generating description of an image is a task to comprehend the scene - a primary goal of the project. In addition to being able to overcome computer vision difficulties in recognizing the items in an image, caption model must be able to capture and convey the relationships between those objects in a natural-language. This is the reason that caption generation has been considered a difficult problem. Since it essentially amounts to imitating the extraordinary human capacity to condense enormous volumes of crucial visual information into descriptive language, it's a crucial problem for algorithms used in machine learning. The utilization of CNNs and RNNs as the foundation of our strategy, given their proficiency in extracting hierarchical features from images. By using the strength of deep learning, we aim to prepare a robust model capable of discerning objects, relationships, and sentiments depicted in images. The output captions are envisioned to transcend mere annotations, encapsulating the essence of visual content with human-like fluency and coherence.

The traditional RNN is insufficient for encoding long-term dependent information in video sequences. For long sequential data, long term reliance remains an unresolved issue, despite the fact that LSTM partially addresses it. To get over this restriction, we separate the input video into separate frames and produce the final caption while preserving the context between each frame. Text to speech is then used to turn the output captions into speech.

II. LITERATURE SURVEY

The automatic production of captions for photographs is a critical task in computer vision and scene interpretation, representing the nexus of natural language expression and visual perception. This hard task requires models that are not just good at identifying things in photos, but also good at explaining their relationships in a logical, human-like manner. This problem is critical to machine learning algorithms because it entails simulating the human capacity to communicate large amounts of visual information in a concise manner using descriptive language. The image caption generation problem has attracted a lot of research interest recently, partly due to the availability of huge classification datasets and advances in neural network training. With two variants examined in this paper—a "hard" stochastic attention mechanism trainable through maximizing a variational lower bound, or REINFORCE and a "soft" deterministic attention mechanism trainable through conventional back-propagation methods—attention mechanisms have emerged as crucial components in this framework. Incorporating attention improves the model's performance and makes it possible to see where the model is focused when creating captions.

The primary objective of video understanding is to automatically describe the contents of a video in natural language. Unlike image captioning, which deals with static images, video understanding poses a more formidable task because of the intricate nature of information within dynamic sequences. The challenge lies in capturing not only the spatial contents, such as objects and scenes, but also the temporal dynamics, encompassing actions, context, and flow. This paper addresses the limitations of traditional video captioning models and their inadequate performance in understanding video context. Conventional deep learning-based video captioning models typically employ gated recurrent neural networks (RNN), such as LSTM or gated recurrent unit (GRU), within an encoder-decoder architecture. Despite their utility, these models face challenges such as Inability to generate natural captions for long videos with diverse and complex events and will also lead to insufficient memory and Lack of robust context understanding. While LSTM and GRU partially address the long-term dependency issue, it remains unsolved for lengthy sequential data.

III. METHODOLOGY

Caption generation from photos and videos requires combining CNN and NLP algorithms. The process for creating captions from photos and videos includes:

Step 1. Data Collection and Pre-processing:
Gather a large dataset of images with corresponding captions. Datasets MS-COCO is used for this purpose. Pre-process the images (resize, normalize pixel values) and tokenize the captions.

Step 2. Feature Extraction:
Convolutional Neural Networks (CNNs) like CLIP are used to extract features from the pictures. In order to obtain a feature vector, remove the last layer of categorization. The image is represented more succinctly by this feature vector.

Step 3. Sequence Generation:
Using Recurrent Neural Network (RNN) or transformer-based model to produce captions by using the image features. Alternatively, we use CLIP that is capable of understanding both images and text, making it suitable for cross-modal tasks.

Step 4.NLP:
Rewrite the output word tokens such that the captions are legible by humans. Apply any further post-processing to improve the captions' readability.

Step 5 Text to speech:
Create an audio version of the output text by using the gTTS text-to-speech library. By connecting to Google Translate's API, the Python gTTS (Google Text-to Speech) package makes text-to-speech implementation simpler. By defining the text and language, creating a gTTS object, and saving the result as an audio file (often MP3), it allows for rapid text-to-speech conversion. This Applications requiring voice synthesis will find the compact library to be a convenient option because it allows for control over speech speed, supports many languages, and enables flexibility in output formats.

A. SYSTEM DESIGN FOR IMAGE PROCESSING:

The proposed project aims to create a system for automatic caption generation from images using CNNs and RNNs. This project is envisioned as a significant step forward in enhancing the synergy between computer vision and natural language processing, paving the way for advancements in image understanding and interpretation.

PyTorch is the framework used by the picture captioning system during training. Captions and pre-processed photos are loaded using MS-COCO. An LSTM-based caption generator and CLIP for image feature extraction make up the model architecture. An optimizer is used in the training loop to reduce loss. The loss determines which model should be saved. In addition to evaluating learning curves and assessment metrics, the system displays model performance using random samples taken from the validation set. PyTorch, CLIP, and RNN are components of the tech stack that are used for deep learning, image processing, and text metrics, respectively.

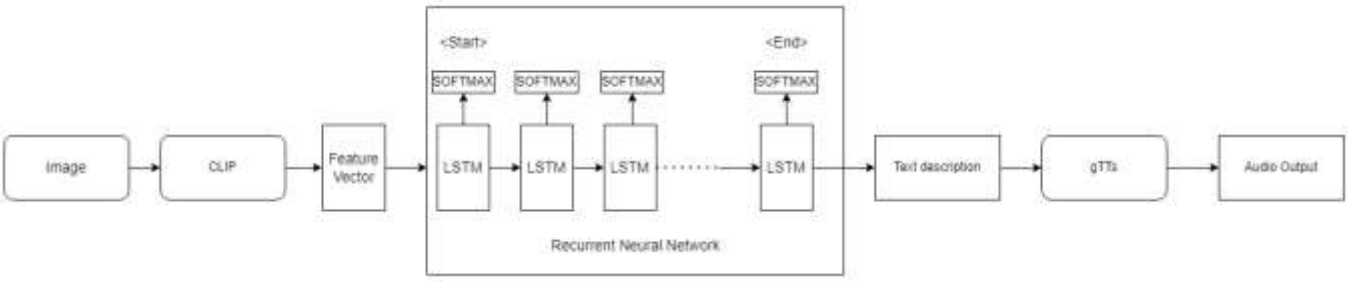

FIG 1 Image Processing System Architecture

B. SYSTEM DESIGN FOR VIDEO PROCESSING:

The traditional RNN is insufficient for encoding long-term dependent information in video sequences. For long sequential data, long term reliance remains an unresolved issue, despite the fact that LSTM partially addresses it. To ensure that get over this restriction, we separate the input video into separate frames and produce the final caption while preserving the context between each frame. Text to speech is then used to turn the output text into speech.

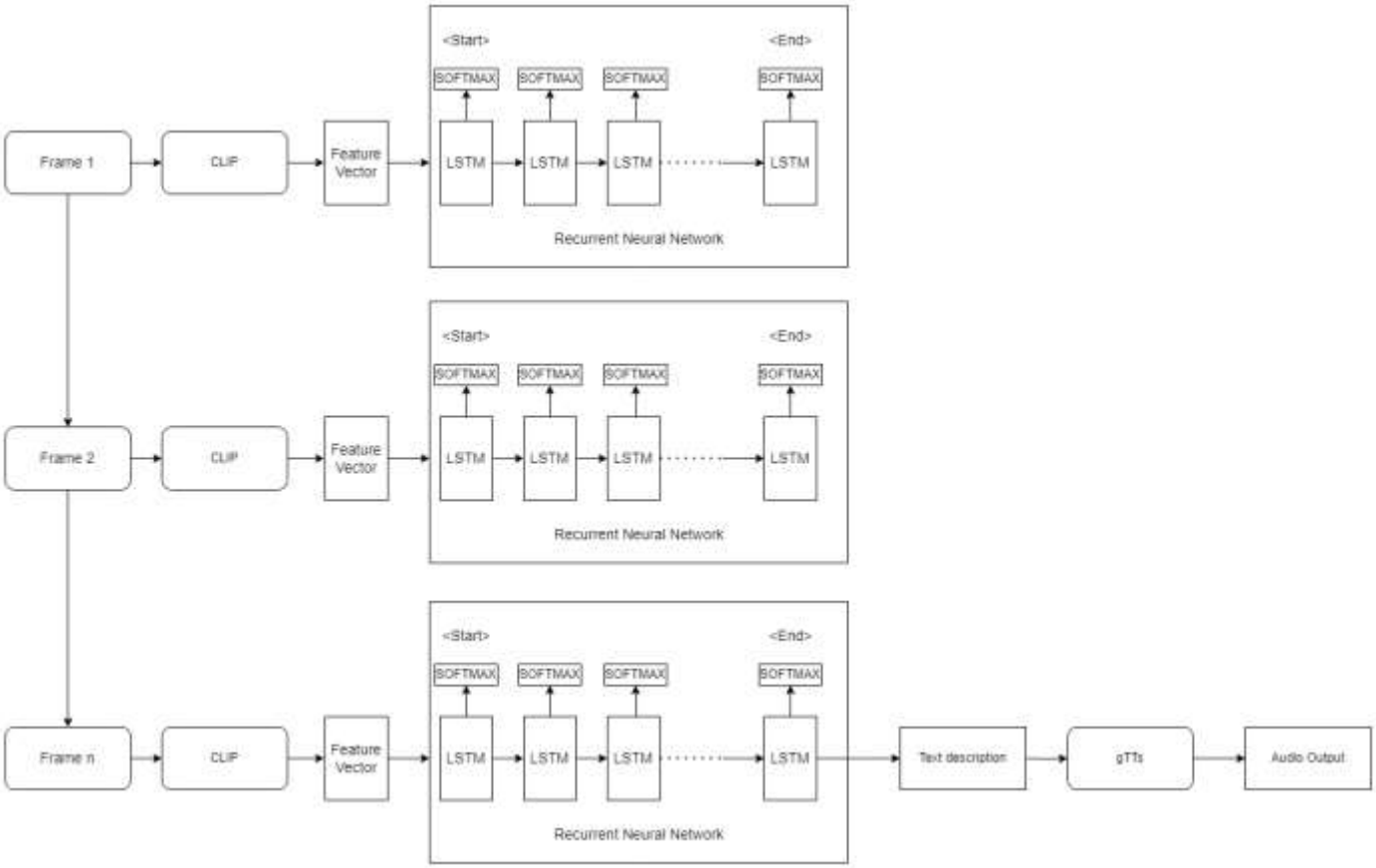

FIG 2 Video Processing System Architecture

C. FLOWCHART

Data Gathering and Preprocessing: Compile a substantial image library with relevant captions.
Resizing photos, standardizing pixel values, and tokenizing captions are examples of preprocessing data.

Feature Extraction:
Using a CNN model, like CLIP, to extract high-level characteristics of the images. These features capture the essential visual content of the image.

Caption Generation:
Employ a recurrent neural network (RNN) or a transformer-based model to generate captions according to the extracted image features. These models are adept at processing sequential data like sentences and can translate the visual features into coherent language descriptions.

Text-to-Speech:
Integrate a text-to-speech library like gTTS to convert the output captions into spoken audio. This step is especially beneficial for those with visual agnosia, as it allows them to use the information through audio instead of relying on text.

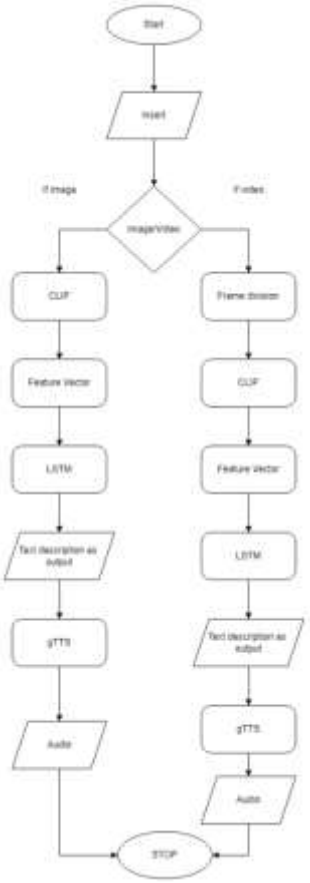

FIG 3 Work flow diagram

IV. INTERFACE

Gradio is a powerful tool that simplifies the creation of interactive interfaces for machine learning models. In our project, Gradio

plays a crucial part in crafting a user-friendly interface. The back end and front end is both managed by gradio, as it gives ready-

made components to deploy the trained model.

FIG 4 Interface for creating captions for images

FIG 5 Interface for creating captions for videos

V. RESULTS AND DISCUSSION

In summary, a significant advancement in the field of assistive technology for those with visual impairments has been made with the development of a caption generation system specifically designed for individuals with visual agnosia. This project has demonstrated the ability to automatically generate descriptive captions from images and videos by utilizing computer vision algorithms and natural language processing techniques. This can help individuals with impaired visual recognition bridge the comprehension gap between visual content and visual content. When such a system is successfully implemented, it not only helps individuals with visual agnosia comprehend and interpret visual cues better, but it also encourages their independence and autonomy in daily tasks. Research into and improvement of caption generating systems show significant promise in improving the inclusivity and accessibility of information for people with various visual impairments as technology develops, ultimately leading to a better quality of life and well-being.

REFERENCES

[1] Kelvin Xu et al (2016) "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." arxiv.org April 2016.

[2] Jonghong Kim, Inchul Choi and Minho Lee, (2020) "Context Aware Video - Caption-Generation with Consecutive Differentiable Neural Computer." mdpi 2020.

[3] Karen Simonyan, Andrew Zisserman (2015) "Very Deep Convolutional Networks for Large-Scale Image Recognition" International Conference on Learning Representations (ICLR), 2015.

[4] Zhe Gan, Chuang Gan,Yunchen Pu, Kenneth Tran, Lawrence Carin, Li Deng, "Semantic - Compositional Networks for Visual Captioning," ArXiv, 2017.

[5] S. Li, Z. Tao, K. Li and Y. Fu, "Visual to Text: Survey of Image and Video Captioning," vol. 3, no. 4, pp. 297- 312, Aug. 2019.

[6] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga and M. Bennamoun, "Text to Image Synthesis for Improved Image Captioning," in IEEE Access, vol. 9, pp. 64918-64928, 2021.

[7] Tiwary T, Mahapatra RP (2023) An accurate generation of image captions for blind people using extended convolutional atom neural network. Multimed Tools Appl 82:3801–3830.

[8] Chu Y, Yue X, Lei Y, Sergei M, Wang Z (2020) Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention. Wirel Commun Mob Comput 2020:8909458–7.

[9] Feng J, Zhao J (2022) Context-fused guidance for image captioning using sequence-level training. Comput Intell Neuroscie 9743123:9.