IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ENHANCING TRAFFIC SCENE AND UNDERSTANDING THROUGH IMAGE CAPTIONING AND AUDIO

[1]Sejal Pawar, [2]Shruti Mulay, [3]Jivani Suryawanshi, [4]Vaishnavi Walgude, [5]Prof. K. V. Patil

[1]Student, [2] Student, [3] Student, [4] Student, [5] Professor

[1]Department of Information Technology,

[1]Bharati Vidyapeeth's College Of Engineering for Women, Pune, India

Abstract: Understanding what's happening in traffic scenes is crucial for systems that help control traffic or drive autonomous vehicles. The project focuses on developing an automatic captioning system for traffic images. To do this, we use modern techniques like spotting pedestrians, cars, and recognizing traffic signs. Leveraging computer vision and natural language processing techniques, our system generates descriptive text based on image content and also provides audio. We enhance accessibility by converting these captions into natural-sounding audio descriptions, benefiting visually impaired users. It does this by using special types of computer programs called Convolutional Neural Networks (CNN), You Only Look Once (YOLO) and Long Short-Term Memory (LSTM) models. Our project aims to automate traffic image annotation, bridging the gap between visual content and textual and audio descriptions.

Index Terms – Convolutional Neural Networks (CNN), You Only Look Once (YOLO), Long Short-Term Memory (LSTM).

I. INTRODUCTION

The advancement of image captioning methods holds promise for enhancing the interpretability of autonomous vehicles, contributing to the development of Smart Transportation Systems that address the multifaceted challenges of modern society, encompassing socio-economic, environmental, and safety considerations. These methods, focusing on specific tasks like traffic sign recognition, vector detection, and pedestrian detection, represent a significant stride towards addressing the complexities of the transportation landscape. While existing research has primarily concentrated on single or multiple target detection, the current endeavor aims to automatically generate captions for images based on their content. Achieving this necessitates a comprehensive understanding of image content followed by the articulation of semantic information in a grammatically correct manner, demanding the integration of computer vision and natural language processing techniques.

This paper endeavours to produce automated descriptions by learning image contents, supplemented by a mechanism for converting generated captions into synthesized voice using a Text-to-Speech (TTS) engine. This system holds potential for enhancing Transportation Systems as well as facilitating applications tailored for individuals with visual impairments. Leveraging an encoder-decoder model for caption generation and employing text-to-speech technology, this work utilizes the Flicker 30K dataset to drive advancements in descriptive image interpretation and accessibility.

II. Literature Survey

In [1] The Zhaowei QU study, as mentioned by the authors, introduced an image captioning model focused on accurately recognizing various types of traffic objects and comprehending traffic situations by leveraging all available information. Instead of relying solely on keywords, they employed Long Short Term Memory (LSTM) networks to generate natural language suggestions or driving operation strategies. This approach aimed to provide more nuanced and contextually rich descriptions, enhancing the capability of autonomous systems to understand and navigate complex traffic environments.

In [2] this article, Chuan Wu presents an innovative approach to image captioning network designed specifically for modeling traffic scenes. The key concept introduced is the integration of element attention into the encoder-decoder mechanism. This enhancement aims to generate scene captions that are not only more accurate but also more contextually sensible. By incorporating element attention, the network is able to focus on specific elements within the traffic scene, allowing for a more detailed and relevant description in the generated captions. This proposed method holds the potential to significantly improve the quality of scene captions, thereby enhancing the overall performance of systems tasked with understanding and interpreting traffic environments.

In [3] According to the authors, Wang et al. introduced a Spatial CNN model, as discussed in the article. This model extends traditional deep convolutional neural networks by adopting slice-by-slice convolutions within feature maps, rather than the conventional layer-by-layer convolutions. By doing so, it facilitates communication between pixels across both rows and columns within a layer, allowing for more comprehensive information exchange. This approach enhances the network's ability to capture spatial dependencies within the data, potentially leading to improved performance in tasks such as image processing and scene understanding.

In [4] This work aims to leverage a combination of machine learning, genetic algorithms, soft computing techniques, and deep learning algorithms to analyze vast amounts of data in the transportation system. By employing these advanced methodologies, the complexity of the data analysis process is significantly reduced. This streamlined approach ultimately facilitates the proper training of autonomous vehicles, ensuring that they can navigate and operate effectively in diverse real-world scenarios. Through the integration of these sophisticated algorithms, the transportation system can harness the power of big data to optimize efficiency, safety, and reliability, thereby advancing the development and deployment of autonomous vehicle technologies.

III. Proposed System

The proposed system aims to generate both text captions and audio for images. The system uses algorithms CNN, YOLO and LSTM networks for generating text-based captions and a text-to-speech (TTS) engine for generating audio descriptions.

- The system uses the Flicker dataset for training and evaluation purposes.
- The proposed system consists of several stages, including image detection, image augmentation, image feature extraction, text cleaning, tokenization, and LSTM-based captions generation.
- The LSTM model is trained on the pre-processed text and visual attributes to generate descriptive captions for both text and audio modalities.
- The proposed system uses a TTS engine to convert the text-based captions into audio descriptions. The TTS engine generates natural-sounding audio descriptions that capture the essence of the image.
- The system provides traffic understanding by image captioning and audio.

IV. Architecture

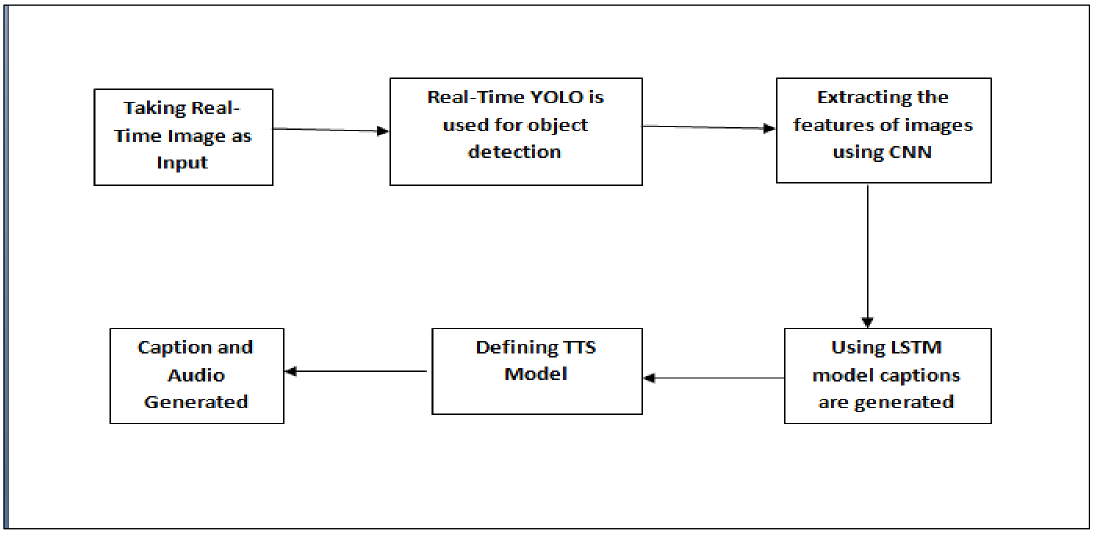
Taking Real-Time Image as Input
Real-Time YOLO is used for object detection
Extracting the features of images using CNN
Using LSTM model captions are generated
Defining TTS Model
Caption and Audio Generated

Fig.1. Architecture

The aim is ought to be to provide clear descriptions of the images in order to apply the image captioning method to the understanding of traffic scenes. The architecture combines object detection, feature extraction, caption generation, and text-to-speech (TTS) to create a comprehensive traffic image captioning system.

The system begins with a traffic image captured by a camera or sensor. YOLO processes the input image and performs real-time object detection.

It divides the image into a grid of cells and predicts bounding boxes around objects. Each bounding box includes class probabilities.

A pre-trained CNN (such as VGG, ResNet, or MobileNet) extracts high-level features from the detected objects. These features represent visual information about the objects' appearance and context. The CNN output serves as a rich representation of the image.

Then the LSTM model generates descriptive captions for the traffic scene based on the detected objects and scene analysis. It can describe objects, their positions, and the overall context.

The generated captions are converted into audio using a TTS system. TTS synthesizes natural-sounding speech from the textual descriptions. The audio output provides real-time information about the traffic scene.

The final output includes both the visual caption (displayed on a screen) and the audio description (played through speakers).

System Requirements:

- Processor - Intel i3
- Speed - 3.1 GHz
- RAM - 4 GB
- Hard Disk - 20
-

IV. Algorithms

4.1 You Only Look Once (YOLO)

YOLO revolutionized real-time computer vision tasks by providing accurate and efficient object detection.
YOLO is a single-stage object detection model that directly predicts bounding boxes and class probabilities in a single pass through the neural network.
Unlike multi-stage detectors (such as Faster R-CNN), YOLO doesn't rely on region proposals or anchor boxes. Each grid cell predicts multiple bounding boxes along with the confidence scores for those boxes and the probabilities of different object classes within those boxes. YOLO then combines these predictions to generate the final set of detected objects. YOLO's speed and accuracy make it ideal for real-time applications. YOLO achieve real-time object detection, meaning it can detect objects in video streams or live camera feeds.

4.2 Convolution Layer (CNN):

The layer in feature extraction from an input image is called convolution (image). Convolution uses small squares of input data to learn visual attributes, preserving the link between pixels. Using filters such as identity, edge detection, sharpen box blur, and Gaussian blur, convolution of an image with several filters.
Pooling
In situations where the photos are too big, pooling layers would lower the number of parameters. Spatial pooling sometimes referred to as down sampling or sub sampling, lowers each map's dimension while keeping crucial details.
Fully Connected Layer
In this layer Feature map matrix will be converted as vector (x1, x2, x3 …). With the fully connected layers, we combined these features together to create a model.
Softmax Classifier
Finally, we have an activation function such as softmax or sigmoid to classify the outputs.

4.3 Long Short-Term Memory:

LSTMs, or Long Short-Term Memory networks, represent a specialized form of recurrent neural networks (RNNs) frequently employed in image captioning tasks. Their popularity stems from their unique capability to capture long-range dependencies within sequences of data. Unlike traditional RNNs, LSTMs are equipped with memory cells that facilitate the storage and retrieval of information over extended time steps.

This characteristic proves invaluable in the context of image captioning, where generating coherent and contextually relevant captions is paramount. By effectively retaining important information over prolonged sequences, LSTMs enable the generation of captions that accurately reflect the content and context of the images they describe.

4.4 Text-To-Speech (TTS):

Integrating Text-to-Speech (TTS) technology with image captioning provides a means to convert textual image descriptions into natural language speech, enhancing accessibility for users. This integration enables individuals to receive audio descriptions of images, facilitating comprehension and accessibility. The audio output provides real-time information about the traffic scene.

V. Conclusion

The deep learning model presented in this paper aimed at enhancing image captioning methods by leveraging visual attention mechanisms. Our DNN model explores the relationship within visual attention and learns the mechanism of attention transmission through a customized LSTM model. By storing and propagating visual attention in a matrix-form memory cell and filtering attention values with a reconstructed output gate, our model effectively captures and utilizes visual attention information.

Through integration with a language model, both generated words and visual attention areas are endowed with memory in a unified space. We embed our DNN model into three classical attention-based image captioning frameworks and demonstrate its superiority through experimental evaluations on the Flicker datasets.

Furthermore, our system integrates a text-to-speech (TTS) engine, enabling the conversion of text-based captions into natural-sounding audio descriptions. By providing both textual and audio descriptions, our system enhances traffic understanding through image captioning, making crucial information about the traffic environment accessible to a broader range of users.

REFERENCES

[1]WEI LI 1,2, ZHAOWEI QU 1 , HAIYU SONG 1,2, PENGJIE WANG 2 , AND BO XUE2, "The Traffic Scene Understanding and Prediction Based on Image Captioning", 2021, IEEE Access.

[2]Chuan Wu, Yaochen Li, Ling Li, Le Wang, Yuehu Liu, "Caption Generation from Road Images for Traffic Scene Construction", 19 Oct 2020, IEEE.

[3]P. Aishwarya Naidu1 *, Satvik Vats2 , Gehna Anand3 , Nalina V. 4, "A Deep Learning Model for Image Caption Generation'. International Journal of Computer Sciences and Engineering on 6th June 2020.

[4]V.Geetha, C K Gomathy, T. Harshitha, P. Vijay Nagendra Varma, "A Traffic Prediction for Intelligent Transportation System using Machine Learning," in International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-10 Issue-4, April 2021.

[5]Gauri Rao1 , Bhavya Divecha2 , Jenish Joshi 3 , Anishk Jaiswal 4, "Traffic Light Management System Using Image Processing", July 2022| IJIRT | Volume 9 Issue 2 | ISSN: 2349-6002.

[6]Xiaoyuan Liang, Xusheng Du, "Deep Reinforcement Learning for Traffic Light Control in Vehicular Networks", IEEE, 2018.