Text Representation Enhancement using Multi-Modal Attention Mechanism for Text Visual Question Answering

Esther Oduntan Ph.D and Joseph Domguia
AI|Machine Learning|Data Science

ABSTRACT

Text Visual Questions and Answering is an integral part of communication.Researchers in Natural Language Processing and Computer Vision have been used various architectures such as Bottom-up and Top-Down Visual Attention Mechanism, Visual-BERT, M4C to mention a few. In this research, we looked into the application of multi-modal attention mechanism to improve the question and answering system. Hence,it was observed that the quality of answer generated from the pre-trained MC4 model with an improved OCR will enhance the performance for text visual question and answering system.

INTRODUCTION

Visual Question Answering (VQA) system can be defined as an algorithm that takes as input an image and a natural language question about the image and generates a natural language answer as the output. This is a computer vision task where a system is given a text-based question about an image, and it must infer the answer. The main idea in Text VQA is that the search and the reasoning part must be performed over the content of an image. The system must be able to detect objects, it needs to classify a scene and needs world knowledge,commonsense reasoning and knowledge reasoning are necessary.

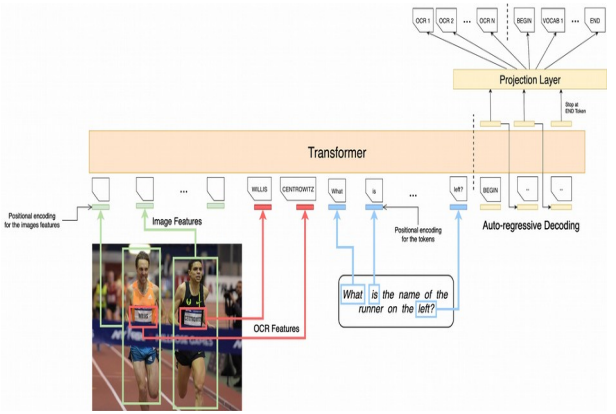
RELATED WORKS

"Bottom-up and top-down attention for image captioning and visual question answering," by P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang(2018).

"Multimodal grid features and cell pointers for scene text visual question answering," by L. Gómez, A. F. Biten, R. Tito, A. Mafla, and D. Karatzas(2020)

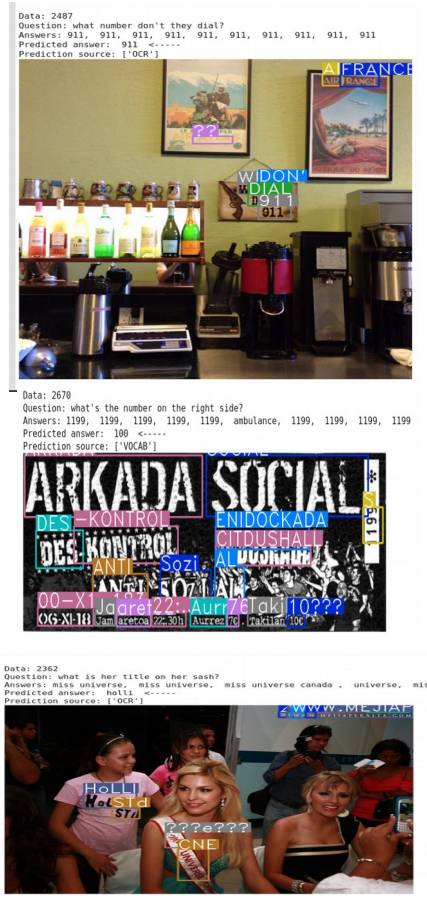
METHODOLOGY

The method used in this study, is a multi-modal technique that involves more than one input and outputs. The inputs are the questions, images and optical character reader (OCR) extracted features. The images features were extracted using the Faster R-CNN model, the contextual word embedding features from the questions were extracted using global vec(GLO Vec) and text on images were extracted using the google ocr-app. All the input features were passed into the transformer. The dataset that was used is the TextVQA dataset.



The Multimodal OCR-Transformer Based VQA Architecture

RESULTS



CONCLUSION

This project has been able to implement a multi-modal attention mechanism for visual question answering system, by integrating Google OCR and Rosetta OCR on an M4C pre-trained model. Out of the three approaches intended to use in enhancing text representation, two of the approaches were implemented, while the third approach of training from scratch and fine-tuning the pre-trained M4C model will be examined for future work