CSE 6367 2023

PROJECT PROPOSAL

Presented to:

Marnim Gallib

Proposal by:

Nelson Joseph
1002050500

Problem Description

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

The goal of the this project proposal is to create an image captioning model with text-to-speech functionality to help people with vision impairments comprehend their environment. To create compelling descriptions for photographs that can be spoken, the model will employ deep learning techniques. The goal is to develop an intuitive application that offers real-time audio descriptions, boosting independence and navigation for those who are blind or visually impaired.

Related Work

- Semi-Autoregressive Image Captioning (Paper)
- DeeCap: Dynamic Early Exiting for Efficient Image Captioning (Paper)
- SpeechCLIP: Integrating Speech with Pre-Trained Vision and Language Model (Paper)
- Show, Translate and Tell (Paper)
- Guided Open Vocabulary Image Captioning with Constrained Beam Search (Paper)
- AVLnet: Learning Audio-Visual Language Representations from Instructional Videos (Paper)
- Show, Translate and Tell (Paper)
- NICE: CVPR 2023 Challenge on Zero-shot Image Captioning (Paper)
- A Picture is Worth a Thousand Words: A Unified System for Diverse Captions and Rich Images Generation(Paper)

Datasets

- MSCOCO
- Flickr30k
- Flickr10k
- COCO Captions
- Conceptual Captions

Final Deliverables

- Image Captioning Model
- Text-to-Speech Module
- Testing and Evaluation Reports

References

- Image captioning with visual attention
- Current challenges and limitations of image captioning
- How to Develop a Deep Learning Photo Caption Generator from Scratch
- Generating image captions from the camera feed
- A Real-time Image Caption Generator based on Jetson nano
- Exploring Deep Learning Image Captioning