Are Visual Question Answering (VQA) Models Ready for Navigational Assistance to Blind and Low-Vision Users?

Elena Pearce[1], Md Touhidul Islam[2], Imran Kabir[2], Syed Masum Billah[2], and Md Alimoor Reza[1]

1. Drake University

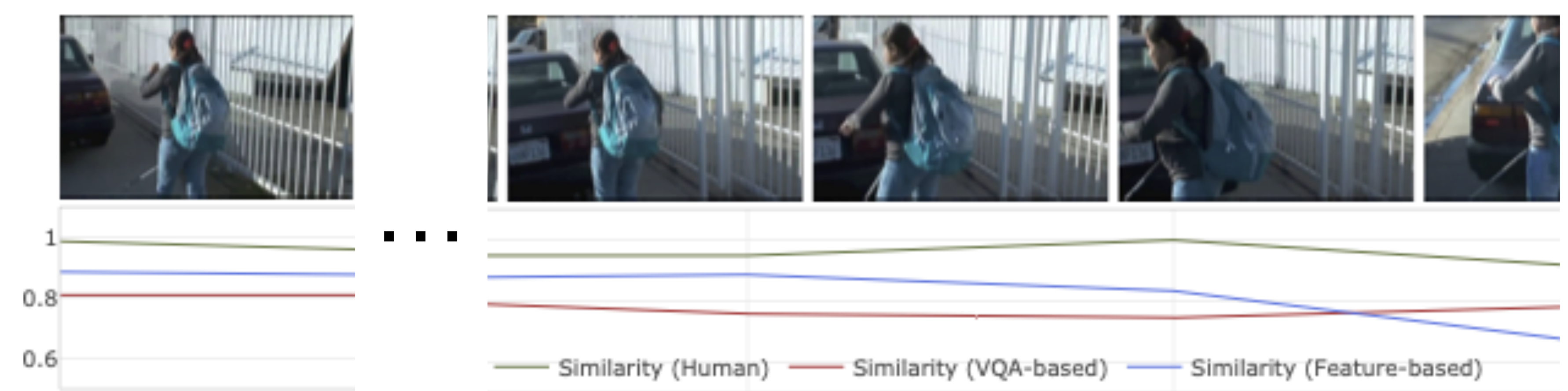
2. The Pennsylvania State University

Motivation

- Visual Question Answering (VQA) models aim to predict an open-ended answer based on input image and a textual query about its content [1]
- VQA models are promising for blind and low-vision (BLV) users as they could benefit by interacting with the models in a dialog-style conversation to learn the image content, e.g., what objects are present, the number of occurrences of a specific object, spatial relationships of different objects, etc.
- VQA models can also be used for navigational assistance to BLV users, who can learn about the obstacles on their pathways

Problem Statement

- Evaluated the robustness of existing VQA models by testing against human annotation
- Construct a data set comprised of several annotated key frames of video segments along with ground truth for each video segments
- Analyze the key frames with VQA models and use the ground truth tables to evaluate the accuracy of the models



Dataset and Taxonomy

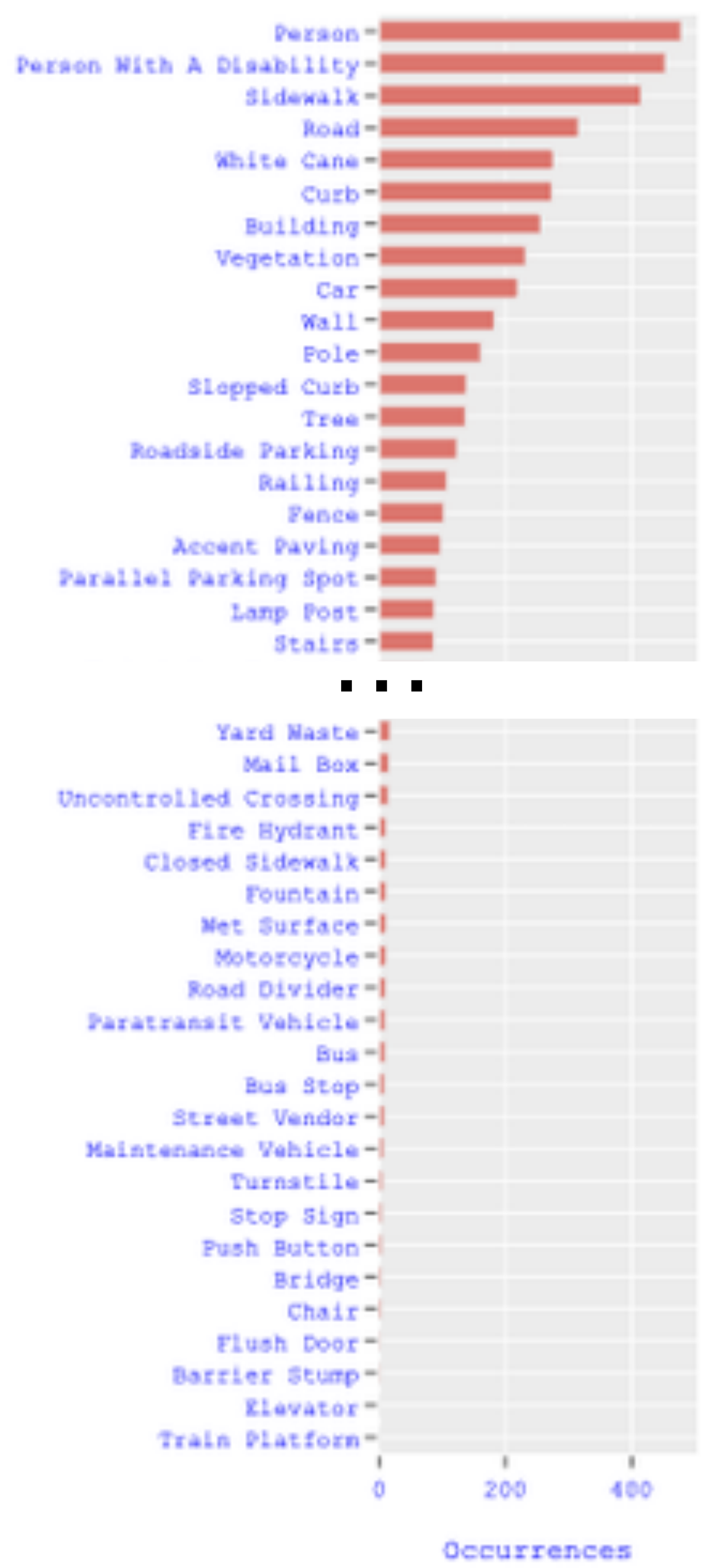
- The dataset was created by collecting free, publicly available videos from YouTube and Vimeo
- Annotating objects with accessibility impact was done by inputing the time in which an important object appeared
- Ground truths for the videos were made by reviewing keyframes and comparing the objects present to those on the taxonomy that was compiled from the list of annotated objects

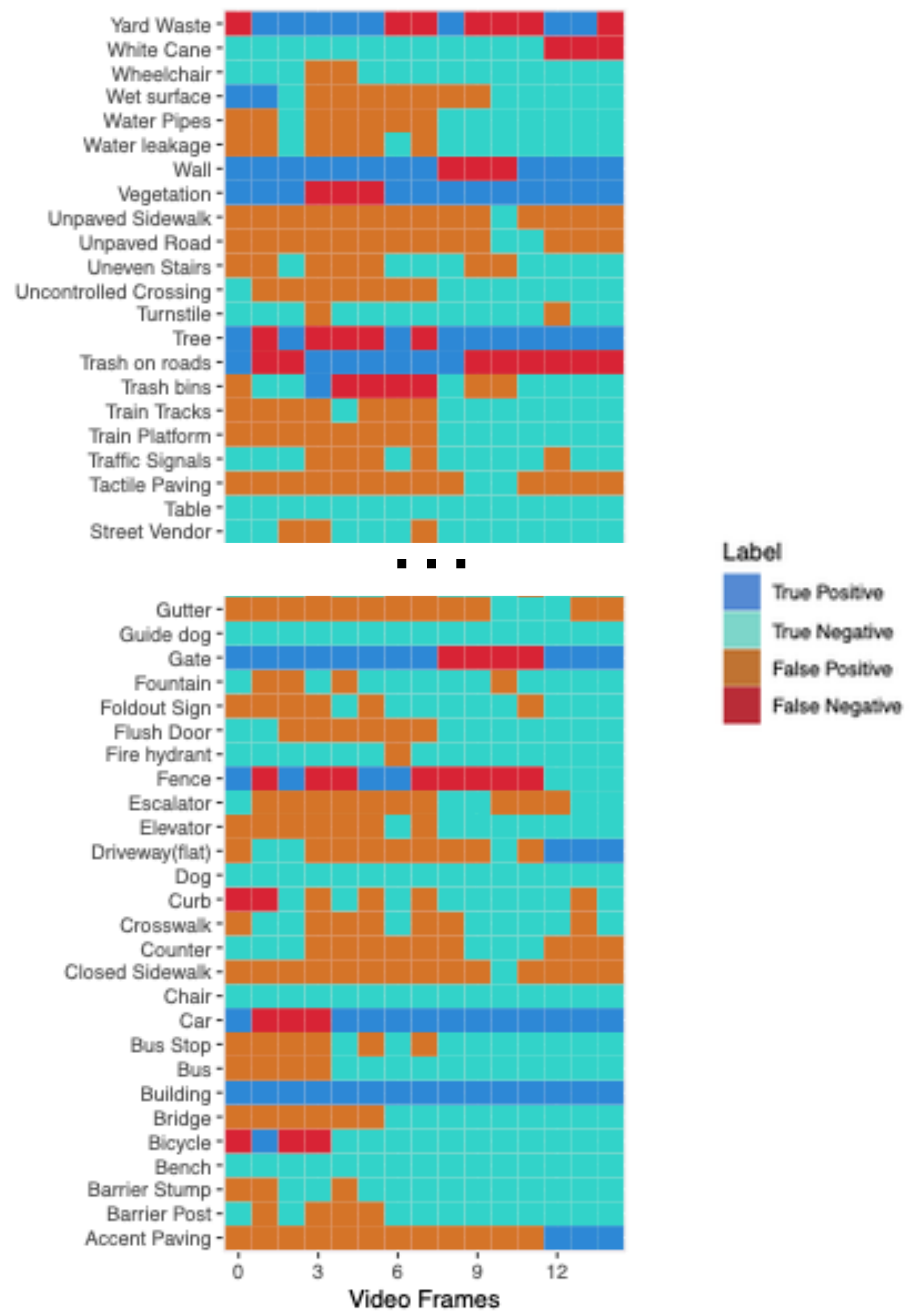
ID	Parent Concept	Accessibility-related objects
1	Attributes of a sidewalk and driveway	Accent Paving, Driveway (flat), Puddle, Raised Entryway, Sidewalk, Sidewalk Pits, Sloped Driveway, Tactile Paving, Brick Paving, Cobblestone Paving, Unpaved Sidewalk, Wet Surface
2	Obstructions likely to be detected by a white cane	Fire hydrant, Gutter, Vegetation, Tree, Brick Wall, Fence, Trash Bins, Lamp Post, Pole, Mailbox
3	Obstructions less likely to be detected by a white cane	Closed Sidewalk, Barrier Post, Barrier Stump, Foldout Sign, Bench
. . .	. . .	. . .
9	Intersection	Pedestrian Crossing, Slopped Curb, Intersection, Crosswalk, Curb, Bridge, Uncontrolled Crossing
10	Objects on the road shoulder	Road Shoulder, Roadside Parking, Parallel Parking Spot, Paratransit Vehicle
11	Objects on the road	Road, Unpaved Road, Bus, Car, Motorcycle, Road Divider
12	Traffic signals and street signs	Traffic Signals, Stop Sign, Sign, Sign Post, Push Button, "Use the Other Door" Sign, Toilet Sign
13	Objects related to building exits and entrances	Gate, Flush Door, Doorway

Visual Question Answering (VQA)

- The rapid advances in computer vision and natural language processing (NLP) techniques and the availability of large-scale datasets largely contributed to a growing interest in VQA eg, VilBERT, UNITER, and LXMERT
- In this work, we used two VQA models: GPV-1 and BLIP
- General Purpose Vision (GPV-1)[2]: capable of solving a variety of tasks, such as VQA, localization, image captioning, etc. This flexible and end-to-end trainable model does not require any modification to network architecture for adapting to a new task
- BLIP[3]: Vision language pre-training with additional noisy image-text pairs collected from the internet has been effective in boosting the performance of various vision language tasks, including VQA

Results





Conclusion and Future Work

- We found VQA models exhibit sequential-frame answer inconsistency, where their answers to our dataset's questions differ significantly despite the visual information across two framesremaining nearly identical. Our conclusion was VQA models are not ready for BLV users yet
- Increasing the robustness of VQA models in the future

References

1. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Zitnick, and D. Parikh. VQA: Visual Question Answering. In IEEE International Conference on computer vision (ICCV), 2015
2. T. Gupta, A. Kamath, A. Kembhavi, and D. Hoiem. 2022. GPV: Towards General Purpose Vision Systems. Conference of Computer Vision and Pattern Recognition (CVPR), 2022
3. J. Li, D. Li, C. Xiong, and S. Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning (ICML), 2022

Acknowledgements: This work was partially supported by Penn State University IST Seed Grant