Integrating vision and language modalities for visually challenged people

Vision-language models (VML) have emerged as transformative tools in assisting visually impaired individuals by bridging the gap between visual and textual/auditory understanding. VLMs help assist people with blindness and low vision (pBLV) by providing comprehensive scene understanding and detailed environmental descriptions by offering real-time assistance for daily tasks and navigation and help in tasks like reading, object identification etc.

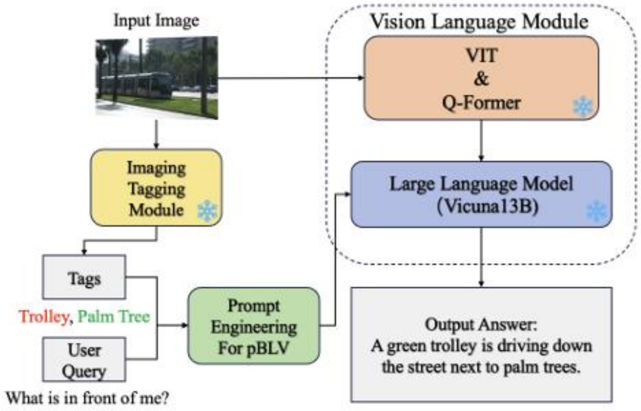
Development of large scale vision language models, improvement in modern computer vision techniques, integration of multimodal AI systems that can handle various data types concurrently along with great advancements in hardware like GPUs, processors, cameras etc have made it possible to train large datasets on large models.[1]

Multi model foundation models like VisPercep provides detailed environmental descriptions and risk assessments for visually impaired users, and wearable device like Alris combine sophisticated cameras with natural language processing for real-time auditory descriptions.

Components of the VisPercep Architecture[2]

Image Tagging Module:

- Recognize Anything Model (RAM) to identifies all common objects in the captured images.
- Employs a pre-trained image encoder to extract high-level features, followed by an attention mechanism to focus on relevant objects.
- Maps features to a set of object categories or tags, to create understanding of the scene.



Prompt Engineering Module:

- Combines user queries and the identified tags to generate contextually relevant prompts.
- Tailored for visually impaired users by focusing on clarity and respect in responses.
- Prompts include descriptions like: "The image may contain elements of {tags}." or personalized instructions based on user queries for risk assessment and scene description.

Vision-Language Module:

- Based on InstructBLIP, a vision-language foundation model.
- Processes input images and prompts to generate comprehensive textual descriptions and analyses.

- Uses a Vision Transformer (ViT) for encoding images and a Q-Former mechanism for contextual embeddings. These are combined with prompt embeddings to generate descriptive outputs.

System Workflow

- Captures images using a smartphone and processes them alongside user queries.
- Outputs detailed scene descriptions, risk assessments, and object localizations.
- Results are delivered as text or synthesized audio for accessibility.

Total Loss = Image Tagging Loss + Vision-Language Loss + Contrastive Loss + Prompt Engineering Loss

- Image Tagging: classify and tag objects correctly
- Vision-Language: contextually relevant text descriptions or answers to user queries based on the image and the prompt
- Contrastive Loss: correct pairs of images and text are close in the embedding space, while incorrect pairs are far apart.
- Prompt Engineering: Reinforcement Learning to tailor responses to the needs of pBLV

Evaluation on the Visual7W Dataset

Visual question-answering with grounding, 7W categories (ex : what, where, who), 47,300 images and more than 327,900 question-answer pairs, evaluates scene understanding, object localization, and risk assessment.

Inference Time: VisPercep demonstrates lower inference times for both the image tagging module and the vision-language module compared to prior methods.

- Scene Understanding: 8.85/10, Object Localization: 8.60/10, Risk Assessment: 9.40/10

Evaluation on the VizWiz Dataset

designed for assistive technology, focusing on real-world challenges faced by pBLV, 31,000 images with 40,000+ questions, Questions are categorized into: Unanswerable: Cases where the image lacks enough information, Yes/No: Binary answers, Number: Quantitative responses, Other: Open-ended answers, Reflects real-world conditions, including varying image quality, occlusions, and lighting

Average Scores:

- BLEU_1: 25.43, BLEU_2: 14.52, METEOR: 19.09, ROUGE-L: 35.76, CIDEr: 53.98

Limitations

- Dynamic and Complex Environments: Factors such as varying lighting conditions, weather changes, and moving objects can affect the system's performance and potentially lead to false alarms.
- Inherent Limitations of AI Models: AI-based systems are not infallible as they operate within the confines of their training data and algorithms, which might not cover every possible real-world scenario or object encountered by pBLV. This limitation can further lead to inaccuracies or false positives in object detection and scene interpretation if blurry images are captured through cellphones.
- Potential for hallucination and misalignment between automatic evaluation metrics and human judgment.

Citations:

[1] Granquist, C.; Sun, S.Y.; Montezuma, S.R.; Tran, T.M.; Gage, R.; Legge, G.E. Evaluation and comparison of artificial intelligence vision aids: Orcam myeye 1 and seeing ai. J. Vis. Impair. Blind. 2021, 115, 277–285.

[2] https://pmc.ncbi.nlm.nih.gov/articles/PMC11122237/ (Hao, Y.; Yang, F.; Huang, H.; Yuan, S.; Rangan, S.; Rizzo, J.-R.; Wang, Y.; Fang, Y. A Multi-Modal Foundation Model to Assist People with Blindness and Low Vision in Environmental Interaction. J. Imaging 2024, 10, 103. https://doi.org/10.3390/jimaging10050103)