VQA for Visually Impaired

Mahmoud Khalil, Zanming Huang, Qipeng Zou
Boston University Department of Electrical & Computer Engineering (ECE) & Systems Engineering (SE)

Introduction

A natural application of artificial intelligence is to help the blind and visually impaired (BVI) to overcome their daily visual challenges and allowing them to live an independent life through AI-based technologies. In this regard, one of the most promising tasks is Visual Question Answering (VQA).
Our work includes:

- Train and validate different VQA architectures to compare performance
- Examine the transferability of general VQA models to the BVI VQA tasks
- Explore the reasons of performance discrepancy of models trained and evaluated on VQA v2 and Vizwiz dataset.

Architectures & Results

Original VQA [Agrawal, et al. 2016] **:**

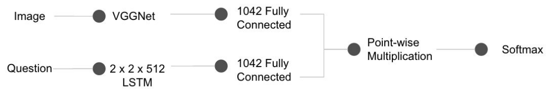


Image → VGGNet → 1042 Fully Connected
Question → 2 x 2 x 512 LSTM → 1042 Fully Connected
→ Point-wise Multiplication → Softmax

Strong Baseline VQA [Kazemi, et al. 2017]**:**

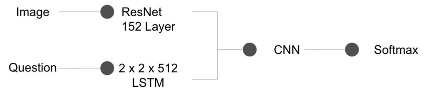
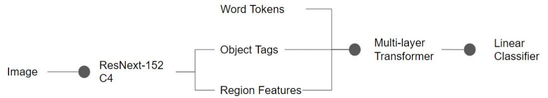


Image → ResNet 152 Layer
Question → 2 x 2 x 512 LSTM
→ CNN → Softmax

VinVL [Zhang, et al. 2021]**:**



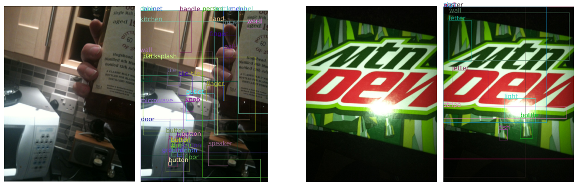
Word Tokens
Image → ResNext-152 C4 → Object Tags, Region Features
→ Multi-layer Transformer → Linear Classifier

	Train VQA v2 & Val. VQA v2	Train VQA v2 & Val. VizWiz	Train Vizwiz & Val. VizWiz
Basic VQA	54.42	49.21	9.78
Strong Baseline	58.54	51.38	10.37
VinVL	74.37	47.31	6.68

Analysis of Datasets

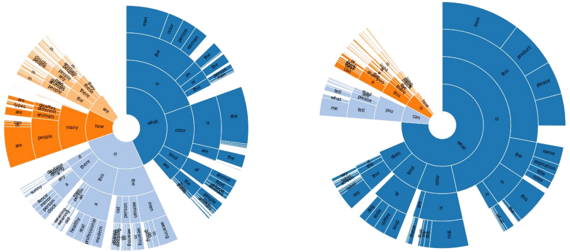
		VizWiz	VQA v2
Purpose		VQA dataset collected by visually impaired people	Largest and most recent dataset containing open-ended questions about images
Size (Images)	Training	20,523	82,783
	Validation	4,319	40,504
	Test	8,000	81,434
Visual Questions		31,173	~1 million

Image and Image Features

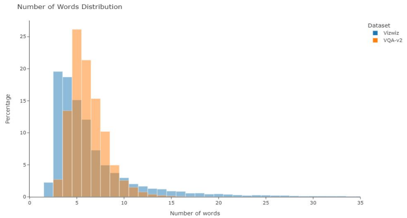


- Images from Vizwiz are blurry and don't contain object of interest this lead to poor performance across all models.
- Image features extracted from VinVL are often mislabeled, introducing false information to the prediction process.

Questions

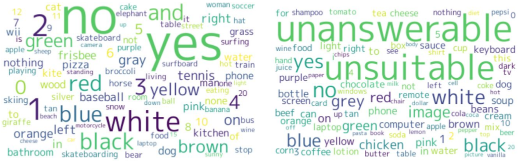


Distribution of the first four words of all the questions in VQA and Vizwiz show discrepancy between question types in both datasets. Suggesting that VQA v2 does not address the needs of the BVI.



Distribution of number of words in both datasets show Vizwiz contains longer questions compared to VQA v2. Upon inspection, the questions in Vizwiz are more natural and conversational.

Answers



Word clouds for answers in both datasets show that the majority of answers in VQA v2 are boolean while, the majority in Vizwiz are “Unanswerable” or “Unsuitable”

Problem Definition
Most models define VQA as a multi-label classification problem with labels from a predetermined answer set. Hence, the model generalizes poorly when facing questions that has out-of-set answers. This could be one of the reasons that caused poor transferability of models between the datasets.

Conclusion & Future work

- We observed that newer model with higher complexity performed better on VQA v2 dataset. However, Basic VQA and Strong Baseline achieved higher accuracy on the Vizwiz dataset than the newer VinVL model.
- We explored the reasons of performance discrepancy in four aspects, including image and image feature, questions, answers as well as problem definition.

In conclusion, the current development trend of VQA models has not taken enough consideration into the needs of BVI. For future works, we hope to develop more inclusive models to address the BVI specific problems mentioned above.

We would like to thank Prof. Osama Al Shaykh and Graduate Teaching Assistant Gilbert for their guidance, advice and support throughout.