Sprint 1
Visual Question Answering
Group: Mahmoud Khalil, Zanming Huang, Qipeng Zou, Zilun Huang

VQA for Visually Impaired

Define your product mission:

To reconnect visually impaired people with the world through technology, and this is achieved through building machines that can replicate the human vision system. The main goal is to understand the challenges of previous technologies aimed to solve this problem and try to use state of the art VQA models to attempt replicate results on a new data set.

Comprehensive literature review: History of technology to assist blind people

https://vizwiz.org/

Crowdsourcing Subjective Fashion Advice Using VizWiz: Challenges and Opportunities - 2010	Paper	The paper describes a study exploring fashion perception among those with vision impairments, as well as discusses an online survey in order to learn how individuals with and without vision impairments get fashion advice. Then, it presents a pilot study using VizWiz, a mobile phone application for people with vision impairments, to test the feasibility of having sighted people answer subjective fashion questions
Visual Challenges in the Everyday Lives of Blind People - 2012	Paper	This paper introduces visual questions that blind people would like to have answered. Questions include a diverse selection of accessibility issues encountered in everyday life. (1) Identification: name, type, brand medicine, currency, media (2) Reading: mail, digital displays, numbers (3) Description: Visual/Physical properties appearance, color, clothing, state (4) Outside of range/Unanswerable

Sprint 1
Visual Question Answering
Group: Mahmoud Khalil, Zanming Huang, Qipeng Zou, Zilun Huang

Answering Visual Questions with Conversational Crowd Assistants - 2013	Paper	Paper introduces a system named Chorus, the system provides the user with answers to sequential questions. This is done by allowing the user to have a reliable conversation with the crowd (People working with the company) about a video stream from the user's phone. The paper discusses the feasibility of the system as well as the user feedback from a few years of user studies.
CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question - 2017	Paper	Paper discusses how existing systems do not account for the fact that a visual question can lead to a single answer or multiple answers. The paper proposes a model called CrowdVerge, that automatically predicts from a question if a crowd would agree on one answer. They then can leverage this prediction to estimate the number of human responses needed for a visual question. Results show that their model provides the same answers with typically 23% less crowd involvement.
Visual Question Answer Diversity - 2018	Paper	This paper also discusses the fact that a visual question can lead to a single answer or multiple answers. They propose a model that predicts the answer distribution that would be expected from a crowd for a given visual question. They also leverage this to predict how many answers are needed for a VQ.
VizWiz Grand Challenge: Answering Visual Questions from Blind People - 2018	Paper	The paper introduces a VQA dataset called Vizwiz and tests its utility. First, it collects the questions, and then implements a filtering process to remove visual questions that compromise the privacy of any individuals; Then, crowdsources answers to support algorithm training and evaluation. In the next step, conduct experiments to characterize the images, questions, and answers and uncover unique aspects differentiating VizWiz from existing VQA datasets. Finally, evaluates numerous algorithms for predicting answers and predicting if a visual question can be answered.

Sprint 1
Visual Question Answering
Group: Mahmoud Khalil, Zanming Huang, Qipeng Zou, Zilun Huang

VizWiz-Priv: A Dataset for Recognizing the Presence and Purpose of Private Visual Information in Images Taken by Blind People - 2018	Paper	The paper introduces the first visual privacy dataset originating from people who are blind to better understand their privacy disclosures. It begins with introducing the first visual privacy dataset originating from this population whose images were taken and shared by blind users Then, for each image, manually annotates private regions according to a taxonomy that represents privacy concerns relevant to their images. It also annotates whether the private visual information is needed to answer the question asked by the user. These annotations are critical because they can decide (1) whether private information is in an image and (2) whether a question about an image asks about the private content in the image. Lastly, the paper does VQA and collects the results in order to test numerous algorithms for both purposes.
“I Hope This Is Helpful”: Understanding Crowdworkers’ Challenges and Motivations for an Image Description Task - 2020	Paper	The paper analyses feedback provided by Amazon Mechanical Turk workers who worked on image captioning tasks for images collected by blind people. The paper aims to understand “What are crowdworkers’ reactions to creating image descriptions ?” by answering 4 questions using the comments: 1) What challenges do workers express with understanding how to complete this task? 2) What suggestions do workers have for improving this task? 3) What kinds of explanations or clarifications do workers give about their work? 4) Why do workers find this task rewarding or interesting?
Vision Skills Needed to Answer Visual Questions - 2020	Paper	The contribution of this paper is three-fold. First it identified the skill needed for assisting populations with visual impairments as well as a visual Turing test for the AI community. The paper categorized the task needed into 4 types: 1) object recognition 2) text recognition 3) color recognition 4) counting The paper found out that real VQA service users

		are mostly concerned with learning about the text and color of common objects, the AI community more strongly emphasizes learning about the type, count, and color of objects in images. Then, it quantifies the difficulty of these skills for both humans and computers on both datasets. In the end, the paper proposed a task of predicting what vision skills are needed to answer a question about an image.

Define MVP:

Input: Image + Question

Output: Answer to a question

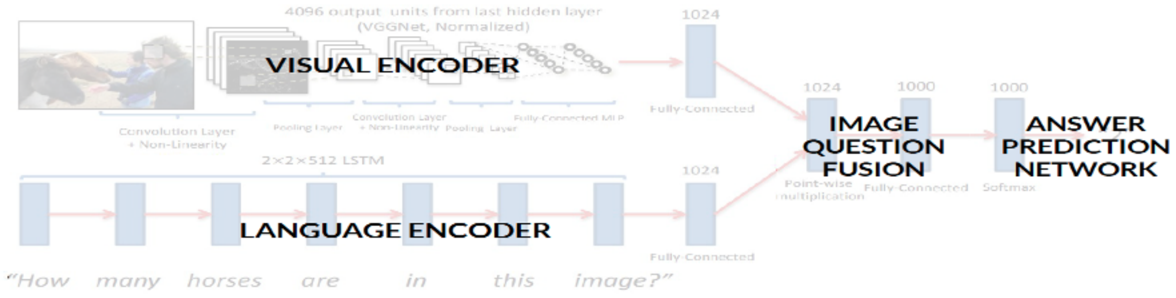
As a user, I want to ask a question concerning an image I have taken and want an answer to the question.

Technologies to evaluate:

The VQA model generally extracts features from the image and the question separately, and then combines the two to do some multimodal fusion (such as element-wise product, MCB, MFB), attention, knowledge supplement and other processing, and finally the answer is output through the classifier .

Previous techniques relied heavily on humans to answer visual questions, however in recent years with the rise of new deep learning algorithms aiming to solve both NLP and image processing problems, there has been more interest in leveraging these new techniques to solve some of the issues of VQA. For a specific picture, if we want the machine to use natural language processing to answer a specific question of the picture, we need to let the machine have a certain understanding of the content, meaning and intention of the image.

Sprint 1
Visual Question Answering
Group: Mahmoud Khalil, Zanming Huang, Qipeng Zou, Zilun Huang



Our aim is to study the most recent VQA deep learning techniques and compare their metrics in order to determine which ones would be more suitable for our application.

All VQA systems consist of four main parts: Image featurization for converting images into their feature representation, question featurization for converting natural language into their embeddings, joint feature representation to combine both the image and question features, and lastly, answer generation to generate the correct answer. Our main goal for sprint 2 is to dig deeper and understand the different techniques used to answer each part of the four components and determine a model that best fits our data, keeping in mind that we want a model that can answer open ended questions.

To help with understanding our work related to doing VQA for the blind people, we intend to use VizWiz, the first goal-oriented VQA dataset to test with algorithms supported to help blind people.

Setup of development environment

Below is a table showing some of the research done in the field of VQA and VLM, and the tools they used for implementation. We observed that the recent papers all used Python+PyTorch to implement their model. Based on this observation, we might choose to use Python+PyTorch going forward.

‘

Sprint 1
Visual Question Answering
Group: Mahmoud Khalil, Zanming Huang, Qipeng Zou, Zilun Huang

Papers	**Tools used**
VQA-LSTM-CNN [Paper] [Code]	Torch (Lua), NLTK
HieCoAttenVQA [Paper] [Code]	Torch (Lua), NLTK
Bottom-up-attention [Paper] [Code]	Caffe (Python)
LXMert [Paper] [Code]	PyTorch
VisualBERT [Paper] [Code]	PyTorch, AllenNLP
Oscar [Paper] [Code]	PyTorch
Uniter [Paper]	PyTorch
MDETR [Paper] [Code]	PyTorch
ResNet-152[Paper][Code]	Pytorch