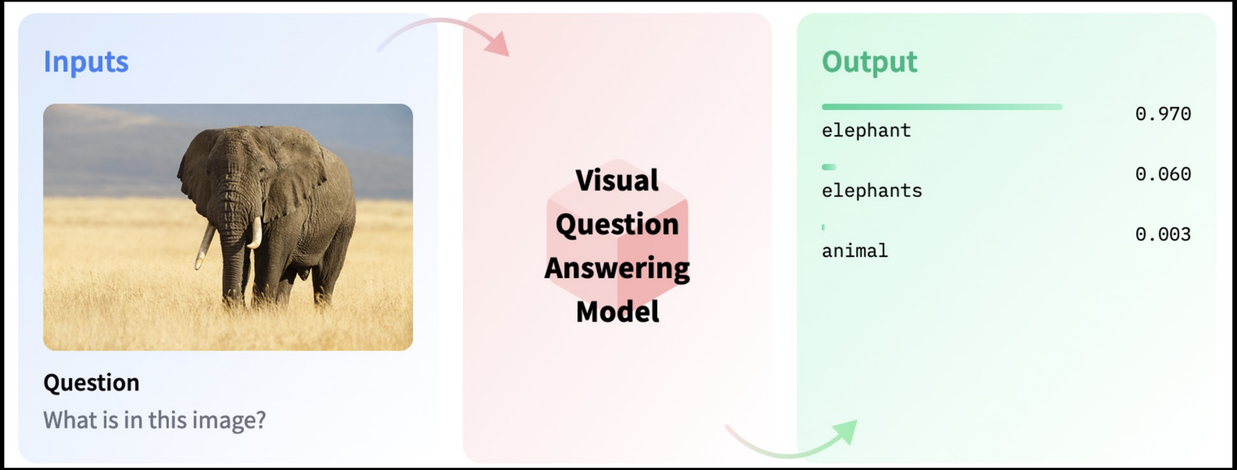
# VizWiz VQA

By: Jack Campbell, Ishan Ranjan, and Rani Shah

# Defining Visual Question Answering

A **Visual Question Answering (VQA)** model take as input an image and and a natural language question about the image and generates a natural language answer as an output.



Inputs

Question

What is in this image?

Visual Question Answering Model

Output

```
elephant          0.970
elephants         0.060
animal            0.003
```

***Use cases:***

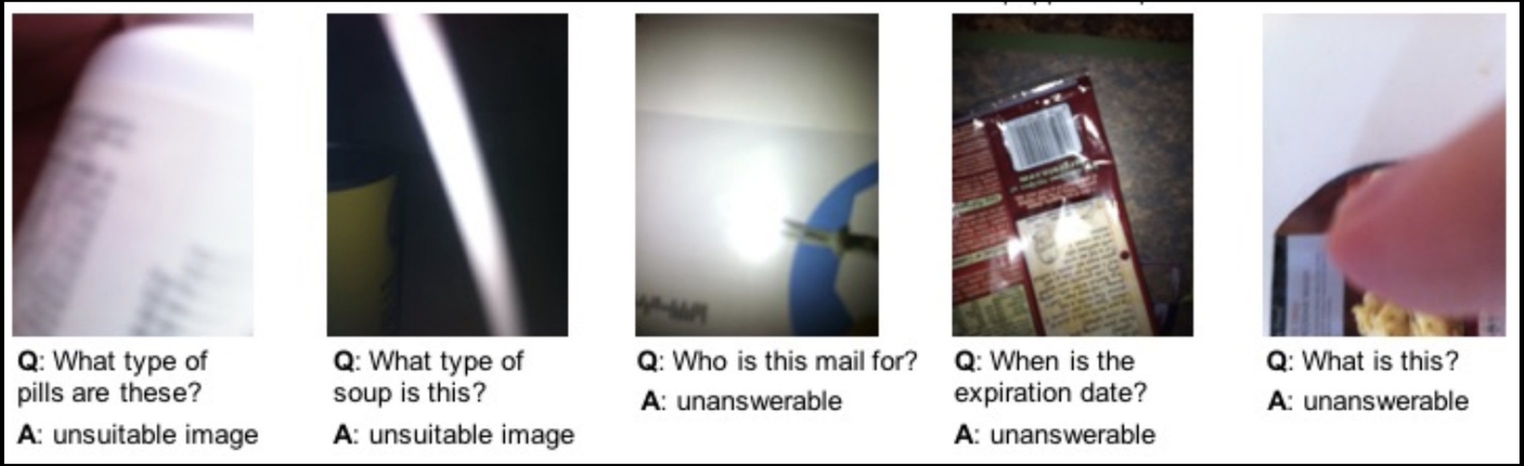
- Education
- Improved image retrieval
- Video search
- Aiding the visually impaired

# VizWiz-VQA Challenge

The VizWiz-VQA challenge task originates from the desire to educate more people about the technological needs of blind people while providing a opportunities for researchers.

**Main Objectives:**

1. Predict the answer to a visual question
2. Predict whether a visual question cannot be answered.



**Q**: What type of pills are these?
**A**: unsuitable image

**Q**: What type of soup is this?
**A**: unsuitable image

**Q**: Who is this mail for?
**A**: unanswerable

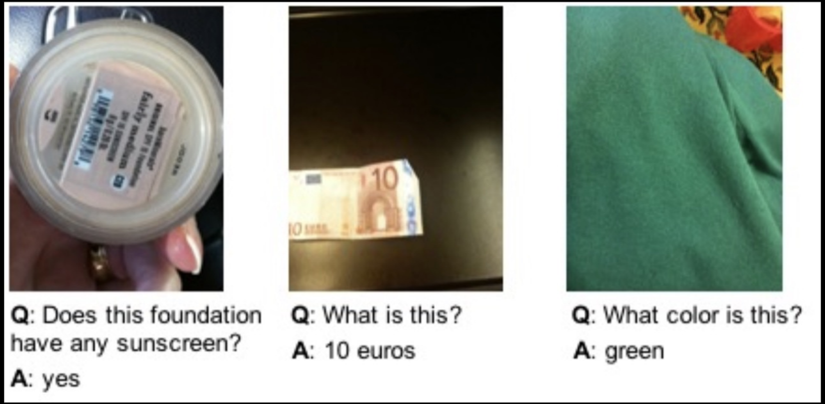
**Q**: When is the expiration date?
**A**: unanswerable

**Q**: What is this?
**A**: unanswerable

# The Dataset

Images taken on mobile phones, paired with questions asked by blind users, and 10 crowdsourced answers per image/question pair.

- **Training Set**: 20,523 image/question pairs & 205,230 answer/answer confidence pairs
- **Validation Set**: 4,319 image/question pairs & 43,190 answer/answer confidence pairs
- **Test Set**: 8,000 image/question pairs



**Q**: Does this foundation have any sunscreen?
**A**: yes

**Q**: What is this?
**A**: 10 euros

**Q**: What color is this?
**A**: green

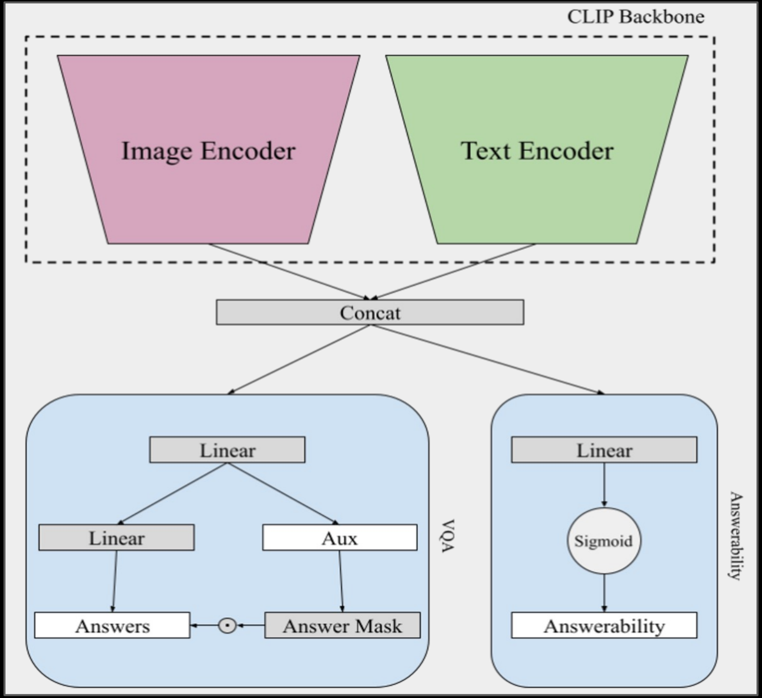
# Model Architecture: *Less is More*

**Model Backbone:** CLIP (Contrastive Language-Image Pre-Training), ViT-like Multi-modal transformer

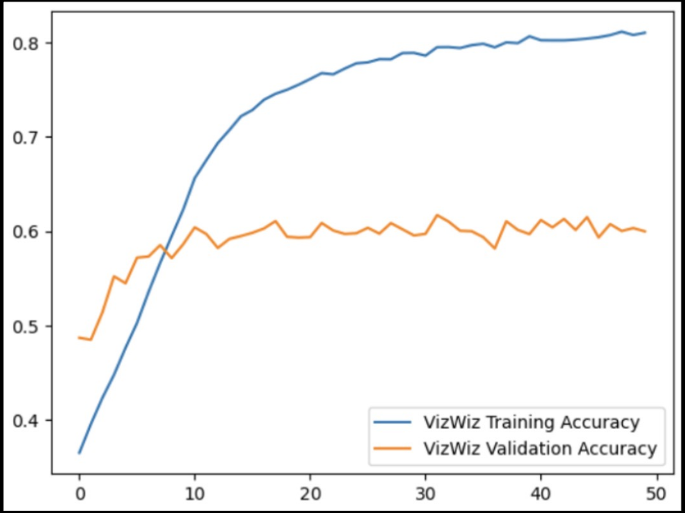
***Key Insight:*** No need to retrain CLIP, utilize both text and image encoder provided and add linear layers for Answering questions and Answerability Tasks

**Answer Generation:** Add two linear layers, as well as auxiliary layer

**Determining Answerability:** Add Linear Layer into Sigmoid Loss to classify Answerability



# Model Results: *Less is More*



**Training**

VizWiz Accuracy: 0.804
Answerability: 0.802

**Validation**

VizWiz Accuracy: 0.615
Answerability: 0.798

# Proposed Work

**Architecture:** Suitable model architecture derived from work by Deuser et al. using ‘Less is More’  design principles

**Models:** Test other models besides CLIP

**Fine-tune hyperparameters:** learning rate, weight decay, epochs, neuron dropout rate, and optimizer choice

# Findings

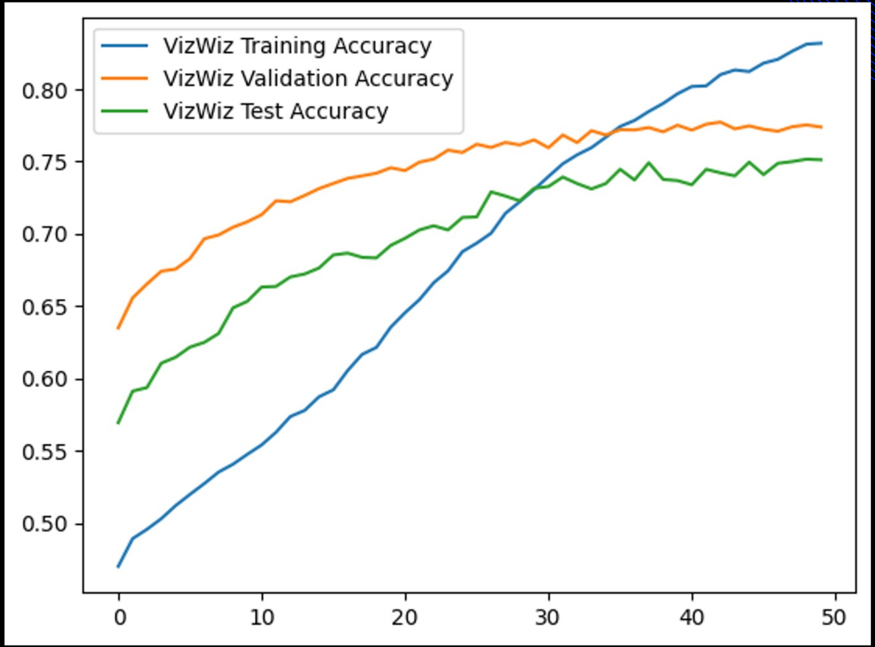
VizWiz Accuracy Score:

- Validation: 0.774
- Test: 0.751

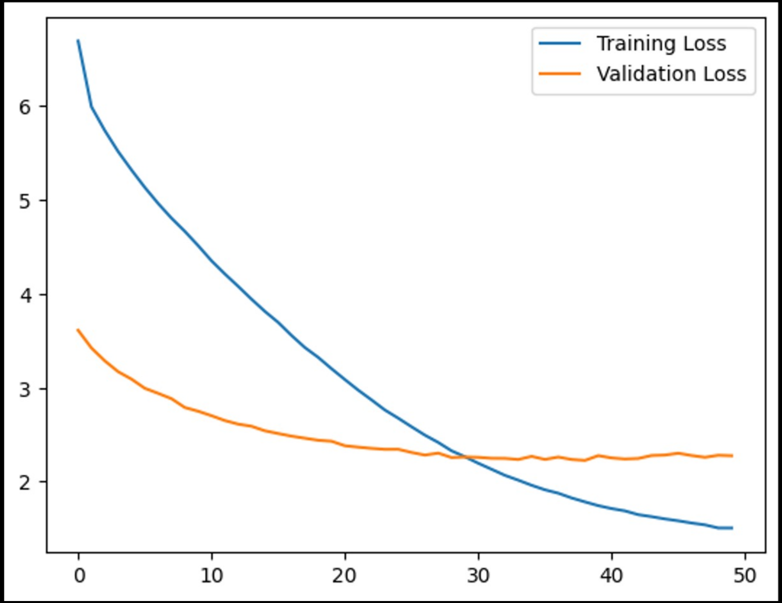
Answerability Score: 0.798

**Final Parameters:**

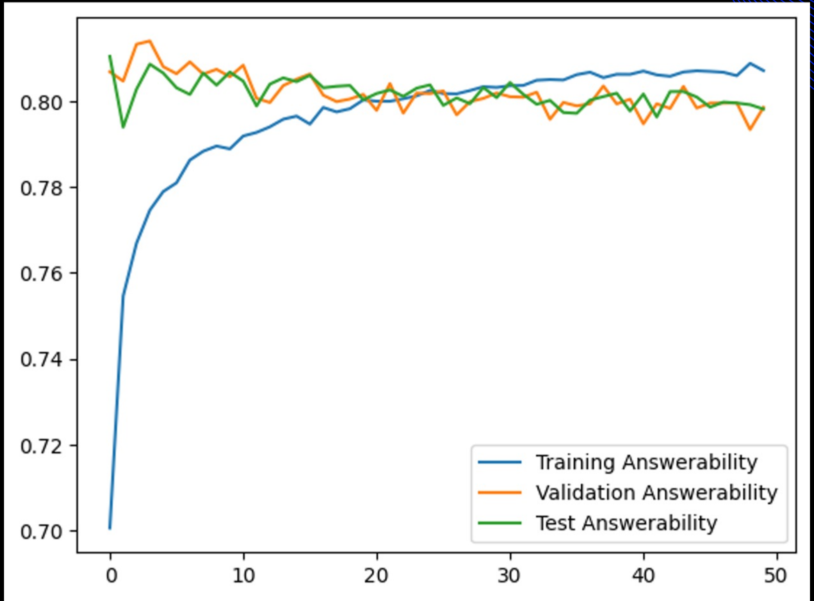
- Model: CLIP
- Optimizer: AdamW
- Learning Rate: 1e-4
- Dropout Rate: 0.5 and 0.5
- Weight Decay: 0



# Findings



Training Loss
Validation Loss
6
5
4
3
2
0
10
20
30
40
50



0.80
0.78
0.76
0.74
0.72
0.70
Training Answerability
Validation Answerability
Test Answerability
0
10
20
30
40
50

# Discussion

**Implications:**

- Improved assistive technology for aiding blind people
- Dismantling accessibility barriers
- Education about needs of blind people

**Future Work:**

- Improve score further
    - Changing linear layers
    - Improving answerability calculations

# Questions?