Multihop and Multimodal Question Answering

Carnegie Mellon University

Deigant Yadava Dheeraj Pai Joao Monteiro Vinay Nair

{dyadava,dmohanda,jmonteir,vinayn}@cs.cmu.edu

Motivation

- **Multihop and Multimodal Question Answering** (MMQA) is the cognitive process of retrieving and combining relevant information from diverse knowledge sources (e.g., textual, visual, and audio) to answer a given question
- While previous approaches do fairly well in retrieving the relevant sources, **alignment** is still a bottleneck for MMQA on WebQA dataset.

SOTA answers "*A water-related object is present in the image*" for the question : "What water related object is present in the image?"

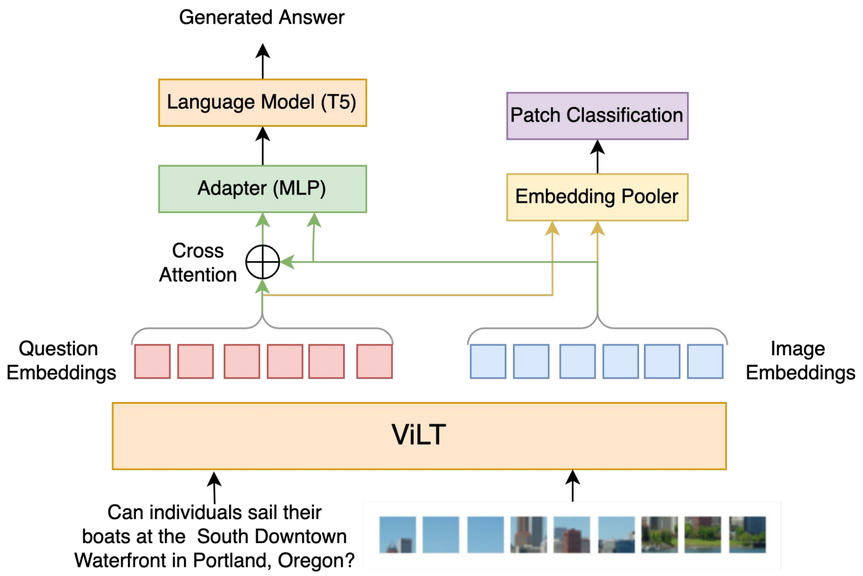
Dataset

WebQA

- Emulates the way humans do web search; aggregate multiple modalities to get a solution.
- Each example has a query, and has a a set of positive sources and distractor sources from which it must extract the answer.
- Evaluation Method:
  - QA quality:
    - Fluency: BARTScore
    - Accuracy: Keywords Overlap

Method

- Key Idea : Train a model to jointly learn the alignment between the question and the image patches as well as to generate the answer.



- ~10% of the dataset was annotated to train the patch classification module.

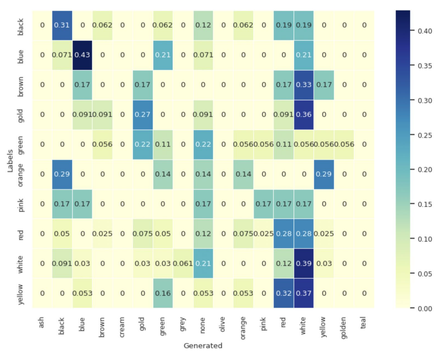
Annotation example

Q: How many strings does a Paraguayan harp have?

- Train the Language model using Log Likelihood Loss and the Patch Classifier using the Cross Entropy Loss

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{V} y_{i,j}\log(p_{i,j})$$

Preliminary Results



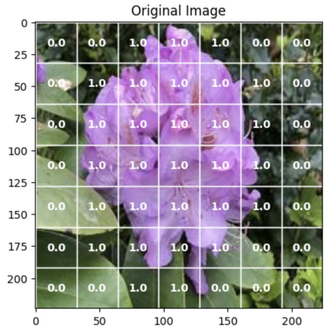
Our Method overfits less on color than Solar (SOTA)

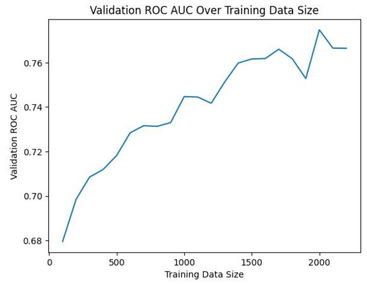
Class	Accuracy (Ours)	Accuracy (Solar)
Others	0.72	0.77
YesNo	0.32	0.49
shape	**0.22**	**0.14**
color	**0.31**	**0.32**
choose	0.75	0.85
number	0.36	0.42

Approximate results for image questions

Analysis and Takeaways

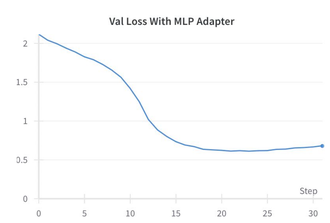
- The performance of the patch classification network scales with the amount of annotated data that is available.

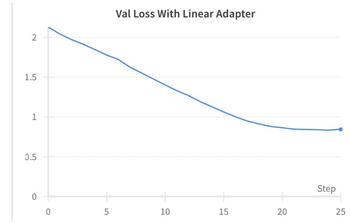




Q: Does a Minnetonka Rhododendron flower have petals in a cup shape?

- A non-linear adapter works better than a linear adapter to convert the embeddings from the space of ViLT to T5





- The T5 model still overfits to the textual data that is being provided by the Questions
  - This could be because the golden answer consists of a lot of redundant words/tokens from the question.
- Better reasoning alone is not enough to solve the WebQA task. **Chain of thought** reasoning using **MiniGPT-4** led to errors when the model did not focus on the right parts/elements of the image

Future Work

- Pre-training T5 with other vision-language related tasks
  - Image Captioning
  - Other VQA datasets
- Using image segmentation techniques instead of patching to find the continuous segments that are most relevant
- Using CLIP Embeddings instead of ViLT embeddings so that the embeddings of the images and the text are better matched.
- Curate the dataset better to remove redundant wording from the golden answer.