# Multimodal Model for Visual Question Answering

Xueyan Li

Supervisor:
Prof Bill Byrne, Weizhe Lin
Special thanks to
Howard Mei, Eric Chen

## Outside Knowledge Visual Question Answering Dataset (OKVQA)

Requires :
- Image understanding
- Language understanding
- Large knowledge resource

Q: Which American president is associated with the stuffed animal seen here?

A: Teddy Roosevelt

**Outside Knowledge**

Another lasting, popular legacy of Roosevelt is the stuffed toy bears—teddy bears—named after him following an incident on a hunting trip in Mississippi in 1902.

Developed apparently simultaneously by toymakers ... and named after President Theodore "Teddy" Roosevelt, the teddy bear became an iconic children's toy, celebrated in story, song, and film.

At the same time in the USA, Morris Michtom created the first teddy bear, after being inspired by a drawing of Theodore "Teddy" Roosevelt with a bear cub.

Current leaderboard

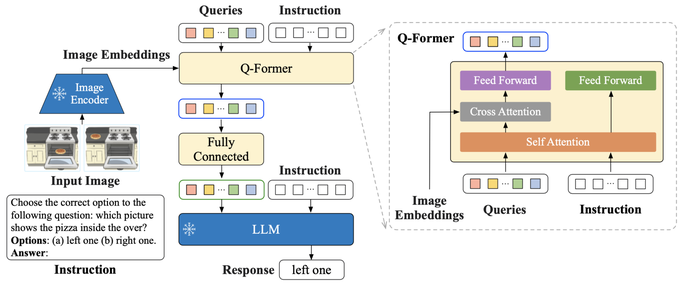
| Model | Accuracy | Method |
|---|---|---|
| PaLM-E-562B | 66.1 | |
| InstructBLIP | 62.1 | |
| Prophet | 61.1 | |
| PromptCap | 60.4 | |
| REVIVE | 58.0 | |
| RA-VQA | 54.48 | |
| Flamingo80B | 50.6 | |
| TRiG | 50.50 | |
| PICa | 48.0 | |
| LaKo | 47.01 | |
| BLIP-2 ViT-G FlanT5 XXL | 45.9 | |

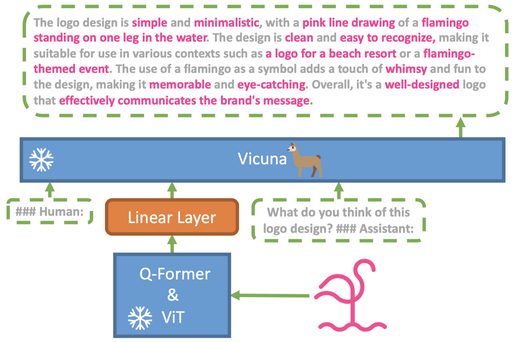
Main methods used
- Visual-language model
- External document retrieval
- Few-shot learning

## Visual Language Models

### InstructBLIP

- Visual information extracted is dependent on instruction
- Different image embedding generated for the same picture with different questions
- FlanT5-XXL and Vicuna-7B are the best performing LLMs for OKVQA



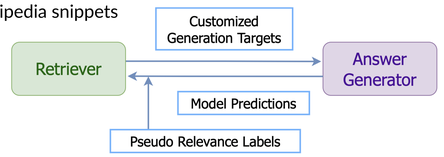

### MiniGPT4

- Uses non-instruction fine-tuned Q-former from BLIP2
- Single linear layer projects image features to LLM embedding
- Freezes Q-Former and ViT
- Uses Vicuna as LLM backbone, which achieves 90% ChatGPT's performance

## External Document Retrieval

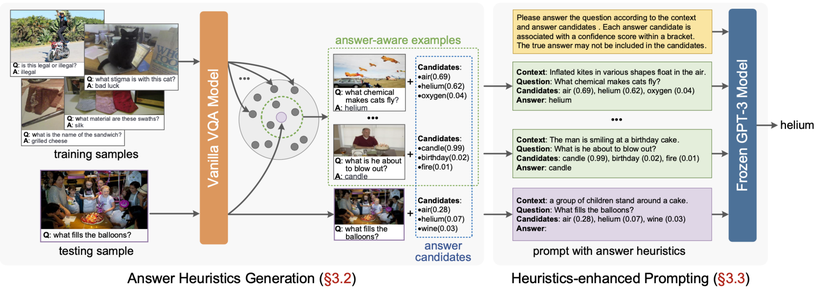
Challenges:
- Need to find relevant documents that contains the correct answer
- Language model needs generate answer based on documents

Sources of knowledge:
- Google search results
- Wikipedia snippets




**RAVQA** (Retrieval Augmented Visual Question Answering)
- Joint training of document retriever and answer generator
- Maximize cosine similarity between a relevant document and text description of image
- Triple loss term that
  - Encourages retrieval of document with correct answer
  - Punishes retrieval of documents without correct answer
  - Adds answer generation loss

## Few Shot In-Context Learning




**Methodology of Prophet**

**Advantages**
- No training required
- Maximally utilize the capability of large language models (GPT3, Vicuna)
- Easily adaptable to other datasets

**Limitations**:
- Some very large models used (GPT3) are not open source
- Replies on language model's implicit knowledge
- Not able to have up to date information/knowledge base
- Precise prompt engineering and ensembling might be required

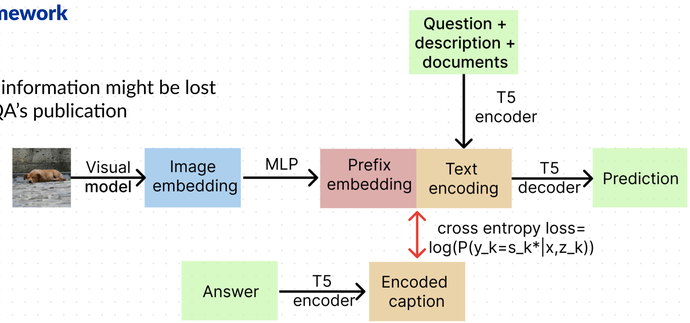
**Possible solution**
- Incorporate document retrieval
- Fine-tune end-to-end with LoRA

## Results

**Goal: Incorporate visual features into RAVQA framework**

**Limitation with RAVQA:**
- Image captioning is done separately and necessary information might be lost
- Better language models have come out since RAVQA's publication

**Solution:**
- Directly add image embedding to text embedding
- An linear layer maps image space to text space
- Use Q-Former pretrained with instruction from Instruct BLIP so that image embedding is question dependent
- Experiment with new language models such as Flan-T5, LLaMA




| Vision model | Language model | Question + descriptions + documents | Question + image prefix | Accuracy |
|---|---|---|---|---|
| None | T5-large | ✗ | ✗ | 26.85 |
| | | ✓ | ✗ | 51.63 |
| | Flan-T5-large | ✗ | ✗ | 28 |
| | | ✓ | ✗ | 54.23 |
| Q-Former | T5-large | ✗ | ✓ | No improvement |
| | | ✓ | ✓ | |
| | Flan-T5-large | ✗ | ✓ | |
| | | ✓ | ✓ | |
| ViT | T5-large | ✗ | ✓ | 27.11 |
| | | ✓ | ✓ | 52.09 |
| | Flan-T5-large | ✗ | ✓ | **36.71** |
| | | ✓ | ✓ | 54.22 |

**Results**
- Output of ViT trained from CLIP improves performance of Flan-T5-Large by 8%
- However, the same ViT does not improve T5-Large
- Q-Former does not bring any improvement despite being a better vision model

Conclusion so far...

**Under-defined visual features will be ignored by language model! Pre-training required!**

## Future work

- Pretrain Q-Former linear layer with Conceptual Captions or VQA 2.0
- Use InstructBLIP checkpoint to get image embedding with question as input
- Experiment with fine-tuning very large language models (>3B parameters) end-to-end with LoRA
- Few-shot in-context learning with frozen large language models that incorporate external documents retrieved

## References

1. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge
2. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning
3. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models
4. Prompting Large Language Models with Answer Heuristics for Knowledge-based Visual Question Answering
5. Retrieval Augmented Visual Question Answering with Outside Knowledge