Visual Question Answering over Text, Tables & Images

University of Southern California

Group 50 - Baladitya Swaika, Bowen Shi, Henil Shelat, Prashant Vibhor Agarwal, Sunit Ashish Vaidya

USC Viterbi School of Engineering

INTRODUCTION

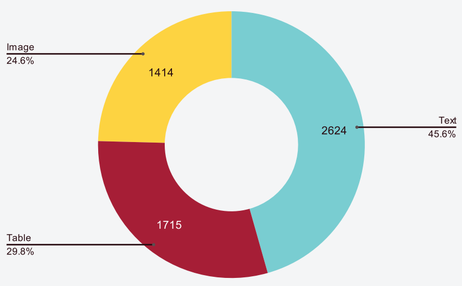
Motivation

- Current state-of-the-art models in VQA use deep learning models that **do not rely on common sense reasoning**
- Leverage **multi-modal knowledge sources** to answer complex questions with the help of **knowledge graphs**

Dataset

- **MMConvQA** - Passages + Tables + Images

Questions
5735

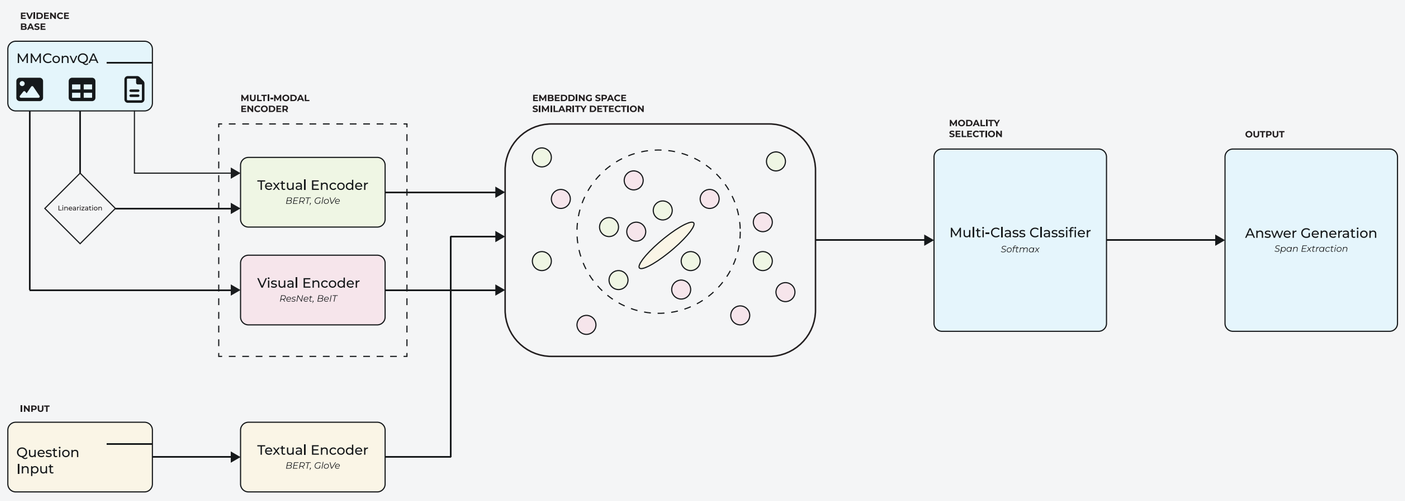


METHODOLOGY

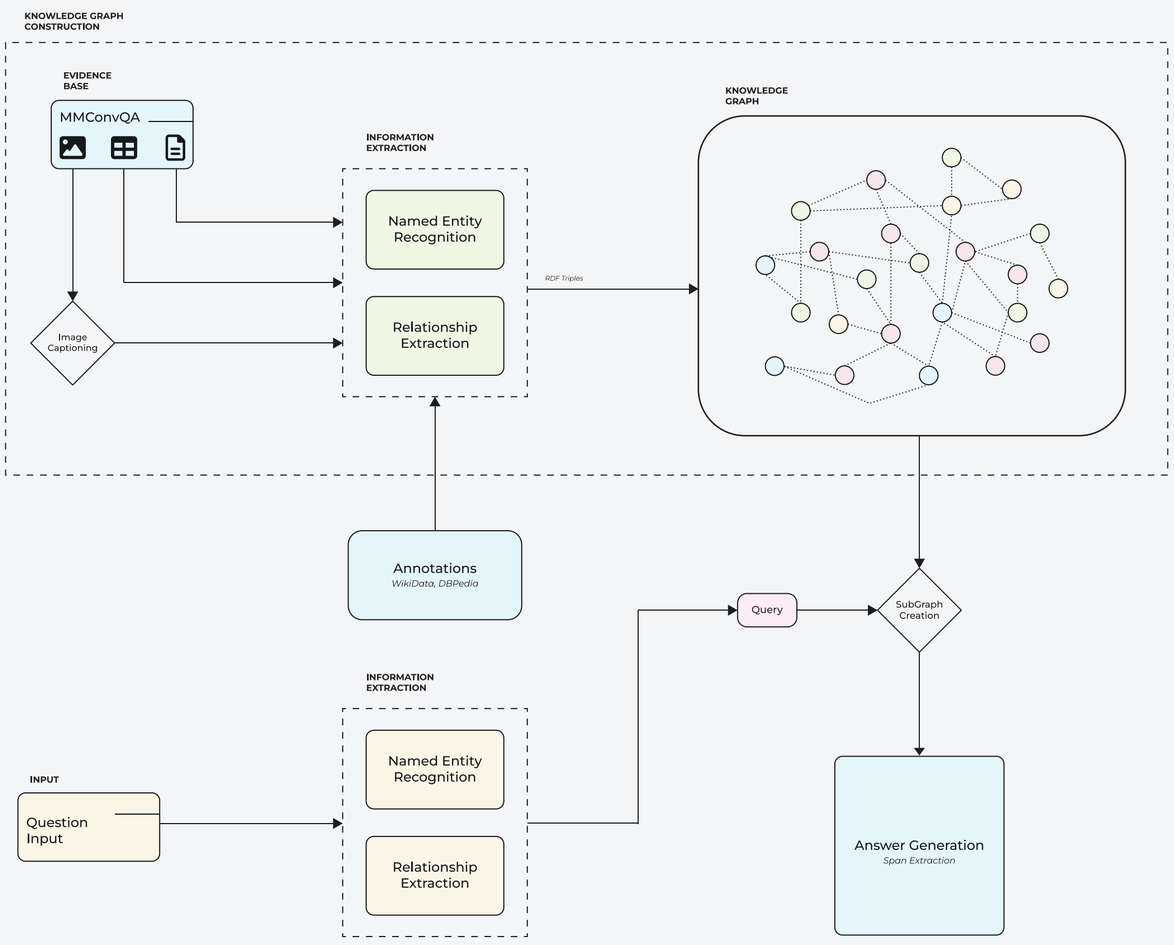
- We use the Multi-modal Conversational QA system with Adaptive Extractors as our baseline for multi-modal QA tasks. Its major components are question understanding, multi-modal evidence retrieval and adaptive answer extraction.
- Textual and image embeddings were created and trained on gold questions and related data from the knowledge graph. The question representations were obtained by encoding questions using the pre-trained BERT model with masked language modeling objective.
- Baseline model uses various encoders to transform multi-model evidence into unified knowledge space and generates correct answers using extraction and modality detection for different sources based on their ranks.
- To improve on the baseline model, we created a knowledge graph from multi-modal evidence sources and augmented it in place of the modality detection module. We provide the node embeddings with question embeddings to generate natural language spans.

MODELS

Baseline



KG-Based

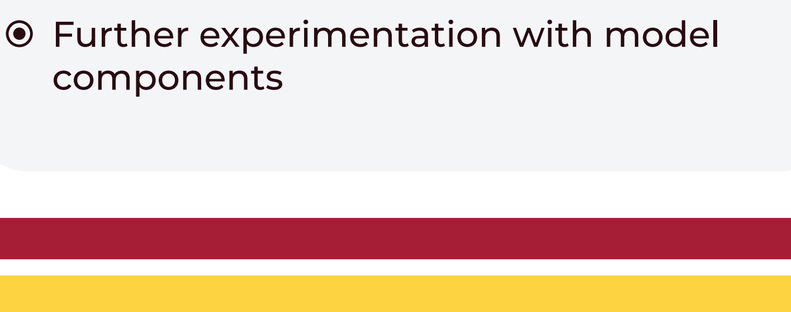


RESULTS

- Word-level F1 score is used to measure the accuracy of generated answers with reference answers

Model	F1-Score
Baseline	0.305
KG-Based	0.364

- Modality analysis on performance



FUTURE WORK

- Incorporate more modalities like video and speech
- More training could help improve performance
- Further experimentation with model components