Visual Question Answering (VQA)

Team Members:

- Onkar Kunte
- Shashidhar Reddy
- Jianlin Lin

Yeshiva University®

Introduction

Research Focus: The study explores challenges in Visual Question Answering (VQA), advancing from static image recognition to multi-modal problem-solving by integrating visual and textual data.

Scope and Approach: It targets complex educational content, particularly machine learning lectures, by creating a specialized dataset of lecture materials with open-ended question-answer pairs.

Impact: The tailored VQA models enhance comprehension of intricate educational topics and advance academic question-answering systems.

Objectives

Dataset and Model Development: Create a comprehensive dataset and develop advanced multimodal VQA models tailored to analyze textual and visual data effectively.

Addressing Educational Challenges: Enhance VQA systems to handle complex educational content, improving comprehension of intricate materials like deep learning lectures.

Advancing AI Research: Contribute to the field of artificial intelligence by exploring innovative techniques for integrating visual and linguistic data, with a focus on challenging academic scenarios.

Data Creation

1. **Image Collection:**
    - Extracted images from **‘Deep Learning’** lecture slides.
    - Ensured consistency by resizing all images to 224×224224 times 224 pixels.
2. **Transcript Collection:**
    - Used advanced tools like **Deepgram** and **Google Speech-to-Text API** for audio transcription.
    - Manually reviewed and corrected transcription errors due to noisy recordings.
3. **Query and Response Formulation:**
    - Generated 10 question-answer pairs based on slide content.
    - Created contextual prompts, ensuring relevance and completeness.


```
{
  "instruction": "Are there any further questions allowed after the introduction?",
  "context": "Week 11, Post-Introduction Q&A",
  "response": "Yes, after the introduction, there is an opportunity for students to ask further questions.",
  "category": "closed_qa",
  "week": 11,
  "page": 1
},
{
  "instruction": "What does the instructor plan to do after the introduction?",
  "context": "Week 11, Lecture Agenda",
  "response": "The instructor plans to proceed with the study and the test lecture after the introduction.",
  "category": "information_extraction",
  "week": 11,
  "page": 1
},
```

Data Preprocessing

Data Integration and Organization: Merged datasets of question-answer pairs, transcripts, and images were combined using attributes like week and page numbers. Columns were renamed for clarity, such as changing "instruction" to "question" and "response" to "answer."

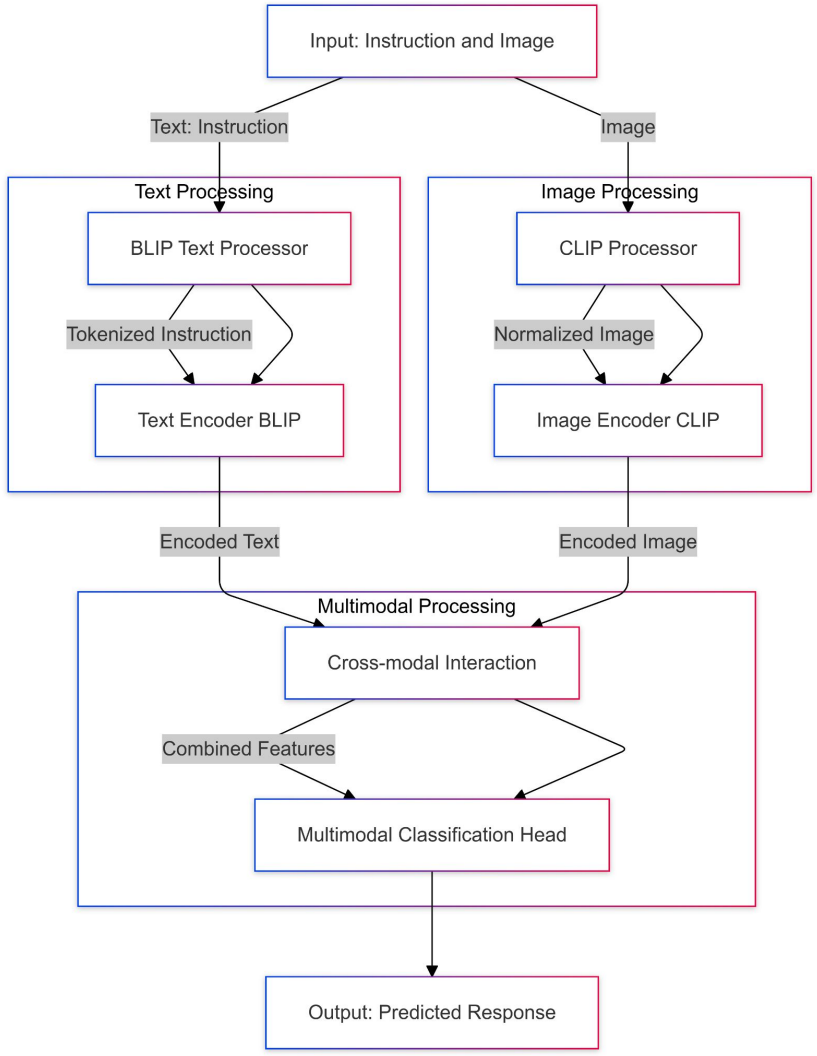
Dataset Partitioning: The data was split into training (90%, 8,681 samples) and validation (10%) sets to enable model training and performance evaluation.

Dataset Preparation: Images were resized to 224x224 pixels, textual prompts were created by combining questions and transcripts, and text exceeding model token limits was trimmed. Responses were formatted to align with the model's requirements for training and validation.

Methodology: BLIP-CLIP

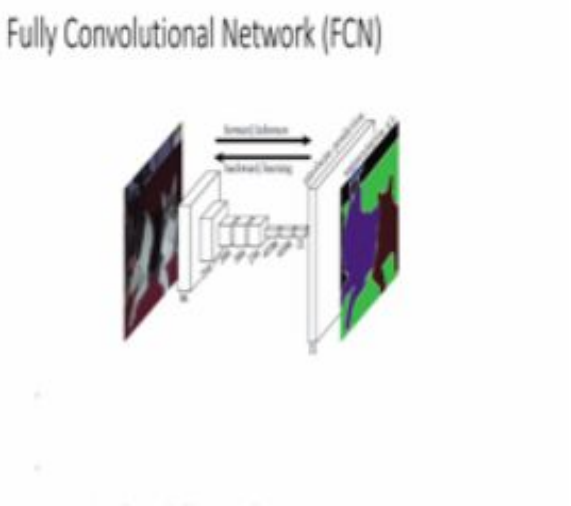
Single Slide Points: How Our Model Works

1. **Input**:
   - Takes both **Instruction (text)** and **Image** as inputs.
2. **Text Processing**:
   - Uses **BLIP Text Processor** for tokenization and normalization.
   - Encodes the instruction with the **BLIP Text Encoder** to generate text features.
3. **Image Processing**:
   - Processes the image with the **CLIP Processor** (resizing and normalization).
   - Encodes the image with the **CLIP Image Encoder** to generate image features.
4. **Multimodal Fusion**:
   - Combines features from text and image through a **Cross-modal Interaction Module** for deeper understanding.
5. **Output Generation**:
   - Features are passed to a **Multimodal Classification Head** to produce the final **Predicted Response**.
6. **Highlight**:
   - Leverages both BLIP and CLIP models for robust text-image understanding and response generation.



Results

```
Question:  what is an fcn and what does it do?
Predicted Answer:  an fcn ( fcn ) can be adapted for deeper convolutional layers by
Actual Answer:  a fully convolutional network ( fcn ) is a type of neural network designed for semantic image segm
entation. it processes images on a
```



Results

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU score	Cosine
BLIP	0.2350	0.0408	0.1884	0.0138	0.3414
BLIP-CLIP	0.393	0.19137	0.34126	0.1207	0.42

Future work

- Interactive UI
- Get more accuracy using LLM

Thank you!