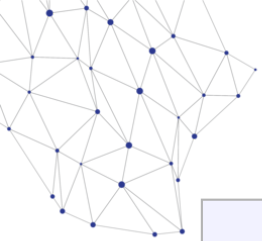


Visual Question Answering

Transforming Visual Data into Knowledge:

Taha Bouhafa

Loubaba Malki L'Hlaibi

2nd Year Big Data and Artificial Intelligence

27/11/2024

Prof. Dr.Belcaid Anass

1. Introduction to Visual Question Answering :

Definition:

Visual Question Answering (VQA) is a task where an AI system is given an image and a question. The system analyzes the image and generates an accurate answer by understanding both the visual content and the language of the question.

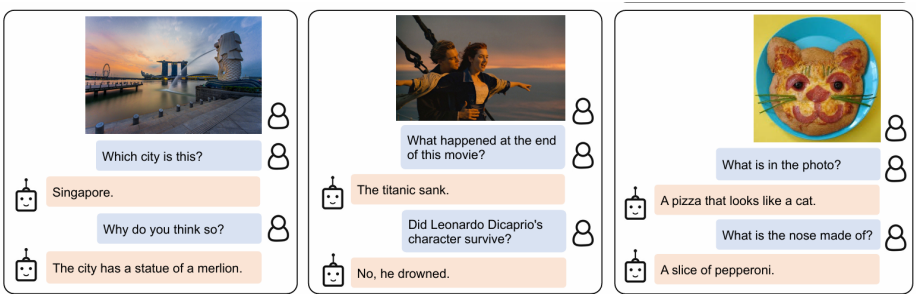


Figure 1: : Examples of questions and answers in VQA tasks, showcasing different question types.

1. Introduction to Visual Question Answering :

Table 1: Computer Vision subtasks required to be solved by VQA

CV Task	Representative VQA Question
Object recognition	What is in the image?
Object detection	Are there any dogs in the picture?
Attribute classification	What color is the umbrella?
Scene classification	Is it raining?
Counting	How many people are there in the image?
Activity recognition	Is the child crying?
Spatial relationships among objects	What is between the cat and the sofa?
Commonsense reasoning	Does this person have 20/20 vision?
Knowledge-base reasoning	Is this a vegetarian pizza?

2. Applications :

2. Applications :

- **Medical VQA:**

Organ System

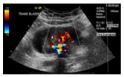
Q: What is the organ system?
A: Gastrointestinal

Object/Condition Presence

Q: Is there gastric fullness?
A: yes

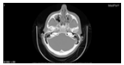
Positional

Q: What is the location of the mass?
A: head of the pancreas



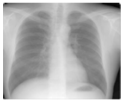
Modality

Q: what imaging method was used?
A: us-d - doppler ultrasound



Plane

Q: which plane is the image shown in?
A: axial



Q: Airspace opacity?
A: Yes
Q: Fracture?
A: Not in report

Q: Lung lesion?
A: No
Q: Pneumonia?
A: Yes

Figure 2: Samples of images and question-answer pairs.

Source: Medical Visual Question Answering: A Survey. (2023)

2. Applications :

- **VQA in Video Surveillance scenarios:**

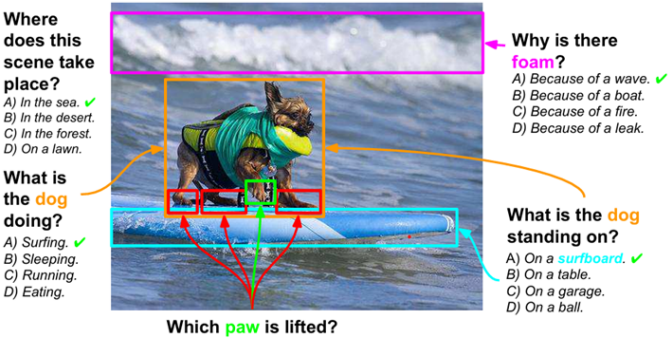


Figure 3: Diverse questions to acquire detailed information on images

Source: *Visual7W: Grounded Question Answering in Images. (2016)*

2. Applications :

- **VQA and Advertising:**

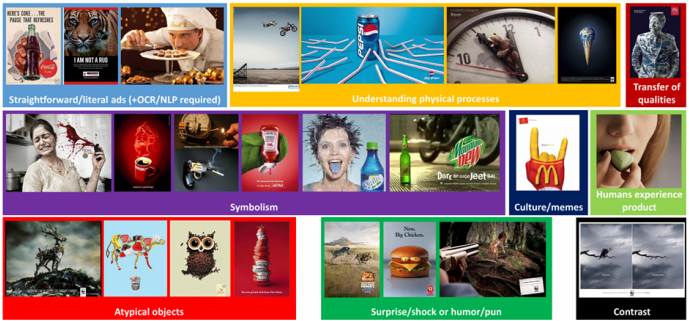


Figure 4: Examples of ads grouped by strategy or visual understanding required for decoding the ad.

Source: *Automatic Understanding of Image and Video Advertisements. (2017)*

3. Datasets for VQA :

- **COCO VQA V2.0 2017 Dataset**:
  https://visualqa.org/index.html
- **Visual7W**(Visual Genome Questions):
  https://ai.stanford.edu/ yukez/visual7w/
- **CLEVR**(Compositional Language and Elementary Visual Reasoning):
  https://cs.stanford.edu/people/jcjohns/clevr/

COCO VQA V2.0 2017 Dataset

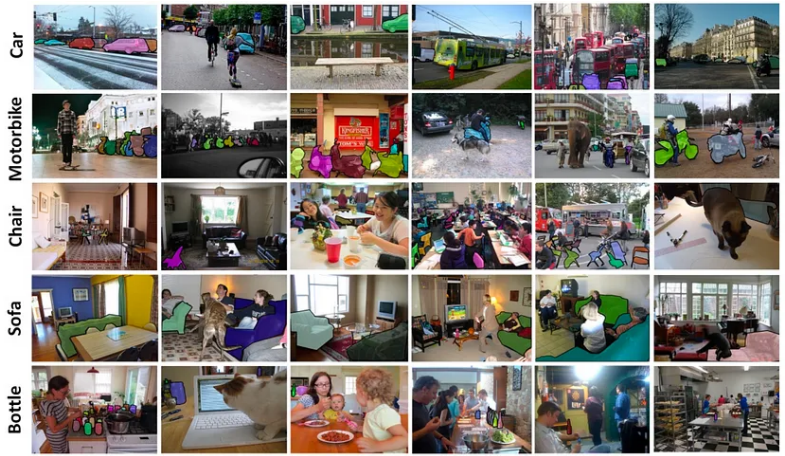


Figure 5: : COCO VQA V2.0 Dataset Overview

COCO VQA V2.0 2017 Dataset

- **Description:** Visual Question Answering (VQA) v2.0 is a dataset containing open-ended questions about images. These questions require an understanding of vision, language, and commonsense knowledge to answer. It is the second version of the VQA dataset.

Images	Questions	Annotations	Questions per Image	Split (Training/Testing)
250,000+	1M+	7M+	5.4 questions average	82,000+ / 80,000+

Table 2: Statistics of the Visual7W Dataset

COCO VQA V2.0: Examples

Figure 6: Examples of images from the dataset with questions and answers.

Visual7W Dataset

Q: What endangered animal is featured on the truck?

A: **A bald eagle.**
A: A sparrow.
A: A humming bird.
A: A raven.

Q: Where will the driver go if turning right?

A: **Onto 24 ¾ Rd.**
A: Onto 25 ¾ Rd.
A: Onto 23 ¾ Rd.
A: Onto Main Street.

Q: When was the picture taken?

A: **During a wedding.**
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church service.

Figure 7: : Visual7W Dataset Overview

Visual7W Dataset

- **Description:** Visual7W is a large-scale visual question answering (VQA) dataset that includes object-level groundings and multimodal answers. Each question in the dataset begins with one of the seven Ws (what, where, when, who, why, how, and which), designed to test a model's understanding of various types of inquiries.

Images	QA pairs	Object Groundings	Objects	Split (Training/Testing)
47.000+	327.000+	561.000+	36.000+	42.000+ / 5.000+

Table 3 : Statistics of the COCO VQA V2.0 Dataset

Visual7W : Examples

Q: Who is under the umbrella?

A: Two women.
A: A child.
A: An old man.
A: A husband and a wife.

Q: Why was the hand of the woman over the left shoulder of the man?

A: They were together and engaging in affection.
A: The woman was trying to get the man's attention.

Q: How many magnets are on the bottom of the fridge?

A: 5.
A: 2.
A: 3.
A: 4.

Figure 8: Examples of images from the dataset with questions and answers.

CLEVR Dataset

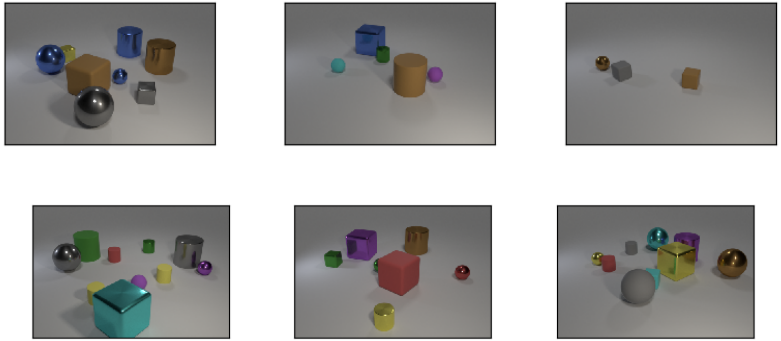


Figure 9: CLEVR Dataset Overview

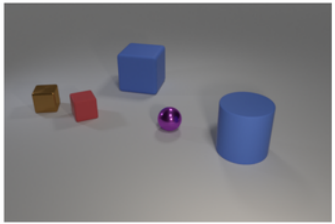
CLEVR Dataset

- **Description:** A synthetic dataset aimed at testing a model's compositional and logical reasoning skills.Questions involve reasoning about shapes, colors,sizes,and spatial relationships in 3D-rendered scenes.

Category	Images	QA pairs	Split(Training/testing)
CLEVR	70.000+	700.000+	60.000+/10.000+

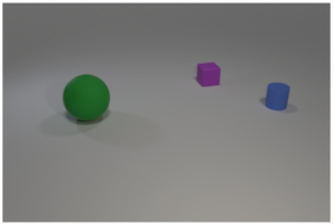
Table 4 : Statistics of CLEVR Dataset

CLEVR: Examples



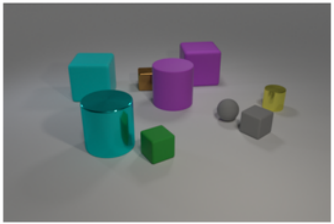
Q: Are there more brown shiny objects behind the large rubber cylinder than gray blocks?
A: yes

Q: What color is the matte object to the right of the large block?
A: blue



Q: Are there any other things that are the same shape as the large green thing?
A: no

Q: The matte thing that is both in front of the purple cube and to the left of the blue rubber cylinder is what color?
A: green



Q: What number of cubes are the same color as the rubber ball?
A: 1

Q: Is the big cyan metal thing the same shape as the brown thing?
A: no

Figure 10: Examples of images from the CLEVR Dataset with questions and answers.

4. Evaluation :

4. Evaluation :

Definition:

An answer is deemed 100% accurate if at least 3 workers provided that exact answer.

$$\text{Acc(ans)} = \min\left(1, \frac{\#\text{humans provided ans}}{3}\right)$$

4. Evaluation :

Definition:

An answer is deemed 100% accurate if at least 3 workers provided that exact answer.

$$\text{Acc}(\text{ans}) = \min\left(1, \frac{\#\text{humans provided ans}}{3}\right)$$

Example: A robotic arm is assembling a car part.

- Human-provided answers:
  - Automation (5 times)
  - Programming (3 times)
  - Sensors (2 times)
- Predicted answer: Automation
- Accuracy (Automation):
  $\text{Acc}(\text{Automation}) = \min\left(1, \frac{5}{3}\right) = 1.0$

4. State-of-the-Art Algorithms :

4. State-of-the-Art Algorithms :

Table 5: Performance Analysis of VLP architectures in VQA. The models are evaluated on the test-dev and test-std split of the VQAv2 dataset.

Model Name	Test-Dev	Test-std
PaLI	84.30	84.30
BeiT-3	84.20	84.00
VLMO	82.88	82.78
One-Peace	82.60	82.50
BLIP-2	82.30	82.30
Flamingo	82.00	82.10
OFA	82.00	82.00

Source: *From Image to Language: A Critical Analysis of Visual Question Answering (VQA) Approaches, Challenges, and Opportunities. (2024)*

BeiT-3 :

BeiT-3 :

- **Architecture:**

BeiT-3 :

- **Architecture:**

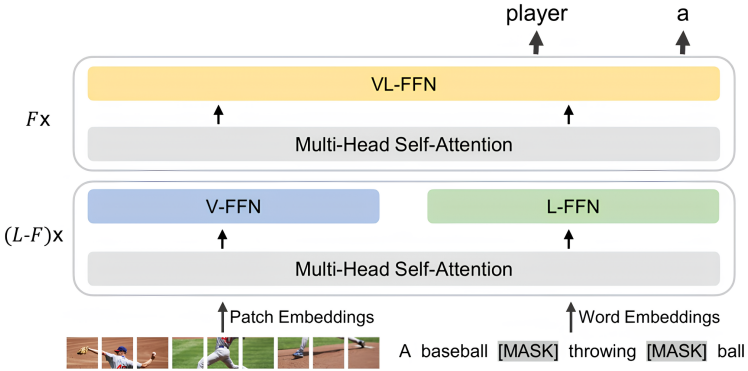


Figure 12: Fusion Encoder for Masked Vision-Language Modeling and Vision-Language Tasks

Source: *Wang et al., Image as a Foreign Language: BEIT Pretraining. (2022)*

BeiT-3 :

BeiT-3 :

- **How is BeiT-3 used in VQA tasks:**

BeiT-3 :

- **How is BeiT-3 used in VQA tasks:**
  - **Multiway Transformer for Joint Encoding:** A transformer-based architecture that jointly processes image and question embeddings to model interactions between the modalities.

BeiT-3 :

- **How is BeiT-3 used in VQA tasks:**
  - **Multiway Transformer for Joint Encoding:** A transformer-based architecture that jointly processes image and question embeddings to model interactions between the modalities.
  - **Multimodal Embedding:** Visual features of the image and semantic embeddings of the question are extracted, then concatenated into a unified vector representation for the Multiway Transformer.

BeiT-3 :

- **How is BeiT-3 used in VQA tasks:**
  - **Multiway Transformer for Joint Encoding:** A transformer-based architecture that jointly processes image and question embeddings to model interactions between the modalities.
  - **Multimodal Embedding:** Visual features of the image and semantic embeddings of the question are extracted, then concatenated into a unified vector representation for the Multiway Transformer.
  - **Classifier for Answer Prediction:** A classification layer predicts the answer based on the output of the Multiway Transformer.

BeiT-3 :

- **How is BeiT-3 used in VQA tasks:**
  - **Multiway Transformer for Joint Encoding:** A transformer-based architecture that jointly processes image and question embeddings to model interactions between the modalities.
  - **Multimodal Embedding:** Visual features of the image and semantic embeddings of the question are extracted, then concatenated into a unified vector representation for the Multiway Transformer.
  - **Classifier for Answer Prediction:** A classification layer predicts the answer based on the output of the Multiway Transformer.
  - **End-to-End Fine-tuning:** The model is fine-tuned on the VQA v2.0 dataset as a classification problem, optimizing for the most frequent 3129 answer candidates.

BLIP-2 :

BLIP-2 :

- **Architecture:**

BLIP-2 :

- **Architecture:**

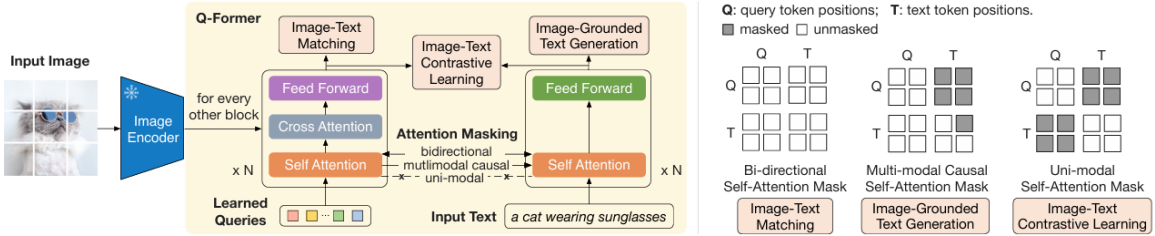


Figure 11: (**Left**) Model architecture and first-stage objectives jointly optimize learnable query embeddings to extract visual features relevant to text. (**Right**) The self-attention masking strategy for each objective to control query-text interaction.

Source: *Li et al., Salesforce (2023).*

BLIP-2 :

BLIP-2 :

- **How is BLIP-2 used in VQA tasks:**

BLIP-2 :

- **How is BLIP-2 used in VQA tasks:**
  - **Q-Former for Feature Extraction:** Extracts visual features from frozen image encoders using 32 learnable queries to focus on task-relevant information.

BLIP-2 :

- **How is BLIP-2 used in VQA tasks:**
  - **Q-Former for Feature Extraction:** Extracts visual features from frozen image encoders using 32 learnable queries to focus on task-relevant information.
  - **Dual Transformer Design:** Utilizes an Image Transformer for visual feature extraction via cross-attention and a Text Transformer for encoding and decoding text, with the output compressed into a fixed-size vector (32×768) for efficient reasoning.

BLIP-2 :

- **How is BLIP-2 used in VQA tasks:**
  - **Q-Former for Feature Extraction:** Extracts visual features from frozen image encoders using 32 learnable queries to focus on task-relevant information.
  - **Dual Transformer Design:** Utilizes an Image Transformer for visual feature extraction via cross-attention and a Text Transformer for encoding and decoding text, with the output compressed into a fixed-size vector (32×768) for efficient reasoning.
  - **Classifier for Answer Prediction:** Optimized on VQA tasks to align extracted visual features with textual inputs.

BLIP-2 :

- **How is BLIP-2 used in VQA tasks:**
  - **Q-Former for Feature Extraction:** Extracts visual features from frozen image encoders using 32 learnable queries to focus on task-relevant information.
  - **Dual Transformer Design:** Utilizes an Image Transformer for visual feature extraction via cross-attention and a Text Transformer for encoding and decoding text, with the output compressed into a fixed-size vector (32×768) for efficient reasoning.
  - **Classifier for Answer Prediction:** Optimized on VQA tasks to align extracted visual features with textual inputs.
  - **End-to-End Fine-tuning:** Optimized on VQA tasks to align extracted visual features with textual inputs.

5. Challenges and Future Directions :

- **Challenges:**

5. Challenges and Future Directions :

- **Challenges:**
  - **Dataset Size Limitations:** Despite growth, VQA datasets still lack the volume and diversity needed for robust model training. Larger datasets are crucial for better accuracy and generalization.

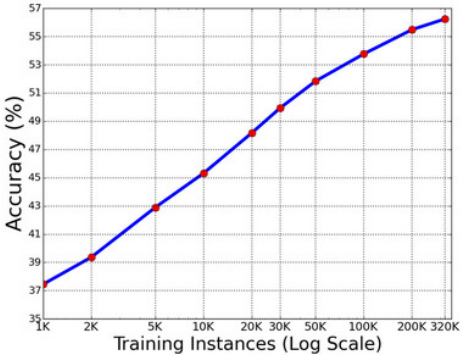


Figure 13: Graph showing the impact of dataset size on model accuracy for VQA.

5. Challenges and Future Directions :

- **Challenges:**
  - **Language Bias in VQA:** A key challenge in VQA is language bias, where models prioritize linguistic patterns over image content, ignoring visual cues. This reduces robustness, accuracy, and interpretability, hindering effective VQA performance.

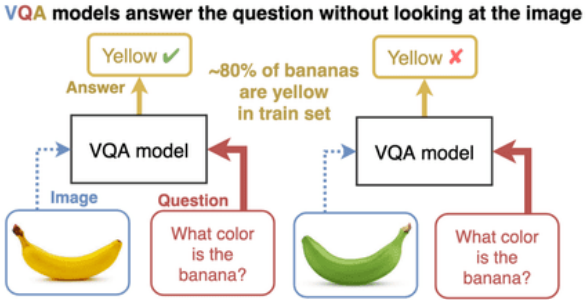


Figure 14: Example of language bias in VQA.

5. Challenges and Future Directions :

- **Future Directions:**

5. Challenges and Future Directions :

- **Future Directions:**
  - **Reducing Unimodal Biases in VQA:** Future models should integrate better multimodal learning techniques to balance the use of visual and linguistic information, minimizing over-reliance on a single modality.
  - **Multimodal Temporal Reasoning for Video Question Answering (VQA):** Expanding VQA to video content introduces Video Question Answering (Video VQA), which requires models to incorporate temporal reasoning,analyzing event sequences and how visual and textual features change over time.