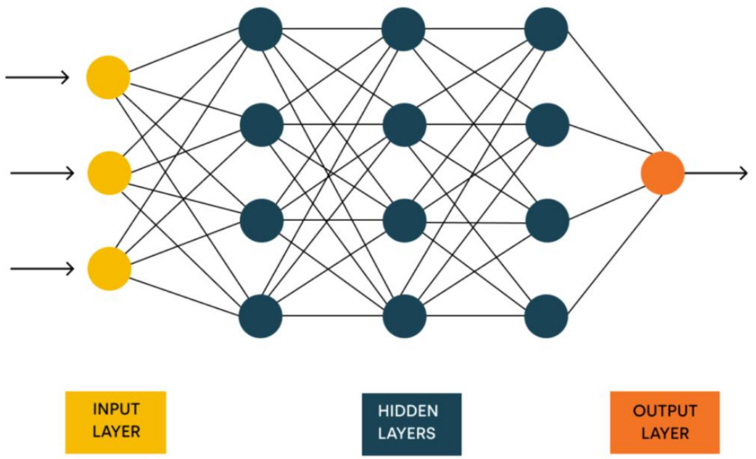
Enhance Visual Reasoning in VQA Tasks through Architectural Choices and Training Strategies

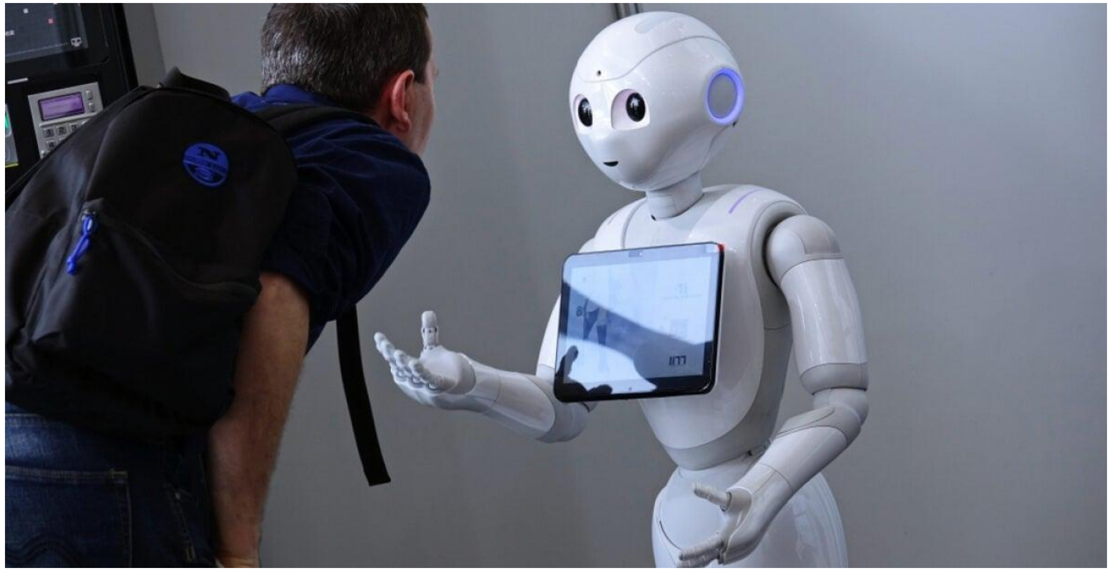
Xiumei Xue

Jan 17, 2025

Motivation

- Current neural network models struggle with tasks requiring spatial reasoning or counting. They often rely on *pattern recognition* rather than true spatial understanding.
- Given this, I am somewhat skeptical of the research path based on the inherent, restrictive approach and am more drawn to exploring interactive learning.
- *Interactive learning* is a paradigm in artificial intelligence where agents learn to perform tasks through interactions with a teacher or environment. This approach is particularly relevant in dynamic or unforeseen context.



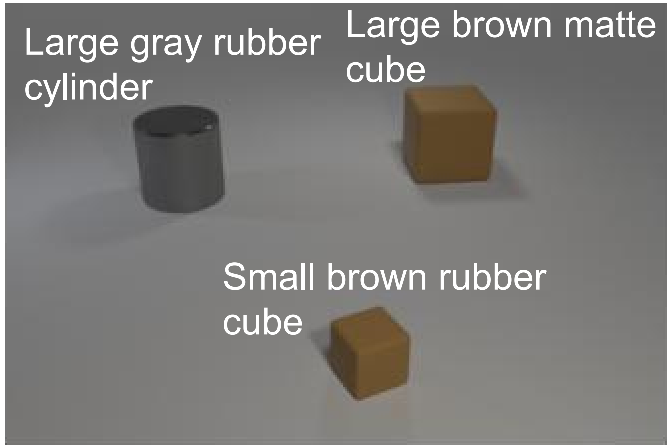


Related Work & Datasets

- **Visual Question Answering (VQA)** is a multimodal task that bridges computer vision and natural language processing, requiring models to reason over both visual and linguistic inputs.

CLEVR (Johnson et al., 2016)

- a diagnostic dataset for compositional language and elementary visual reasoning by providing a series of synthetic 3D rendered images and corresponding complex questions
- helps evaluate VQA models on multi-dimensional benchmarks such as *Question Type*, *Relation Type*, *Question Size*, *Spatial Reasoning*, *Compositional Generalization*, *etc*
- ensures data diversity and balance, while also providing detailed question parsing, allowing researchers to deeply analyze the model's reasoning process



Q: There is a rubber cube in front of the big cylinder in front of the big brown matte thing; what is its size?
A: small
Q-type: query_size
Size: 14

Q: What color is the object that is on the left side of the small rubber thing?
A: gray
Q-type: query_color
Size: 7

Example image and QA pairs in CLEVR

Look into CLEVR

- consists of 3 parts: images, questions, scenes
- images are annotated with ground-truth object positions and attributes (scene graphs)
- questions are represented as *functional programs* that can be executed to answer the question, such as querying object attributes, counting sets of objects, or comparing values

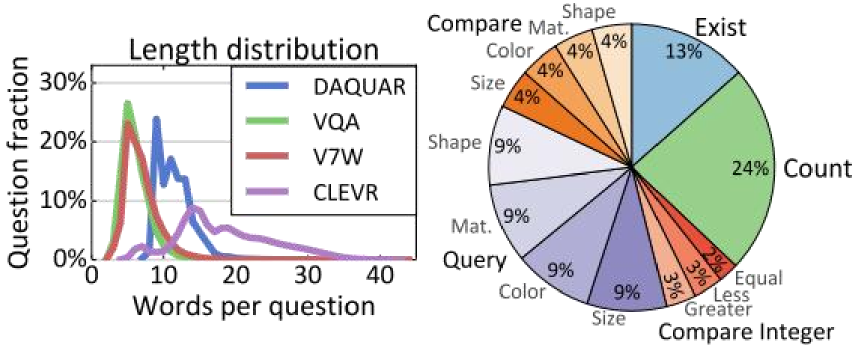

```
{'image_index': 0, 'program': [{'inputs': [], 'function':
'scene', 'value_inputs': []}, {'inputs': [0], 'function':
'filter_size', 'value_inputs': ['large']}, {'inputs': [1],
'function': 'filter_material', 'value_inputs': ['metal']},
{'inputs': [2], 'function': 'unique', 'value_inputs': []},
{'inputs': [3], 'function': 'same_shape', 'value_inputs':
[]}, {'inputs': [4], 'function': 'exist', 'value_inputs':
[]}], 'question_index': 0, 'image_filename':
'CLEVR_val_000000.png', 'question_family_index': 39, 'split':
'val', 'answer': 'no', 'question': 'Are there any other
things that are the same shape as the big metallic object?'}
```


Sample item in *CLEVR_val_questions.json*

**program_length_range 2~25*

Split	Images	Questions	Unique questions	Overlap with train
Total	100,000	999,968	853,554	-
Train	70,000	699,989	608,607	-
Val	15,000	149,991	140,448	17,338
Test	15,000	149,988	140,352	17,335

Statistics for CLEVR



Left - CLEVR questions are generally much longer.
Right - Distribution of question types in CLEVR.

Methodology

- Start from a baseline model, which utilized *LSTM* model to process questions and answers, *VGG16* to extract image features
  => the prelimnary accuracy is around *3%*
- Upgrade the baseline model structure, use *BERT* to encode text and *ResNet-50* to process images
- How to be “interactive”?

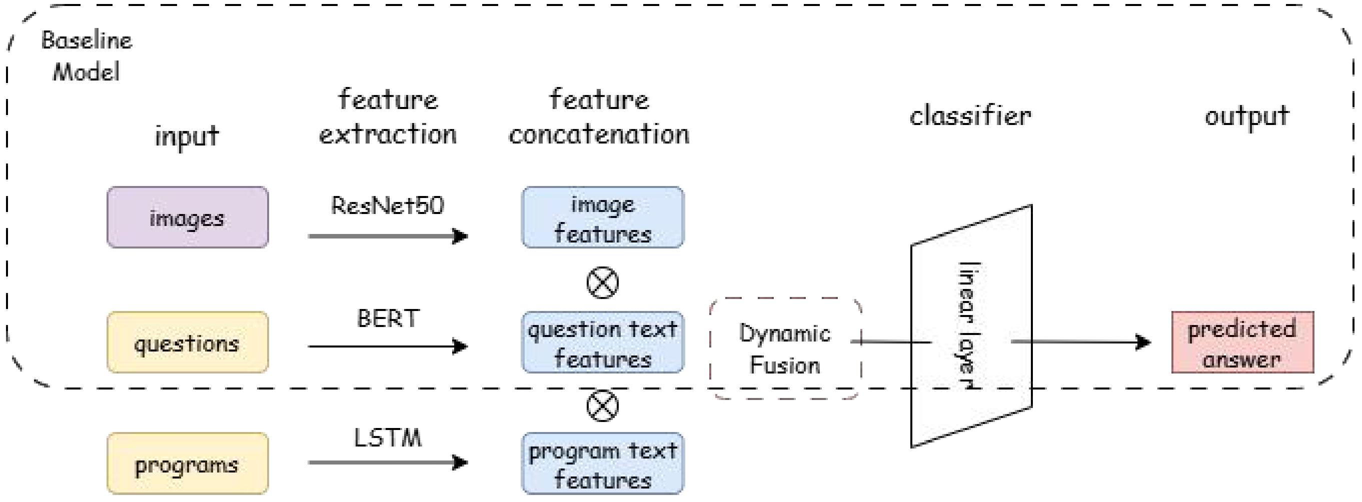
- provide human feedback to realize an *incremental*, *online* learning
- instead of pure correct answer, provide corrections with additional information

- introduce *program* attribute, similar to the concept of *Chain-of-Thinking* (CoT)
- dynamic fusion with multi-head attention
- try *online learning*

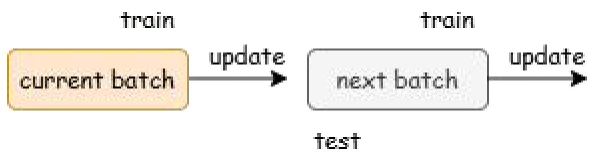
Core challenges: the mismatch between batch-encoded inputs and item-specific human corrections in natural language => to collect human feedback is effort-consuming

Current approach

- AdamW Optimizer, CrossEntropy Loss
- 3 epochs, learning rate=1e-4
- exprimented with $max\_program\_length$ = 3 and 5
- *online learning* with program-augmented model



Model Architecture



sketch map of online learning

**drawn by draw.io*

Results

Performance comparison of different models and subsets

Model	max_program_length	Train Accuracy (%)	Val Accuracy (%)
Baseline Model	3	89.10	48.79
Enhanced With Program	3	95.77	47.57
Online Learning	3	/	62.50
Baseline Model	5	26.04	18.41
Enhanced With Program	5	60.99	46.67
Online Learning	5	/	68.75

a) By controlling the program length, it becomes evident that *performance degrades as the questions become more complex*. This suggests that the model struggles with intricate queries, highlighting a limitation in handling higher levels of abstraction or detail.

b) *Programs serve as valuable additional information*, particularly for relatively complex questions. Though both training and validation accuracies decrease when *max_program_length* is set as 5, the program-enhanced model still demonstrates a significant improvement over the baseline.

c) The generalization capability of model is limited. The results reveal a notable gap between training and validation accuracy, which points to potential *over-fitting*. This discrepancy underscores the model's limited ability to generalize to unseen data.

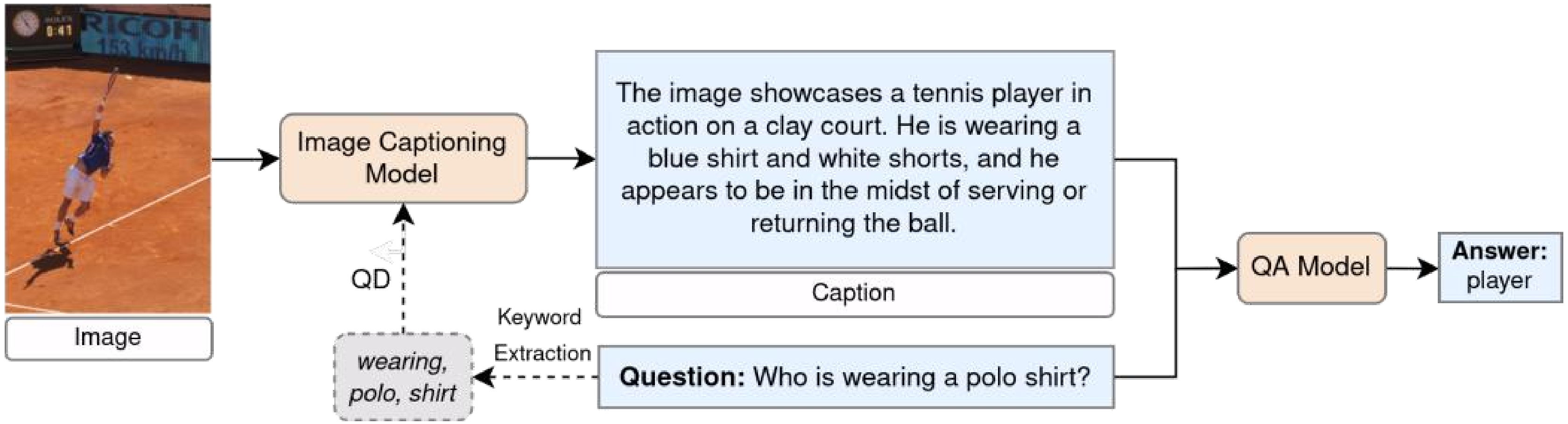
d) While the results of online learning appear promising, they exhibit a degree of *randomness*. This suggests that the online learning process may lack stability or consistency, calling further investigation into its reliability and effectiveness.

Discussion

- **Limitation**
  - The CLEVR dataset is synthetic, making it difficult to generalize to real-world scenarios. Additionally, most real questions lack programming annotations, which limits its applicability.
  - The use of pre-trained models can be computationally expensive, and the current model architectures and hyperparameter configurations may not be the most advanced.
  - The current analysis is limited in scope, focusing primarily on overall performance without considering different task types or dimensions.
- **Future Work**
  - Validate the model on hybrid datasets that combine synthetic and real-world data. This will help assess the model's ability to generalize beyond synthetic environments.
  - Experiment with more efficient pre-trained models (e.g., RoBERTa) and explore state-of-the-art architectures with optimial hyperparameters.
  - Error Analysis-Interactive Learning?

Extended research

- **Question Rephrasing / Enhancements**
  - VQA-Rephrasings (Shah et al., 2019): A new visual question answering dataset and evaluation protocol to measure robustness of VQA models to linguistic variations and a new *cycle-consistency* inspired framework to make VQA models robust to these variations.
  - Question-Driven Image Captions (Özdemir et al., 2024): incorporates image captioning as an intermediary process within the VQA pipeline; explores the efficacy of utilizing image captions instead of images and leveraging large language models (LLMs) to establish a zero-shot setting.



VQA pipeline exploiting general and the proposed question-driven (QD) image captioning as an intermediate step