Introduce a new domain, Visual Query Detection (VQD), which is an expansion of Visual Question Answering and Object Detection, involving image based question answering.

- VQD involves human interaction with deep machine learning models via context-sensitive queries.
- VQD contributes towards Artificial General Intelligence (AGI).

PROBLEM OVERVIEW

- Input: Image of a natural scenario + Question regarding the scene.
- Output: Output with the answer in form of bounding box(es) over the image.
- Snippet of the same image from MS COCO website [1].
- Each person present here is denoted in different colors using a segmentation mask.
- Compared to the picture from our dataset in Fig. 1 where only 2 people are shown, we see 8 different people annotated here.
- Filtering of candidate boxes based on the question is observed.

DATASET CHARACTERISTICS

- As this is a novel task, a new dataset had to be created.
- Images and annotations were extracted from MS-COCO and Visual Genome.
- Questions made by the writing scripts, performing feature extraction over annotations.
- Types of questions:
  - Object Recognition: Show me the person holding a handbag in the image.
  - Color Reasoning: Show me the person wearing orange jacket.
  - Positional Reasoning: Show me the people to the left of the handbag in the image.
  - Absurd: Show me the person wearing a yellow jacket.
- Absurd is a type where we want the model to handle questions which have no correct answer as the output. An AGI targeted feature.

DATASET STATISTICS

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Number of Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object Recognition</td>
<td>145,488</td>
</tr>
<tr>
<td>Color Reasoning</td>
<td>68,094</td>
</tr>
<tr>
<td>Positional Reasoning</td>
<td>14,344</td>
</tr>
<tr>
<td>Absurd</td>
<td>135,620</td>
</tr>
<tr>
<td>Total Number of Images</td>
<td>120,431</td>
</tr>
<tr>
<td>Total Number of Questions</td>
<td>359,536</td>
</tr>
</tbody>
</table>

REFERENCES: