Visual Question Answering

Group member:
VIKASH KUMAR (B200708CS)
FAROOQUI NABEEL AHMED ASHFAK AHMED(B200804CS)
DATTATRAY KHANSOLE (B200706CS)

Guide Name:
CHANDRAMANI CHAUDHARY
Assistant Professor - CSED
National Institute of Technology Calicut

Abstract: Visual Question Answering (VQA) aims to enable machines to comprehend and respond to questions posed about images. The core objective of Visual Question Answering is to develop algorithms that can understand textual questions with respect to visual cues present in the images, and then generate accurate textual answers. VQA is all about making the computer smart enough to understand the picture and give the right answer. This task involves both textual and visual information. Each of these modalities has made tremendous progress in recent couple of years. We will be exploring models with respect to each individual modality and the models which use both visual and textual modalities.

Index

1.Introduction

The main objective of Visual Question answering is Given an image and a natural language question about the image, the task is to provide an accurate natural language answer. Mirroring real-world scenarios, such as helping the visually impaired, both the questions and answers are open-ended. Visual questions selectively target different areas of an image, including background details and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning than a system producing generic image captions. As questions can be multiple choice format or one word answer. Our model will be able to detect the type of question and generate answers accordingly on its own.

2.Literature Survey

In this we have covered major datasets published for validating the Visual Question Answering (VQA) task then we have covered different deep learning models for VQA.

1. **Datasets:**

- **DAQUAR:** A dataset named "Dataset for Question Answering on Real World Images" (DAQUAR) [1] was the first dataset created for Image Question Answering (IQA). It

contains images from the NYU-Depth V2 dataset and includes 12468 question-answer pairs generated from 1449 images.In this questions types cover a range of information about the objects,scenes and action mentioned in the images.

Ex: what is the object on the chair? What is behind the television?

- **VQA Dataset:** The Visual Question Answering (VQA) dataset [2], sourced from the MS-COCO dataset, is one of the largest IQA datasets. It contains over 600,000 questions, both open-ended (multiple different answers may also be correct) and multiple-choice, with at least three questions per image. Answers are categorised into Correct, Plausible, Popular, and Random.

  Ex: Can you describe the person's attire in the photo?

- **Visual Madlibs:** The Visual Madlibs dataset presents a different format for Image Question Answering, including fill-in-the-blank type questions. It focuses on completing sentences related to people, objects, appearances, activities, and interactions.

- **Visual7W:** Based on MS-COCO, the Visual7W dataset consists of over 300,000 question-answer pairs and emphasises seven forms of questions: What, Where, When, Who, Why, How, and Which. It also includes text-based and pointing questions.

- **CLEVR:** CLEVR, a synthetic dataset designed to test VQA systems' visual understanding, features images with objects (cylinders, spheres, and cubes) placed in different sizes, materials, and colours. Questions are synthetically generated based on these objects.This consist of images containing simple 3D objects and corresponding questions about those images.

- **Tally-QA:** Introduced in 2019, Tally-QA is a large dataset for object counting in open-ended tasks. It contains a significant number of questions and images, including complex question types.

- **KVQA:** Developed for common-sense questions, the Knowledge-based VQA dataset requires world knowledge to answer questions. It involves multi-entity, multi-relation, and multi-hop reasoning over large Knowledge Graphs (KG).

2. **Deep Learning-Based VQA Model:**
   - **Vanilla VQA:** Considered a benchmark model, the Vanilla VQA model uses Convolutional Neural Networks (CNN) for image feature extraction and Long Short-Term Memory (LSTM) or Recurrent networks for processing language. It combines these features to classify answers.

   - **Stacked Attention Networks:** This model introduced attention mechanisms using softmax outputs of intermediate question features. Stacked attention helps the model focus on important image regions.

   - **Teney et al. Model:** The Teney et al. Model, winner of the VQA Challenge 2017, incorporates object detection into VQA, employing the R-CNN architecture to improve attention and accuracy.

   - **Neural-Symbolic VQA:** This is particularly for the CLEVR dataset, this model leverages structured

features to filter answers based on question formation and image generation strategies.

- **Focal Visual Text Attention (FVTA):** Designed for VQA in videos, the Focal Visual Text Attention model combines image and question text features using attention mechanisms for improved performance.

- **Pythia v1.0:** An award-winning VQA architecture, Pythia v1.0 reduces computations, uses GloVe word vectors, and combines the predictions of 3D models.

- **Differential Networks:** This model reduces noise and learns interdependencies between features using differences between forward propagation steps. It employs Faster-RCNN for image features and GRU for question feature extraction.

3. MOTIVATION

Here are some works related to visual question answering

- **Stacked Attention Networks for Image Question Answering**
  Working Principle: This work proposes a model that employs stacked attention mechanisms. It uses a bottom-up approach where local image patches and question words are attended to iteratively. The attention weights are used to refine image features and generate the answer through a fully connected layer.

- **Show, Ask, Attend, and Answer: A Strong Baseline for Visual Question Answering**
  Working Principle: This approach combines a CNN for image feature extraction and an LSTM for question processing. The innovation lies in using dynamic parameter prediction to weigh the importance of different modalities (image and question). The model uses both intra-modal and inter-modal attention mechanisms to focus on relevant parts of the image and words for answering questions.

- **Hierarchical Question-Image Co-Attention for Visual Question Answering**
  Working Principle: This work proposes a model that employs stacked attention mechanisms. It uses a bottom-up approach where local image patches and question words are attended to iteratively. The attention weights are used to refine image features and generate the answer through a fully connected layer.

- **Bilinear Attention Networks**
  Working Principle: This work introduces bilinear attention, which captures intricate interactions between different regions in the image and words in the question. The model employs a bilinear attention layer to compute pairwise interactions between image regions and question words, capturing fine-grained relationships. This detailed attention map is then used to predict the answer.

- **Differential Network for Visual Question Answering**
  Working Principle: This approach introduces a Differential Neural Computer (DNC) architecture for

VQA. The model dynamically processes both image and question data through a memory network. It utilizes the memory network's read and write heads to focus on relevant parts of the question and image, allowing the model to reason and generate answers based on the attended information.

- **Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering**
  Working Principle: This model combines bottom-up and top-down attention mechanisms. It first identifies salient image regions using a bottom-up attention mechanism and then utilizes top-down attention to focus on specific regions given the question. By combining these attentions, the model generates fine-grained features, which are then used to answer questions.

4. Methodology

In this section, we outline the methodology of our project,detailing the sequential stages that lead to the implementation of a Visual Question Answering(VQA).

1.Data Collection and Pre-processing:

Image Data: Collect and pre-process the image dataset. This might include resizing, Normalisation, and augmentation.
Text Data: Collect and pre-process the question-answer pairs. Tokenize the questions and answers.

2.Feature Extraction:

Image Features: Extract relevant features from images using techniques like Convolutional Neural Networks (CNNs).
Text Features: Utilise word embeddings like Word2Vec, GloVe

3.Model Architecture:

Merge Modalities: Combine image and text features. Fusion methods like concatenation, element-wise multiplication.
Question Processing: Process the questions using recurrent neural networks (RNNs)
Answer Generation: Design an output module, often a softmax layer, to predict the answer. It could be a classification task or a regression task based on the nature of the answer.

4.Training:

Loss Function: Design an appropriate loss function based on the task (e.g., cross-entropy loss for classification, mean squared error for regression).
Optimization: Use optimization techniques to reduce error.

5.Model Optimization and Fine-tuning:

Hyperparameter Tuning: Tune hyperparameters like learning rate ,batch size and network architecture to optimise the model's performance.
Regularisation: Implement techniques like dropout or L2 regularisation to prevent overfitting.
Fine-tuning: Fine-tune the model based on validation performance to achieve better results.

5. CONCLUSION

In conclusion we studied all relevant datasets which are available for VQA. We also covered available deep learning architecture/models for VQA. Till now we have covered different types of dataset which include given an image and an open-ended natural language question about the image, the task is to provide an accurate natural language answer.

Based on all dataset and deep learning models we have an initial design which include fundamental algorithm CNN for image feature extraction and RNN for processing given questions.

6. References:

1. M. Malinowski and M. Fritz, “A multi-world approach to question answering about real-world scenes based on uncertain input,” Adv. Neural Inf. Process. Syst,vol.2, no January, pp. 1682–1690, 2014.
2. A. Agrawal et al., “VQA : Visual Question Answering,” pp. 1–25
3. Information fusion in visual question answering: A Survey