IJNRD.ORG ISSN : 2456-4184

International Research Journal
IJNRD
Research Through Innovation

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Visual Question Answering using Graph Neural Networks

Piyush Pushkar[1*], Ujjwal Kumar[1*], Aryan[1*], Arjun Kumar Singh[1*], Surabhi Narayan[1*]

[1] Department of CSE, PES University,
100 Feet Ring Road, Banashankari Stage III, Dwarka Nagar, Banashankari, Bengaluru

Abstract : Answering visual questions is a task that has received much attention from two major communities: Computer Vision and Natural Language Processing. Recently, it has been widely accepted as a complete AI function. Wouldn't it be great if machines could understand the content of the pictures and communicate this understanding effectively as human beings? Such technology can be very powerful, either to help a visually impaired person navigate the world of visual acuity, to assist an analyst in extracting relevant information from observation education, to teach a child to play a game on touch screen, to provide viewer information to an art gallery, or to share a robot. This can help elementary school students learn with very little help from their parents or teachers such as recognizing shapes, colors, letters etc. With the advent of computer technology and natural language processing, we are closer to achieving this dream than ever before. Although the main purpose of the project is to model where the computer is required to provide the correct answer to the native language question asked about the input image. The purpose of this program is to assist primary school students with minimal help from others and visually impaired students.

***IndexTerms* - VQA, Scene Graphs, YOLO, Graph Convolutions, GloVe Embeddings**

I. INTRODUCTION

The task of answering questions is one of the tasks that require a deeper understanding of the natural language. Computer science, machine learning, and in-depth learning concepts involving natural language processing (NLP), textual analysis, and information retrieval to develop a model that can provide accurate answer for user queries. Modeling of complex relationships between context and question is necessary when answering a question related to any subject being studied.

Many modern QA systems use in-depth learning structures to better understand the semantics of the text and question, and then use a few methods to capture the relationship between the two. For providing accurate answers model should inculcate external knowledge. Answering visual questions is another area of research that has aroused much interest in recent years (VQA).
VQA can be considered as the development of the concept of machine understanding. It is also a multidisciplinary AI function that combines advanced computer visibility and natural language processing techniques to create a system that can answer the image-based questions. The model tries to find the basic meaning of the image and the semantics and answers questions based on that information. Unlike graphic captions, where unconventional computer-aided visualization and natural language processing are sufficient to develop AI models, functions such as VQA require a complete understanding of advanced techniques in both domains.

The previous AI systems do not have the ability to give answers which can help improve user experience. However, thanks to ongoing research on this topic, it is now possible to try to create a program that is intended to be successful in activities such as VQA. The most complete knowledge of visual and comprehensive thinking skills is often required in such a program. The most advanced version of this system can also have real-time information, which allows it to understand and answer questions with answers that are not explicitly depicted in the image. VQA is a novel problem statement with an interesting concept. To achieve the goal, most of the research in this article uses a variety of in-depth learning projects and learning algorithms in the fields of computer vision, object recognition, and natural language processing.

Most VQA data sets and models available focus on queries that can be answered directly by analyzing the embedded image. When we look at the whole problem, we see that functions like VQA present a wide range of problems which need to be overcome involving all vision, language, and semantic knowledge. It is possible to build different components of the VQA system and integrate them to suit the work being done using the research available in each of these domains. The goal of this project is to create such a VQA system.

We formulate the challenge in this study, followed by a discussion of existing models, associated research, and open datasets for VQA. We then give an overview of our suggested systems, as well as the tests we conducted and the results we acquired.

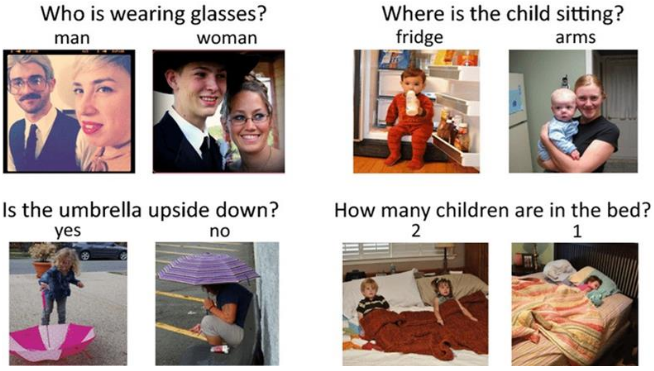

Who is wearing glasses?
man
woman
Where is the child sitting?
fridge
arms
Is the umbrella upside down?
yes
no
How many children are in the bed?
2
1

Fig.1.1 Examples of VQA

II. LITERATURE REVIEW

Visual Question Answering over Scene Graph [1]

This paper proposes to encode a scene graph and a question using Graph Network. Scene graph is one of the ways to extract information from images. Using scene graphs we can represent the objects in the images as nodes and the relation between them by the edges. This is one of the most efficient ways to extract information from the images, this graph can be further used for reasoning, image retrieval, visual question answering and many other tasks. After which the question was encoded to the Memory, Attention, and Composition (MAC) model to generate answers. The paper also advises that encoding of the scene graph based on context is highly important for our problem statement.

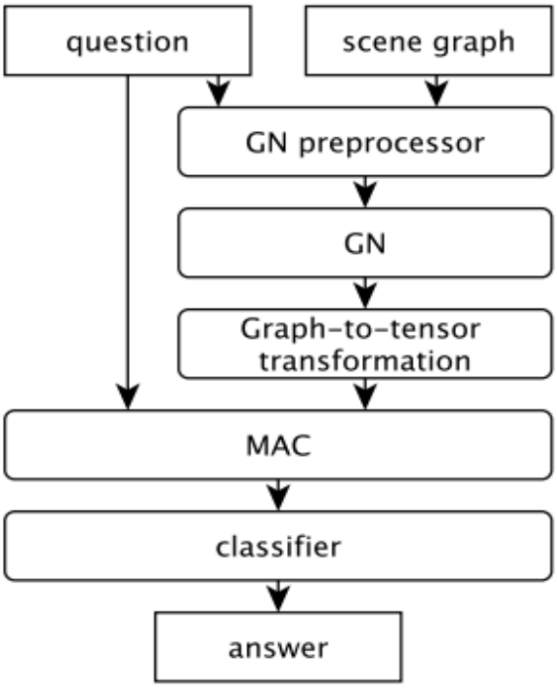


Fig -2.1 High Level Architecture of the Proposed Method

The use of scene graphs had a huge impact on the accuracy of the model. Further we can use attention mechanism for the encoding of questions which will further help us to emphasize on the important words on the question and help in understanding it more clearly. Towards Knowledge-Augmented Visual Question Answering [2]

The proposed model implements a graph-based approach in which scene graphs and concept graphs are used, and uses Graph Attention Networks to learn question-adaptive graph which gives more importance to key knowledge instances.

Image Representation - Faster R-CNN for object detection. Question Representation - To generate question embeddings, bidirectional RNN (GRU) was used to perform self-attention.

Knowledge Retrieval – This step captures the relation between objects in the image, by using both the images and questions. And to grasp the relationship between objects, a scene graph is constructed with the help of objects detected in the Image Representation step. Also, a concept graph is generated to represent the explicit relations present. The different objects which will have been detected from the image will be to use to construct a scene graph which will represent the semantic and the spatial relationship between objects. The different objects detected will be considered as the nodes of the graph and the relationship between them will be represented as the edges between the nodes.

Concept graph is constructed by integrating the explicit relations between knowledge entities. Therefore, in designing the proposed graph, use a question-conditioned attention mechanism and dynamically assign higher weights to those knowledge instances that are mostly relevant to each question, instead of treating all the entities equally.

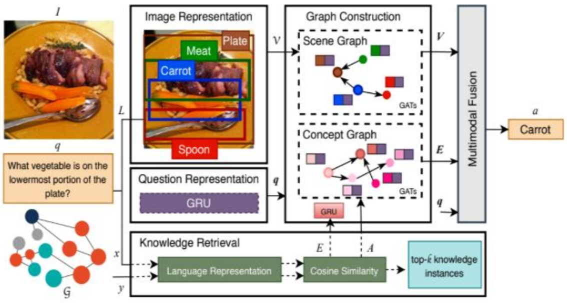


Fig -2.2 High Level Architecture of the Proposed Method : Here we see the fusion of the scene graph constructed with the objects in the image and the Concept graph with the question representation of the image.

Relation-Aware Graph Attention Network for Visual Question Answering [3]

This paper provides an approach using a Relation-aware Graph Attention Network, which converts the image into a graphical representation known as scene graph and used graph attention mechanism to model different type of object relations to frame a question adaptive relation representation.

There are two types of object relations which are considered:
(i) Explicit Relations – Relations which signify the geometric positions and semantic interactions among different objects
(ii) Implicit Relations – Relations which signify the hidden dynamics within image regions.

The proposed approach suggests that using the context from a question, the graph attention is learned through which semantic information from question is injected during the relation encoding step. The learning through the previous step captures both the inter-object relations as well as the semantic relations, to concentrate on particular types of relations for each question.

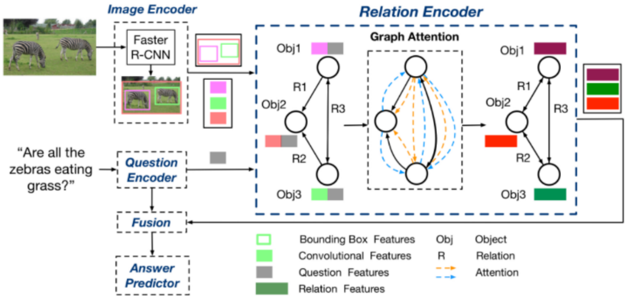
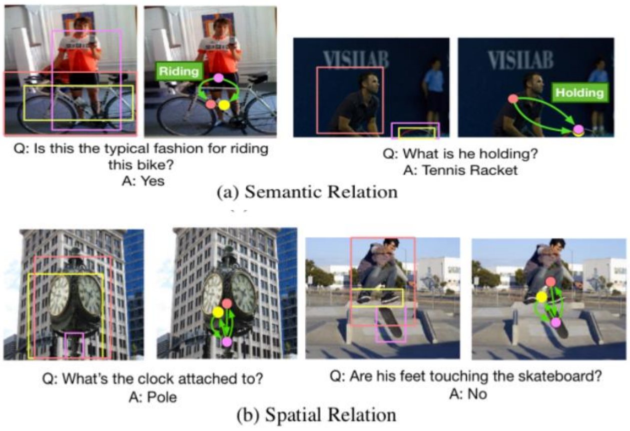


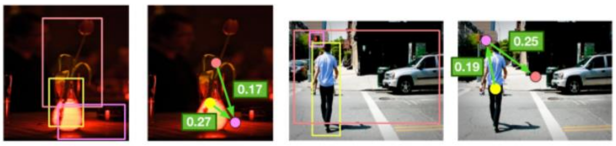
Fig 2.3 Represents the Model architecture for the above approach in which objects are detected using Faster R-CNN and the explicit and implicit relations between the objects are taken along with the question encoder to build up a Relation Encoder.

The ReGAT model considers both the explicit relations as well as the implicit relations and the relation encoder uses graph attention to capture the question – adaptive object interaction. The Relation Encoder is further divided into three parts Semantic, Spatial and Implicit Relation Encoder.


Riding
VISILAB
VISILAB
Holding
Q: Is this the typical fashion for riding this bike?
A: Yes
Q: What is he holding?
A: Tennis Racket
Q: What's the clock attached to?
A: Pole
Q: Are his feet touching the skateboard?
A: No

(a) Semantic Relation

(b) Spatial Relation



Q: Where is the vase?
A: On the table

Q: Should the people be walking according to the light?
A:No

(c) Implicit Relation

Fig 2.4 Examples of Semantic, Spatial and Implicit Relation between the objects pairs in an image.

Zero-shot Visual Question Answering using Knowledge Graphs [4]

Two inadequacies inspire the primary strategy of this paper: Building intermediate models and involving KG queries in a pipeline manner is prevalent in current methodologies, which contributes to error propagation. Second, the majority of them characterize VQA as a classification issue that does not take use of sufficient information of the responses and fails to anticipate unknown answers.

Knowledge Graphs: A collection of three Knowledge Graphs (DBpedia, Concept Net, and WebChild) was employed. It's used to create a previous knowledge connection, which consists of a collection of answer vertices and idea (tool) nodes that help to enhance the links between responses.

Establishment of Multiple Feature Spaces: Find the relation between an answer and the picture question pair it corresponds to by projecting them into a similar feature area and becoming near to each other.

1) The semantic space focuses on the linguistic information inside the picture query, which serves as a guide for projecting triplet relations r in the Knowledge Graph.

2) Object space is more likely a feature space concerning support entity classifier, which concurrently analyses pictures and texts for prominent features, as opposed to standard image classification, which determines the proper category of a given image.

Answer Mask via knowledge: Masking is extensively employed in pre-training language models to improve the machine's interpretation of the text. We obtain discontinuous fusion embedding in two distinct feature spaces using the learnt features, which are used as the foundation for future object and relation analysis.

Fusion Model: To parameterize the fusion function F, different models are used. As a depiction of grid-based visual fusion, we use the Multi-layer Perceptron (MLP) and Stacked Attention Network (SAN). Because of its superior performance, SAN was selected as their framework's fundamental feature extractor. They extract visual characteristics from the layer 4 output of ResNet-152 (14×14×2048 tensor) pre- trained on ImageNet to get access. Meanwhile, to get bottom-up picture area features, use a ResNet-101-based Faster R-CNN that has been pre-trained on the COCO dataset. Every word in the question and response is represented as a GloVe vector of 300 dimensions. For every time step, the series of embedded words in query (average length: 9.5) is fed into Bi-GRU.

What we can apply: We can try to implement the knowledge graph as it helps to predict output of instances which are not seen in the training dataset.

Learning Conditioned Graph Structures for Interpretable Visual Question Answering [5]

This study proposes that the Visual Question Answering issue be modelled as a classification problem, with each training set response representing a class. To answer a query about a picture, the study advises creating a deep neural network that incorporates spatial, image, and linguistic data in a unique way. The CNN detects object characteristics in pictures in the form of a bounding box, and the graph module then creates the relationship between objects to create an adjacency matrix that is dependent on a specific query.

An undirected graph is created in which the objects/entities are considered as nodes or vertices and the relations between the entities as the edges of the graph.

Spatial graph convolutions: They employ Graph Convolution Neural Networks to perform spatial graph convolutions, which operate directly in the graph domain and significantly depend on spatial interactions. They also capture spatial information by using a polar reference frame on the bounding boxes that the Convolution Neural Network has discovered, as well as a patch function that describes the effect of each surrounding node and is resilient to irregular neighborhood structures.

Prediction Layers: Through spatial graph convolution layers, a convolution graph representation H is produced, followed by a global vector representation of the graph hmax through a max-pooling layer across the nodal dimension. A 2-layer MLP with ReLU activations is used to calculate the classification Logits.

Loss Function: The method employs a sigmoid activation function with soft target scores, which has been demonstrated to provide superior predictions.

Conclusion: The performance of this model is highly dependent on the performance of object detectors which can miss or misinterpret objects. Further, the bounding boxes that they define may overlap each other resulting in propagation of erroneous information in the graph. Modeling the VQA problem as a multi-class classification problem strongly limits performance.

YOLO-ing the Visual Question Answering Baseline [6]

This paper proposes a simple model for Visual Question Answering using YOLO object detection. In the proposed model the image features are extracted with the help of Inception V3 which are weighted according to the attention generated from the question and image. Next, they have used YOLO pretrained on MS COCO dataset for 80 different object classes to detect the objects present in the image. The detected objects are then stored in vector form and then passed on to the next step. Each question in the VQA dataset is transformed using one-hot encoding. The words which appear less than six times in the dataset are removed and the same thing is also done for the answers. Finally for appending the question and image features a hyperbolic tangent activation function is used. After passing the output of the gated tanh via a softmax layer, it is reduced to a K dimensional vector.
This vector is known as the attention vector, and it contains the importance of the k-th picture position with the k-th element. The focus vector is then used to calculate the image characteristics weighted sum.
Conclusion: One of the shortcomings is that the object detector used counts the number of objects and stores them as vector form but they do not take into account the position of the objects which is an important aspect to keep the spatial relationship between the objects in mind.

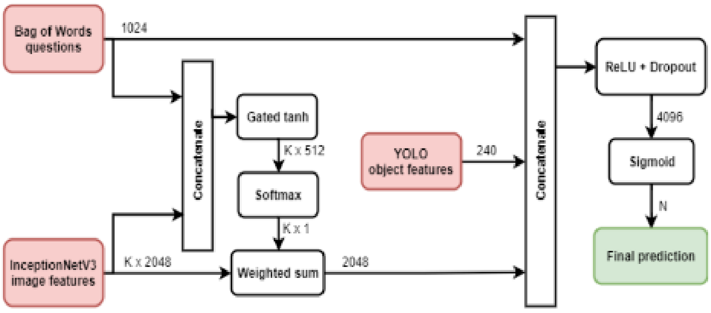


Fig 2.5 Depicts the process of generation of Scene graphs.

Scene Graph Generation via Conditional Random Fields [7]

A novel scene graph generation model for predicting object instances and its corresponding relationships in an image. SG-CRF learns the sequential order of subject and object in a relationship triplet and the semantic compatibility of object instance nodes and relationship nodes in a scene graph efficiently.

Graph R-CNN for Scene Graph Generation [8]

Proposes a novel scene graph generation model called Graph R-CNN, that is both effective and efficient at detecting objects and their relations in images. The model contains a Relation Proposal Network (RePN) that efficiently deals with the quadratic number of potential relations between objects in an image. Also propose an attentional Graph Convolutional Network (aGCN) that effectively captures contextual information between objects and relations.

Exploring Context and Visual Pattern of Relationship for Scene Graph Generation [9]

In order to discover effective pattern for relationship, traditional relationship feature extraction methods such as using union region or combination of subject-object feature pairs are replaced with our proposed intersection region which focuses on more essential parts. Therefore, we present our so-called Relationship Context - Intersection Region (CISC) method.

Unbiased Scene Graph Generation from Biased Training [10]

Today's scene graph generation (SGG) task is still far from practical, mainly due to the severe training bias. This presents a novel SGG framework based on causal inference but not the conventional likelihood. First building a causal graph for SGG, and perform traditional biased training with the graph. Then, propose to draw the counterfactual causality from the trained graph to infer the effect from the bad bias, which should be removed.

III. DATASET

We have used Visual Question Answering (VQA 2.0) for training and validation purposes. Our training dataset consists of 82,783 images and validation dataset consists of 40,504 images.
Each image consists of up to 5 questions and each question has up to 10 answers.

IV. METHODOLOGY

In our Proposed Approach the way in which we have decided to approach the problem of Visual Question Answering is through the use of Graphs. At first, we have detected objects from the given input scene. We went through different methods of object detection and we have decided to go through with using YOLO v5 pre-trained model for detecting objects. But it contains only 80 object classes, so we plan to extend it by training it over the Google Open Images Dataset to extend it to 200 classes, as to detect almost all objects in the given scene.

After Detecting the objects, we plan to generate Scene Graphs using them, to be used further. The objects detected will be used as nodes and the edges between them will represent the relationship between the objects. In this way we will be able to incorporate the spatial relationship between the objects in our model.

For the question-answer pair, we want to employ glove embeddings (GloVe, which encodes the co-occurrence ratio between two words) to get vector representations of the phrases. We combine the output from scene graphs and the question embeddings to produce joint language vision knowledge representations. To deliver a meaningful response, the fusion technique must identify the interaction between the various features. Multi-layer perceptron should be used on the question embeddings and scene graph visual characteristics to build a vector that reflects these two aspects simultaneously. This fusion model's output can be sent to a classifier for predicting responses.

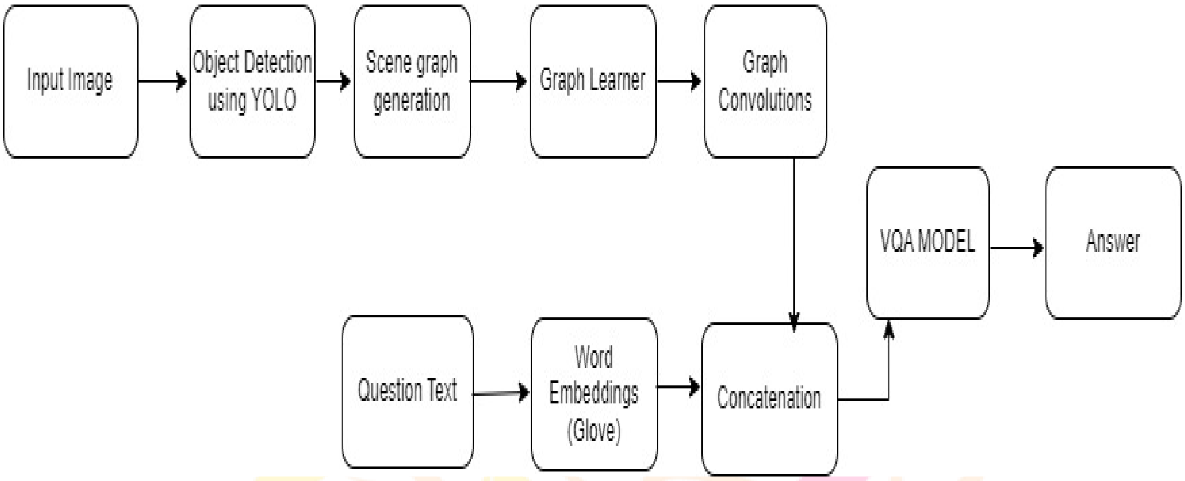


Fig 4.1 Represents the High-Level Design of Proposed Solution

The architecture contains two main parts:
The generation of scene graphs from the input image and Encoding of the Question-answer pairs.

In the first part after some initial pre-processing, like converting all the input images to the same dimension, we pass the input image to our object detection algorithm i.e. the YOLO algorithm. Here our object detector is able to detect some 200 common classes of objects that are present in our dataset and then it is passed to the next step for the generation of scene graphs. The graph generated is then converted to vector form and then passed to the next step.

For the second part of the question-and-answer pair we plan to use GLOVE word embeddings for the encoding of the sentences and extracting the information and finally converting it to vector form.

Finally, we concatenate both of these parts in such a way so that none of the features detected in the previous steps are lost to make our model.

4.1 Pre-processing text

Count the 3000 most frequent words in question and answers on the basis of their occurrences. Tokenize the questions to remove all punctuation marks and separate the words. Combine the question answers on the basis of question, question_id and image_id and create a json file. We have limited the number of words in the question to 14.

4.2 Object Detection

We have used YOLOv5 model for detecting objects. The original YOLO model contained only 80 object classes but we have trained it for further more 200 classes. Our final model detects some 280 object classes.

YOLO works in 2 stages. In the first stage it predicts the possible regions where objects could be present and in the second stage it classifies them into different classes. YOLO works by dividing the image into grids and then does the process of detecting the possible region by calculating intersection over union of different grids which provides the probability of how close the bounding box is to the original object.

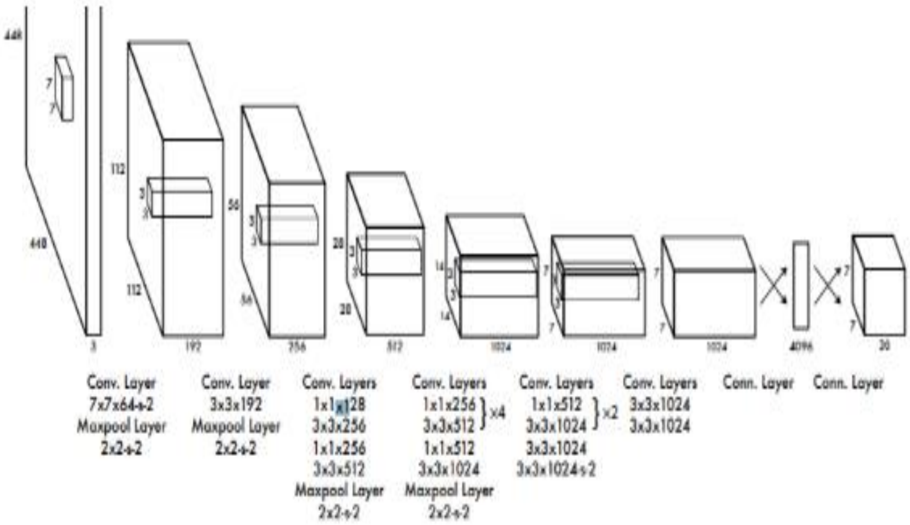


Fig 4.2.1 Architecture of YOLO object Detection Model

The above architecture has 24 convolutional layers followed by 2 fully connected layers. Alternating 1x1 convolutional layers reduce the feature space from preceding layers.

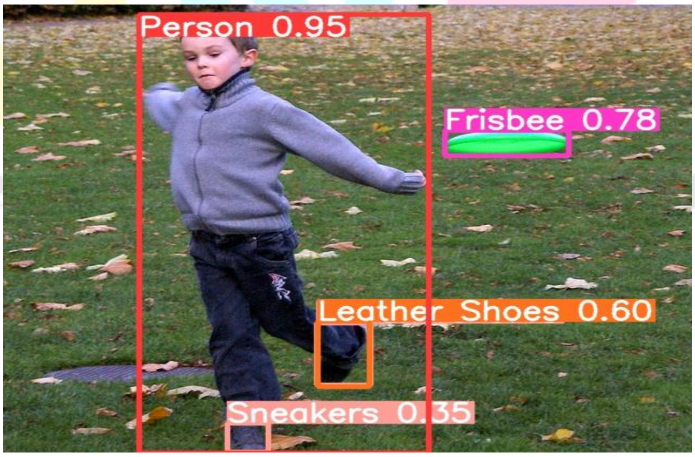


Fig 4.2.2

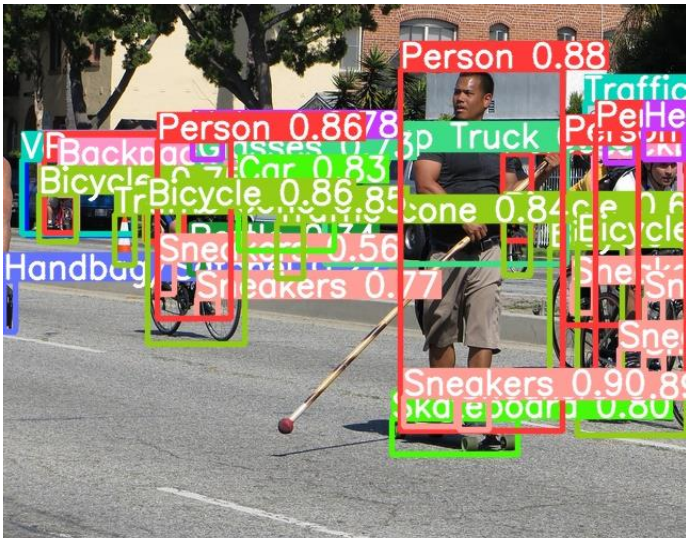


Fig 4.2.3

Fig 4.2.2 and Fig 4.2.3 depicts the objects detected using our YOLO model. The text on top of each of the bounding boxes tells the object and confidence interval of the detected object.

4.3 Scene Graphs

Scene graph is a structured representation of a scene that can explain the objects, attributes, and relationships among the objects present in the input image. Scene graph is used to represent the relationship between different entities that are present in our input image using bounding boxes, centres and the confidence intervals. Using the coordinates of the centre of the bounding box of objects detected we create an adjacency matrix.

We are constructing an undirected graph G = {V, E, A}, where E is the set of graph edges, V corresponds to the detected objects in the image and A is the corresponding adjacency matrix.

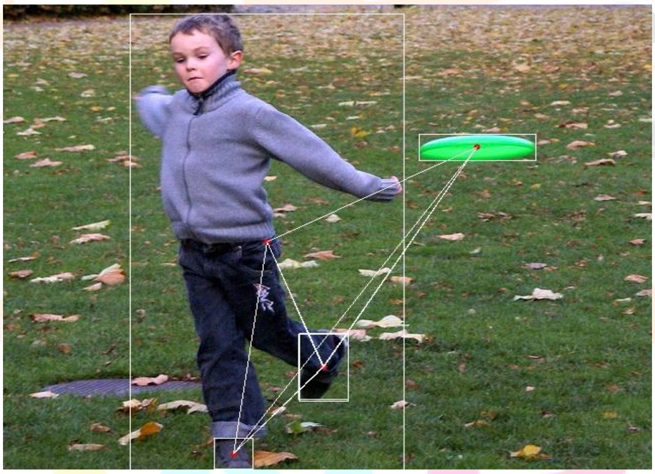


Fig 4.3.1

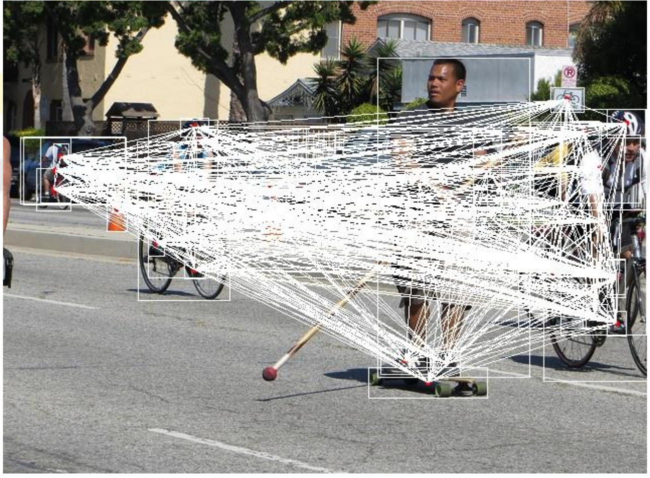


Fig 4.3.2

Fig 4.3.1 and Fig 4.3.2 depicts the scene graph generated using the objects detected in the Fig 4.2.2 and 4.2.3 respectively.

4.4 Glove and Word embeddings

As we know that machines are not able to process textual data so we need to convert it into numerical form so that they can correctly interpret it. Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation. The word embedding is represented as a co-occurrence matrix which is computed using the following formula:

$$P(j|i) = \frac{X_{ij}}{\sum X_{ij}} = \frac{X_{ij}}{X_i}$$

where, X_{ij} encodes global information about the co-occurrence between i and j(global: because it is computed from the entire corpus) and P(j|i) gives the probability about their co-occurrence.
To start, we need to get the word embedding and add them to a dictionary. For each of the tokens generated using the question in the training set, an embedding matrix will be created. Then a word embedding matrix for each word in the word index we previously acquired. A zero matrix will be displayed for words that do not have an embedding in GloVe.

4.5 Graph Learner

The adjacency matrix takes into account the similarities between feature vectors and their relevance with the question. First concatenate each of the N visual features v_n with the question embedding q, which we represent as $[v_n][q]$.
We then compute a joint embedding as:

$$e_n = F([v_n][q]), n = 1,2, \ldots, N$$

where,
q is the question embedding (in tensor format), v_n are the visual features which are the detected objects(the bounding boxes, confidence intervals and their centers)

By concatenating the joint embeddings e_n together into a matrix E, then an adjacency matrix is defined for an undirected graph as A = EET so that $A_{i,j} = e_iTe_j$.

Our model can detect 280 different object classes in an image. For an image with a large number of visual features the detected objects count would be more and it would become computationally difficult to take into account each and every detected object for our adjacency matrix calculation. Hence sparsity is introduced to extract top K objects where k = 36. These top K neighbors are selected based upon the confidence interval of the detected objects in the input image. The confidence interval depict the probability of the object present in the image.

4.6 Graph Convolution

Graph Convolution Network is a type of CNN that works directly with graphs and utilize their structural and spatial information. Out of the different predictions on graphs, we are using the graph-level prediction to predict the answers by aggregating all the graph node features.

$$H^{(l+1)} = \sigma(\tilde{D}^{\frac{-1}{2}}\tilde{A}\tilde{D}^{\frac{-1}{2}}H^{(l)}W^{(l)})$$

Based on the above formula, A is the adjacency matrix A~, D is the degree matrix, W is the weight matrix and H is the hidden layer used in the process of graph convolution.

The GCN process captures spatial information through the use of a pairwise pseudo-coordinate function u(i, j) which defines, two of the detected objects centered at i and j which represents the relative spatial position of the centers which are basically the center of the bounding boxes of the detected objects. This function returns the polar co-ordinate vector (ρ, θ).
The computed weights of the neighborhood features are multiplied with the pseudo co-ordinates that give the gaussian kernel. This kernel and the adjacency matrix are used to perform the graph convolution to learn new object representations.

$$f_k(i) = \sum_{j \in N(i)} w_k(u(i,j))v_j\alpha_{ij}$$

4.7 Output Classifier

We compute classification logits through a 2-layer MLP(Multi-Layer Perceptron) with ReLU activations.

The Visual Question Answering problem can be viewed as a multi-class classification problem, where each class corresponds to one of the most commonly occurring answers present in the training set. For each of the classes, we compute a target score which is then used by the sigmoid activation function to predict the most probable answer(logits).

4.8 Loss Function

We then compute the multi-label soft loss which is simply the sum of the binary cross entropy losses for each of the classes.

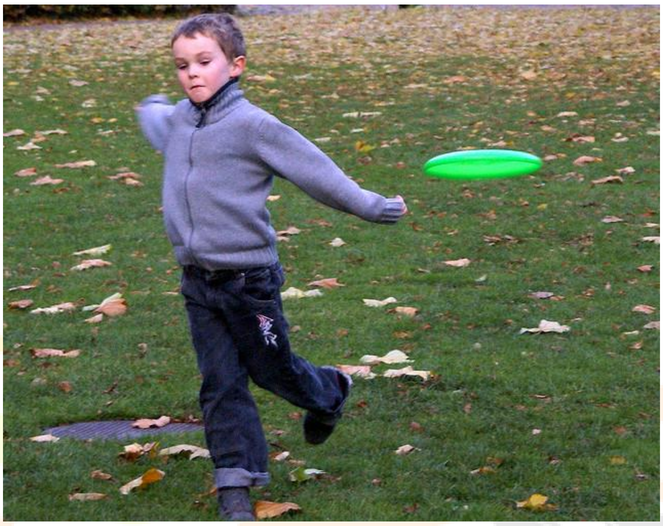
Binary cross entropy compares each of the predicted probabilities to actual class output which can be either 0 or 1. The score is then determined, penalizing the probabilities according to how far they are from the predicted value. This refers to how near or far the value is to the true value.

Binary Cross Entropy is the negative average of the log of corrected predicted probabilities.

$$L(t,y) = \sum_i t_i \log({}^{1}/_{(1+\exp(-y_i))}) + (1-t_i)\log({}^{\exp(-y_i)}/_{(1+\exp(-y_i))})$$

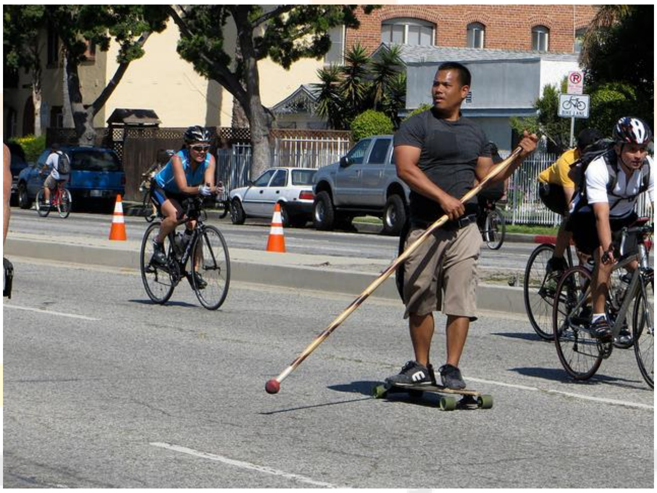
where y is the logit vector (i.e., the output of our model).

V. RESULTS & OUTPUT



Question: "What is the boy holding in his hands
Answer": "frisbee"

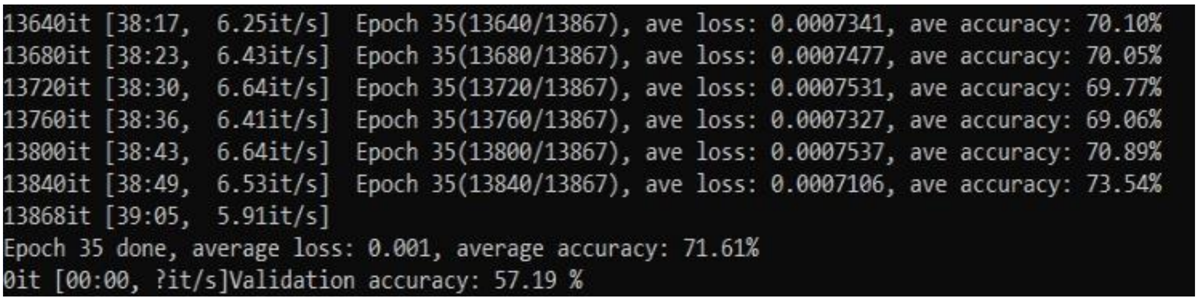
Fig 5.1 Shows that our model was correctly able to interpret that the boy is holding a frisbee in his hands with the help of the objects detected and the scene graph built upon that.



Question: "What is the man doing?"
Answer: "skateboarding"

Fig 5.2 Depicts that on being asked the question what is the man doing our model took in consideration the person with the highest confidence interval to answer that it was skateboarding.

Our Model achieved an **accuracy of 71.61% on training set** and an **accuracy of 57.19% on validation set**.



```
13640it [38:17,  6.25it/s]  Epoch 35(13640/13867), ave loss: 0.0007341, ave accuracy: 70.10%
13680it [38:23,  6.43it/s]  Epoch 35(13680/13867), ave loss: 0.0007477, ave accuracy: 70.05%
13720it [38:30,  6.64it/s]  Epoch 35(13720/13867), ave loss: 0.0007531, ave accuracy: 69.77%
13760it [38:36,  6.41it/s]  Epoch 35(13760/13867), ave loss: 0.0007327, ave accuracy: 69.06%
13800it [38:43,  6.64it/s]  Epoch 35(13800/13867), ave loss: 0.0007537, ave accuracy: 70.89%
13840it [38:49,  6.53it/s]  Epoch 35(13840/13867), ave loss: 0.0007106, ave accuracy: 73.54%
13868it [39:05,  5.91it/s]
Epoch 35 done, average loss: 0.001, average accuracy: 71.61%
0it [00:00, ?it/s]Validation accuracy: 57.19 %
```

Fig 5.3 Shows the Accuracy after training for 35 Epochs

We were able to develop a model which can using novel object detection and using Graph Convolution Neural Network which gives one-word answers for the question related to the image using the visual features of the image.

Per Answer type accuracy of our Model :

Other	56.40
Yes/no	88.89
Numerical value	48.34

Table 5.1 Shows the Per Answer Type accuracy of our Model.

VI. CONCLUSION

We were able to generate one-word answers for the given questions. The use of YOLO algorithm helped us to make our image encoder more accurate and we were to detect even small objects present in the image. The use of Graph Convolution Network helped us to establish the relationship between each pair of objects present in the image. Our model was able to generate better answers compared to CNN based models as GCN works well on irregular structures where the number of nodes may vary and are unordered. CNN on the other hand establishes the relationship between each of the pixels and works well on classification problems. GCN helped us to incorporate various features like spatial and semantic operation which was difficult to learn using CNN. One of the most common problems that we were able to discover was that questions often require answers that are not present in the training dataset and thus give inappropriate answers for those questions.

Although we were able to generate answers for the given question with best possible dataset available but still there were inconsistencies present in the dataset. Some questions like Would you like to fly in this plane? What time is it in the clock? What are the types of fruits that there in the basket whose answers depend on the persons discretion were also present in the dataset.

The dataset that we have used is more generalized and includes images of all types. If there is dataset available related to a specific field or a specific use case then we could have built up a more accurate model.

We were able to detect many objects in the image and this could make our adjacency matrix quite dense so due to computational limitations we were not able to consider all the features of an image. Considering these additional features could further increase the accuracy of our model.

VII. FUTURE WORK

We were able to identify semantic and spatial relationship but we could further work on including Implicit relationship as well.

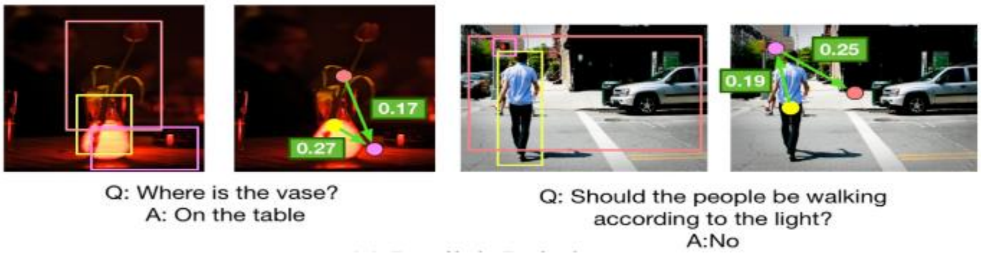


Fig 7.1 Example of Implicit Relations

We could work on extracting the foreground and background features and incorporate it in our model. This could further help us in understanding the relationship between the objects more accurately and also answer the questions based on the implicit relations. Further, we can work on generating more than one-word answers. We can work on counting the number of objects present that are present in the image. For example, how many people are there in the image or the number of plates on the table.

REFERENCES

[1] S. Lee, J. Kim, Y. Oh and J. H. Jeon, "Visual Question Answering over Scene Graph," 2019 First International Conference on Graph Computing (GC), 2019, pp. 45-50, doi: 10.1109/GC46384.2019.00015.

[2] Maryam Ziaeefard and Freddy Lecue. 2020. Towards Knowledge-Augmented Visual Question Answering. In Proceedings of the 28th International Conference on Computational Linguistics, pages 1863–1873, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[3] Linjie Li, Zhe Gan, Yu Cheng, Jingjing Liu; Relation-Aware Graph Attention Network for Visual Question Answering. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10313-10322

[4] Chen, Z., Chen, J., Geng, Y., Pan, J.Z., Yuan, Z. and Chen, H., 2021, October. Zero-shot visual question answering using knowledge graph. In International Semantic Web Conference (pp. 146-162). Springer, Cham.

[5] Norcliffe-Brown, W., Vafeias, S. and Parisot, S., 2018. Learning conditioned graph structures for interpretable visual question answering. Advances in neural information processing systems, 31.

[6] Lanzendörfer, Luca, Sandro Marcon, Lea Auf der Maur, and Team Pendulum. "YOLO-ing the visual question answering baseline." Austin: The University of Texas at Austin (2018).

[7] Cong, Weilin & Wang, William & Lee, Wang-Chien. (2018). Scene Graph Generation via Conditional Random Fields.

[8] Yang, J., Lu, J., Lee, S., Batra, D. and Parikh, D., 2018. Graph r-cnn for scene graph generation. In Proceedings of the European conference on computer vision (ECCV) (pp. 670-685).

[9] W. Wang, R. Wang, S. Shan and X. Chen, "Exploring Context and Visual Pattern of Relationship for Scene Graph Generation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8180-8189, doi: 10.1109/CVPR.2019.00838.

[10] Tang, K., Niu, Y., Huang, J., Shi, J. and Zhang, H., 2020. Unbiased scene graph generation from biased training. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3716-372