Open Images 2019 Visual Relationship

Anjali Kumari, Arijeet Sinha, Melvin Sebastian, Shivi Verma, Gaurav Singal

***Abstract*—Computer vision has come a long way from where we were a few years ago. However, there is a huge gap between the capabilities of a human and that of a machine. Even the enormous datasets currently available the number of categories of objects a computer can identify still cannot be compared to that of humans can. Moreover, computer vision maybe able to identify the objects but it has a long way to go in recognizing the various relationships involved between the various objects. The dataset we are working with to establish visual relationships between the objects is The Open Images dataset which is a collaborative release of around 9 million images annotated with image level labels, bounding boxes, and their visual relationships. This dataset has around 600 classes of objects .We combine a model for object detection and localization along with a model for visual relationship to complete our project.**

***Index words* - Bounding boxes, computer vision, Object detection, Open Images, visual relationship, computer vision.**

I. INTRODUCTION

Visual Relationship has attracted a lot of attention to researchers of late[1]. Visual relationship tells about the relationship between objects in images and investigating on such problems can help us in understanding the images in deep. It greatly helps in the field of image captioning and is an extremely crucial area of research for many tech giants such as Facebook and Google. It also helpful in other areas like object detection and image retrieval. Each images are represented into three labels- <subject, predicate, object>, where two objects appearing in an image are linked through a relation (e.g. man-riding-elephant, man-wearing-hat).

Extracting triples i.e subject, predicate, object from the images is a pretty difficult task. Our goal is to find what

are the objects present in each image and to extract the series of relationships from each image and visualize the different relationships from each image. Our approach is to take images as input and get three labels as output. But the phrase which has the same predicate but different agents is considered as the same type of relationship. For example, the ”clock-on-wall” and “dog-on-sofa”) belong to the same predicate type ”on”, but they describe different semantic scenes

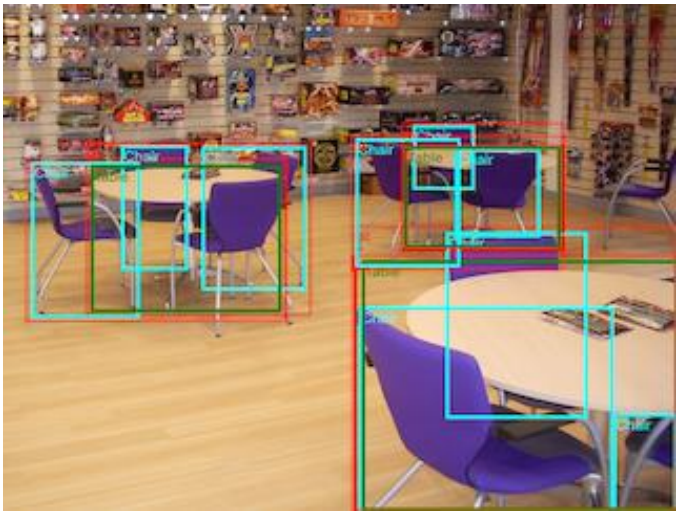

Fig.1 : Chair at table.

On the other hand, the type of relationship between two objects is not only determined by their relative spatial information but also their categories. One major reason why the data scientists struggle to find the right visual relationships is due to the arbitrariness of relationships. There are possibly tens of objects and each object is possibly related to one another in some way spatially or otherwise, the one with meaning is rather arbitrary.

II. RELATED WORK

Visual relationship is not a new concept and a lot of related work has been done. It is an intermediate task connecting vision and vision language tasks, a lot of attempts have been made to understand the use of visual relationship for facilitating specific high-level tasks such as image caption[4,6],scene graph generation [11], image retrieval [10], visual question and answering

(VQA) [2, 1,13], etc. In comparison, our work is dedicated to detect the object and relationship that connects them. [8, 7] attempted to learn four spatial relationships: ”above”, ”below”, ”inside”, and ”around”. [12] detected the physical support relations between adjacent objects. It supports from ”behind” and ”below”. However, we also define some other relationships like “in”, ”at”, ”where”, ”under”, ”holds”, ”plays”, ”interact with”, ”inside of”, ”wears” and “hit”. In [11, 5], each possible combination of visual relationship is treated as a distinct visual phrase class, and the visual relationships detection is transformed to a classification task.The problem with these methods is that they deal with the long trail problem and can detect only a few visual relationships.

Recently, the use cases of deep learning has expanded.It initially consisted of detecting objects, however now it can classify objects into categories. From just being able to detect categories of objects in still images it has expanded to being able to detect objects in live videos[9].Most recently deep learning is being used to get visual phrases.Appearance models for visual phrases improves individual object detection. For example, “a person riding a horse” improves the detection and localization of “person” and “horse”[3].Unlike our work, all previous work attempted to find the handful relationship between objects. Such methods suffers a lot of problems later because they just use a very few number of features.

III.Methodology

A) Object Detection and Visual Relationship Model

The proposed approach consists of two parts. One is for object detection and other is for visual relationship. These two models are combined to give our desired output. Our output must be in words. The goal of our model is to predict the relationship between objects which occur in the image. Object detection model firstly detects what are the objects present in the image and after that it localize each object within the image. The objects in the images are labelled and the different objects present in the image are listed. For ex. Chair, table, cat etc. It gives unique ID to each object and the bounding boxes of each object in the image are found. The second part of the project involves finding the relationships between two objects or the attribute(for example: ‘The chair is wooden.’ Here wooden is the attribute) of a single object.

B) Problem Statement

This problem has two challenges which are as follows:

- Object Detection: predicting a tight bounding box around all object instances of 500 classes.
- Visual Relationship Detection: detecting pairs of objects in particular relations.

C) Approach

The problem of visual relationship track was at first divided into two, object detection and visual relationship. The first problem of object detection was done using a pre-trained model from TensorFlow Hub. The model uses Faster RCNN with InceptionResNetV2. The model found has been pre-trained with the Open Images Dataset v4 making it suitable for our model as we are working with the Open Images Dataset v5 which has the same classes. The model uses MobileNetv2 means that the computing time will be faster, however there is a slight reduction in accuracy. The model is slightly modified so as to produce the results in the required form. A subset of the dataset was used to test the working of the model and the results were found to be suitable and the correct bounding boxes were formed.

Now that the labels and bounding boxes are found. The next step is to find the visual relationships. To accomplish this task the proposed method is to map the provided relation data with the provided bounding boxes and to use that as the dataset for our model to find the visual relationships.

1) Object detection:

The object detection and localization was initially done using MobileNetV2 which is a CNN which can be used for smaller devices. The accuracy is lower than other pre-trained neural network however the speed is much faster.

By making use of SSD or single shot detector, we only need to give a single shot to locate multiple objects in a single image. However the model that has been finally implemented in our model is FasterRCNN+ InceptionResNetV2.The advantage of this model is the much higher accuracy we can obtain.

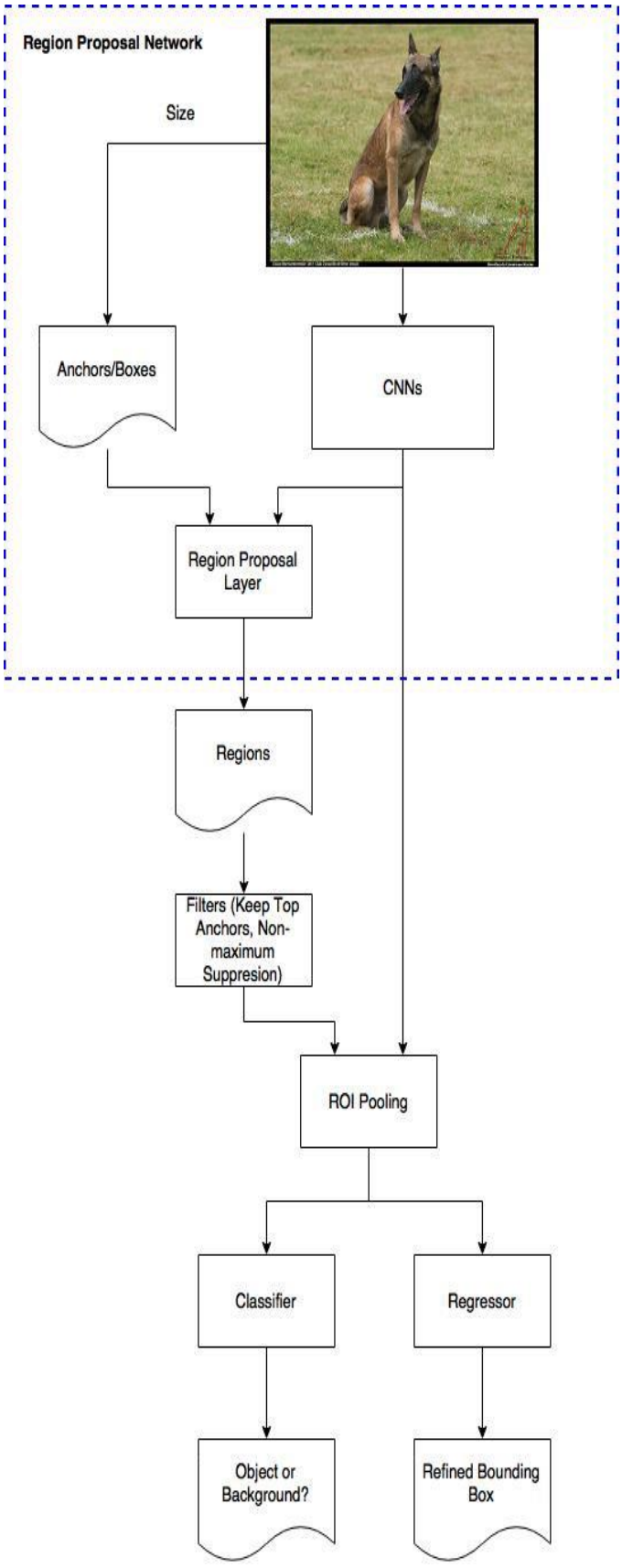

Fig. 2 : Faster RCNN architecture

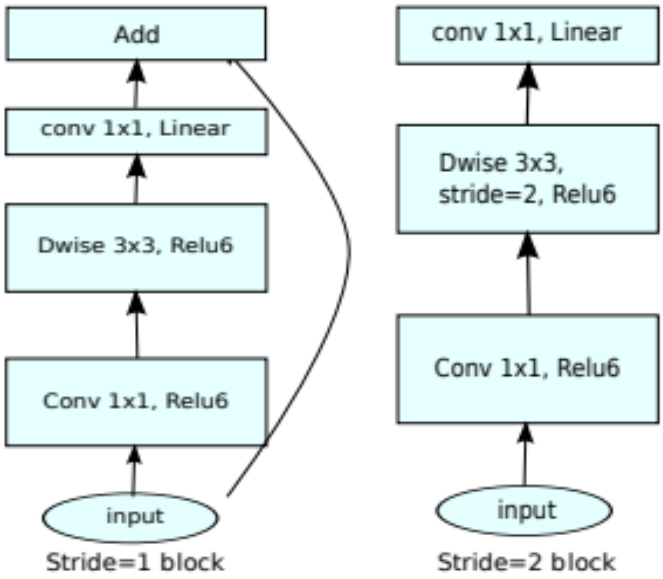

Fig. 3 : MobileNetV2 architecture

2) Visual Relationships:

The main challenge in this project is finding the visual relationships between the objects found detected and making a coherent sentence/ caption for the image. This is done by first, mapping all the objects along with their bounding boxes to all the relationships provided. This is used as the base dataset from which features are defined and a classification neural network is used. There are, however, two parts to this problem as a caption could be a description of a single object or the interaction between two objects. Hence two classification models are formed. The first being to caption the description of a single object (for eg. the chair is wooden). The various qualities or attributes of the objects are limited to 5 classes i.e. wooden, transparent, plastic, textile and leather. Once all the objects are matched with these classes, or a none type. Then a simple multi-image label classifier with six output neurons is used to classify the materials. Therefore all the objects are classified into different categories. The only predicate for attribute captions is ‘is’. Hence the caption formed is given by <object is attribute>.

The other part of the problem is the interaction between various objects. This is done by mapping all the objects, their bounding boxes and relationships together. Then features are defined such as IOU, center to center distance, the locations of the object centers and relative size. These along with the relationships and label names

are fed into a classification model to get the most appropriate relationship. The output is the class of relationship (eg. man riding horse, here the output to be predicted is riding).The model that can be used is InceptionResNetV2.

IV Experimental Results

The object detection done using FasterRCNN+ InceptionResNetV2 gives us a Mean average precision or mAP of 0.58 whereas object detection done with SSD +MobileNetV2 gives a mAP of 0.37. The highest accuracy available for the Google AI Open Images - Visual Relationship track challenge is 0.28.

Object Detection Using	Accuracy
FasterRCNN+InceptionResNetV2	0.58
SSD +MobileNetV2	0.37

V Conclusion

The visual relationship model has been conceptualized however the entire implementation was not completed. The challenges faced include pre-processing the data set which is extremely large and require a lot of resources and computing power. The accuracy of the model can be greatly increased by making use of word embeddings. The concept of zero shot learning can be done using word embeddings which would help the field of visual relationship greatly as we can combine Natural Language Processing and CNN's to get a much more accurate visual relationship model. In this current age of information, it is becoming crucial for computers to be able to connect images with our language and image captioning is an important first step.

Acknowledgment

We would like to express our gratitude firstly to our mentor Dr. Gaurav for his guidance throughout the course of our project. We would also like to thank Bennett University as well as LeadingIndia.ai for giving us the opportunity to make this project reality as well as providing us with the resources to work on our project.

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Deep compositional question answering with neural module networks. In CVPR, 2016.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In ICCV, pages 2425–2433, 2015.

[3] Sadeghi, M.A., Farhadi, A.: Recognition using visual phrases. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 1745– 1752

[4] A. C. Berg, T. L. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, et al. Understanding and predicting importance in images. In CVPR, pages 3562–3569, 2012.

[5] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In CVPR, pages 3270–3277, 2014.

[6] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From´ captions to visual concepts and back. In CVPR, pages 1473–1482, 2015

[7] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In CVPR, pages 1–8, 2008.

[8] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. IJCV, 80(3):300–316, 2008.

[9] Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., Pinkal, M.: Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics 1 (2013) 25–36

[10] N. Prabhu and R. Venkatesh Babu. Attribute-graph: A graph based approach to image ranking. In ICCV, pages 1071–1079, 2015.

[11] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In CVPR, pages 1745–1752, 2011.

[12] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In ECCV, pages 746–760. Springer, 2012.

[13] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In CVPR, pages 5410–5419, 2017.

.