I.J. Mathematical Sciences and Computing, **2019, 2, 19-29**
Published Online April 2019 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijmsc.2019.02.02

***Available online at** http://www.mecs-press.net/ijmsc*

Category Specific Prediction Modules for Visual Relation Recognition

Sohan Chowdhury [a], Tanbirul Hashan [b], Afif Abdur Rahman [c], A.F.M. Saifuddin Saif [d]

[a,b,c,d] *Department of Computer Science, American International University-Bangladesh, Dhaka, Bangladesh*

Received: 18 November 2018; Accepted: 15 February 2019; Published: 08 April 2019

Abstract

Object classification in an image does not provide a complete understanding of the information contained in it. Visual relation information such as "person playing with dog" provides substantially more understanding than just "person, dog". The visual inter-relations of the objects can provide substantial insight for truly understanding the complete picture. Due to the complex nature of such combinations, conventional computer vision techniques have not been able to show significant promise. Monolithic approaches are lacking in precision and accuracy due to the vastness of possible relation combinations. Solving this problem is crucial to development of advanced computer vision applications that impact every sector of the modern world. We propose a model using recent advances in novel applications of Convolution Neural Networks (Deep Learning) combined with a divide and conquer approach to relation detection. The possible relations are broken down to categories such as spatial (left, right), vehicle-related (riding, driving), etc. Then the task is divided to segmenting the objects, estimating possible relationship category and performing recognition on modules specially built for that relation category. The training process can be done for each module on significantly smaller datasets with less computation required. Additionally this approach provides recall rates that are comparable to state of the art research, while still being precise and accurate for the specific relation categories.

Index Terms: Visual Relation Recognition, Deep Learning, Computer Vision.

1. Introduction

In an image, recognition of the objects in the world fails to provide a full understanding [1]. Visual relation information such as "car in front of the door" provides substantially more understanding than just "car, door". Speech description of "person behind the counter" helps the person navigate much better than "person,

*Corresponding author.
E-mail Address:

counter". To the visually impaired person, knowing the visual inter-relations of the objects can provide substantial insight for truly understanding their surroundings [2], this also applies to image recognition. Visual Relation of Objects refers to the contextual relations that two or more objects on an image can have besides their own classification. An object can be above, below, on the left or right of another object, this is the spatial relation of two or more objects. A person may be interacting with an object in some way (e.g. person ride cycle) this is the contextual or action based relation of multiple objects [13]. Today's state of the art systems can recognize and classify individual objects separately with near-human accuracy [4,5].

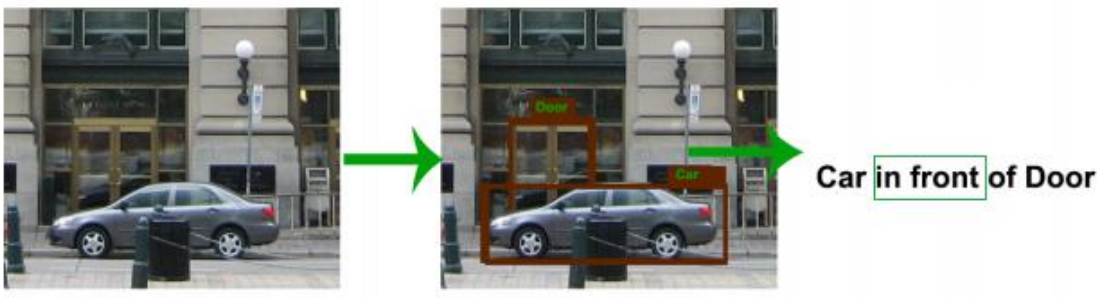

Fig.1. Visual Relation Recognition

The objects can be segmented to their respective pixels as well [6,7]. However, while identifying objects can provide some primary understanding of the image, a large amount of information remains unexplored. Extracting the spatial and visual relationships can provide a deeper understanding into the image. Understanding such relationships are especially crucial in fields such as autonomous driving, robotics (obstacle avoidance), manufacturing (object manipulation), aerial reconnaissance, human-computer interaction, etc.

In recent years, significant progress has been made in this area [2,3,6,8] by novel usages of Convolutional Neural Networks and Recurrent Neural Networks. The advancements in deep learning has opened up new ways to experiment in this domain. Larger datasets such as ImageNet, Visual Genome and Visual Relationship Dataset have made it possible for researchers to use more data-intensive methods of model training and the emergence of these datasets have brought more researchers to this topic.

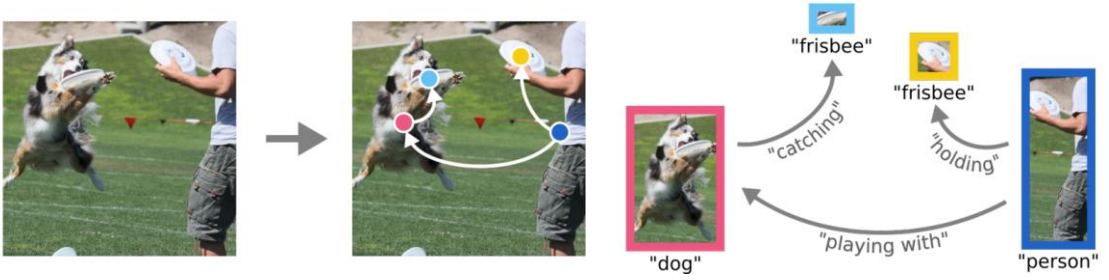

Fig.2. "Scene Graphs are Defined by the Objects in an Image (Vertices) and their Interactions (Edges). The Ability to Express Information about the Connections between Objects Make Scene Graphs a useful Representation for Many Computer Vision Tasks." [12]

Using deep learning methods combined with linguistic models, Fei-Fei Li, et al has achieved significant success and other works have followed this trend of combining linguistic knowledge with visual knowledge in their models. Some researchers considered object recognition and relation recognition as two separate tasks and then attempted to join these interpretations meaningfully [1,5] others approach it as a single task and perform direct image to object-relation retrieval [3]. Some researchers are now also using Reinforcement Learning

techniques to improve upon CONV Net models [4,9]. The most commonly used output representations of the visual relations in the image are natural language descriptions [1,5,10] and scene graphs [3,6,10] with objects as nodes and relations as their connecting edges. Most work in this area relies on the Visual Genome Dataset as it was one of the first large-scale datasets to contain contextual relationship labels along with other labels. Cewu Lu, et al have introduced a new dataset named Visual Relationship Dataset [5]

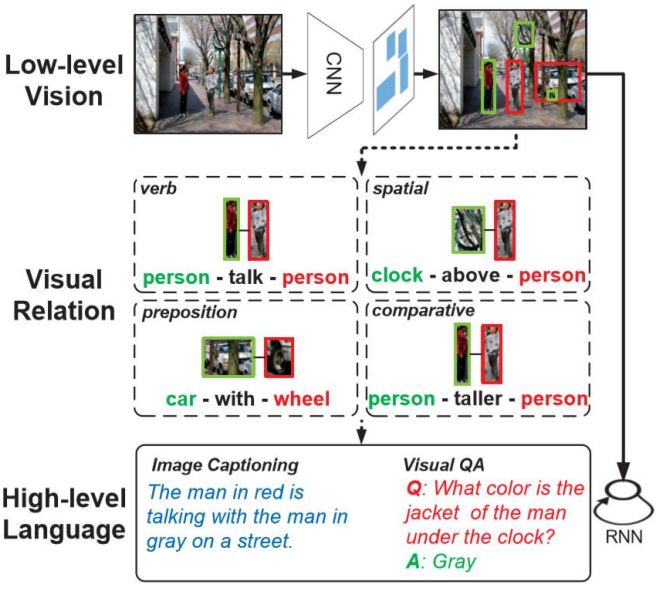

Fig.3. Visual Relations (Dashed Boxes) Presented as Natural Language Descriptions [14].

2. Related Work

The mature field of image recognition has produced some noteworthy datasets such as CIFAR-10, COCO, ImageNet etc. However, datasets which include visual relation information are not so numerous, and the Visual Genome dataset has been repeatedly used as a benchmark for many of the notable works [2,3,6].

This dataset is widely used for training CNN and also to measure and compare the performance and accuracy of approaches to ensure appropriate comparison. Some approaches which approach visual relation recognition as a separate task than object recognition require customization or creation of datasets, whereas single model approaches have used these datasets as raw input to their systems.

Methods benchmarked on the VRD and Visual Genome dataset can be compared directly, and it can be seen that in natural language representations of visual relations Zhang, et al [2] have shown improvements upon the results achieved by Lu, et al [9]. While, even though Xu, et al [6] have shown better performance and accuracy than Liang, et al [10] on graph-based representations of visual relations, their method requires RGB-D depth data, which was not required by the former, and limit the application of their method significantly.

Building upon these works, some have attempted at a speech representation of these outputs, Graves [1,2], Elamri, et al have shown some progress by applying these methods and then putting them through Text to Speech systems and achieved some notable results but much work remains to be done [15]. In this paper we attempt to overcome some of the problems of the mentioned works while also improving the general purpose usage of visual relation recognition.

Table 1. A Chronological Overview of Some Recent Notable Works and their Methods are presented as a Table Below:

Name	Method	Representation	Datasets
Karpathy, et al [12]	"deep neural network model that infers the latent alignment between segments of sentences and the region of the image that they describe"	Natural Language Descriptions	Flickr 8K, Flickr 30K and MS COCO
Lu, et al [9]	"learning visual appearance models for its objects and predicates and using the relationship embedding space learned from language"	Natural Language Caption	VRD and Visual Genome
Zhang, et al [2]	"VTransE Network builds upon an object detection module(e.g., Faster-RCNN), and then incorporates the proposed feature extraction layer and the translation embedding for relation prediction"	Natural Language Caption	VRD and Visual Genome
Liang, et al [10]	"language priors to build a directed semantic action graph G, where the nodes are nouns, attributes, and predicates, connected by directed edges that represent Semantic correlations with explicitly encode the semantic embeddings of previously extracted phrases in the state vector."	Scene Graph	VRD and Visual Genome
Xu, et al [6]	"model passes messages containing contextual information between a pair of bipartite sub-graphs of the scene graph, and iteratively refines its predictions using RNNs"	Scene Graph	Visual Genome and NYU Depth v2
Baier, et al [8]	"The model takes as input a raw image and combines it with a semantic prior, which is derived from the training data. Both types of information are fused, to predict the output, which consists of relevant bounding boxes and a set of triples describing the scene"	Natural Language Descriptions	VRD

3. Proposed Research Methodology

Most visual relation recognition models can have divided into two approaches, some researchers consider object recognition and relation recognition as two separate tasks and then attempt to join these interpretations meaningfully [1,6], others approach it as a single task and perform direct image to object-relation retrieval. We approach this task with a divide and conquer approach and attempt to break down the task of relation recognition into smaller, comparatively approachable computer vision tasks.

3.1. Proposed Model

We propose a multi-module model which combines object detection, object segmentation, masking, and relation category estimation, and a collection of fully connected Conv. Nets for relationship detection of two subjects. Most current research in this domain are using a single final module for prediction the relation of the

objects, while this is preferred from a performance perspective, it makes the work of the network very expansive. In that approach, for any combination of two objects, all relationship outcomes are considered and evaluated. But if human perceptions are considered, when recognizing the relation of two objects, humans do not generally consider all relations. There is an understanding of what limited combination of relations two subjects or objects can have depending on what the objects are. So we propose a model where first objects are detected from an image, then for each pair of objects the probability of contextual relations between them is considered using a linguistic subject, predicate, object probability analysis. Then only the most probable combination of relations is considered and a segmented image containing only the two objects in consideration are put through specialized modules or networks to detect which relation actually exists in the image.

3.2. Model

- Use Mask RCNN [17] to detect objects and segment them.
- Calculate all non-repeating pair combination of objects.
- Remove pairs that cross a proximity threshold.
- For each pair calculate which relation category is most probable using subject, predicate, object term frequency-inverse document frequency.
- Make new segmented images containing only the pairs, apply masking, and run the new images through the module specialized for detecting that relation category.
- Output the relationship with the highest confidence for each pair.

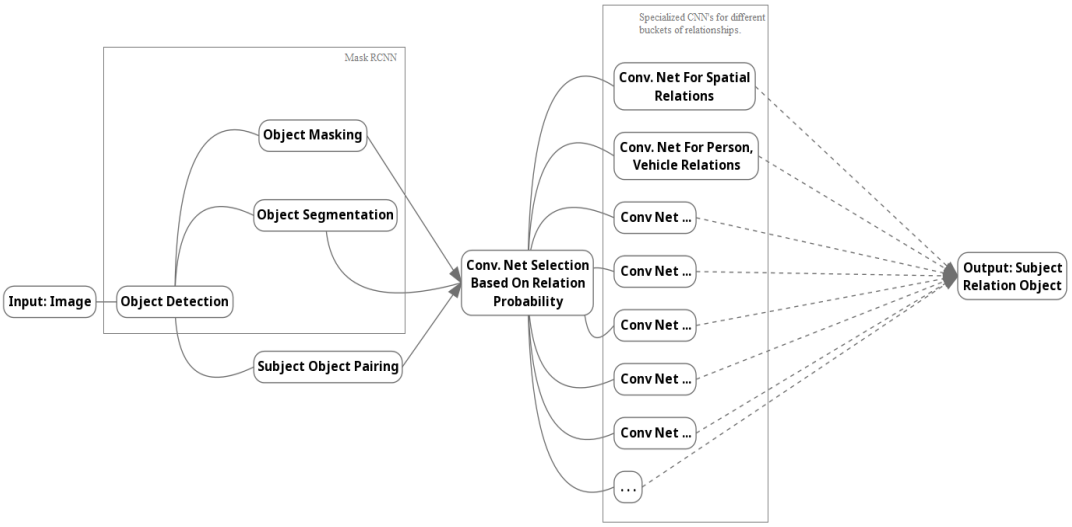

Fig.4. Object Segmentation, Linguistic Relation Category Estimation, and Relation Prediction Pipeline

3.3. Mask RCNN

Mask RCNN [17] by Kaiming He, Georgia Gkioxari, et al, is a flexible and general framework for object instance segmentation. Their approach efficiently detects objects in an image and also produces segmentation masks for each object.

3.4. Relation Category Estimation

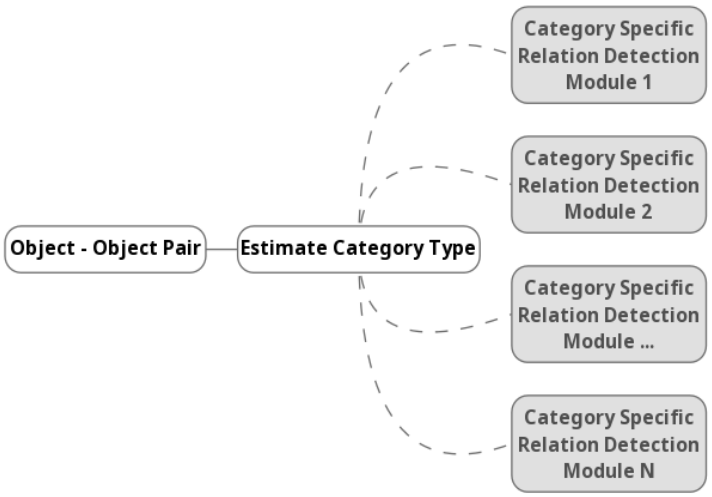

Fig.5. Linguistic Relation Category Estimation

When two objects are car and motorcycle, the most probable relation between them falls into one category: spatial. But if the objects are person and motorcycle, then the person may be riding it as well. Using term frequency-inverse document frequency for object, predicate, object probability calculation, given two objects, the probable category of the relation between them can be estimated.

3.5. Relation Category Specific Modules

For each relation category, a separate module is trained using segmented images of the two objects with transparent color masking applied for determining the order. For example, the spatial relation category prediction module has only four output confidences: front, behind, left, right

If we consider even these only 4 relations, combined with 11 selected relevant classes: The probable combinations are $11C2 \times 4 = 220$. But we break this task into two parts, detect from 11 object classes, and then detect from 4 relation classes. By applying such a divide and conquer model, the task becomes vastly less complicated, and accuracy increases as well.

Only the module for spatial relation recognition is trained and tested in this research.

4. Experimental Results and Discussion

4.1. Dataset

The Mask RCNN [17] model is trained on MS COCO; pre-trained weights are used for this module. For training relation recognition, a custom dataset is used by taking 27000 images using the MS COCO python API, and then validation is done on 526 images. The train, validation split is 48:1.

4.2. Evaluation

Our experiment is validated on 526 images where relation does not exist over 96 images and 480 images contain at least one relation. Our model identifies 390 relations accurately and 82 relations are wrongly detected.8 relations cannot be identified at all. Relation detection accuracy is 84.38%.

Table 2. Experimental Outcome overview

Images Tested	576
Relation exist	480
Relation does not exist	96
Relation detected correctly	390
Wrongly relation detected	82
The relation cannot be detected	8

Based on our findings we generate a binary matrix where we divide our findings into two classes such as relation exist and relation does not exist where wrongly detected relation considers as it identified as relation does not exist. Here "+" is for Relation exist and "-" is for Relation does not exist.

Table 3. Binary Metrics

		Predicted Class		Total instances
		+	-	
Actual Class	+	390	90	480
	-	0	96	96

Our Proposed model has an 84.38% detection rate, 81.25% Recall rate and 100% of Precision rate. Relation existence percentage is 83.33% where missed detection is only 18.05% and has a mAP of 0.74. Comparing with other models our proposed model outperforms all of them. Feature-based relation extraction with SVM [16] had a 49.5% of Precision and 63.1% of Recall where our model has an 81.25% Recall rate and 100% of Precision. Visual appearance module combined with language module [9] got 82.7% of Precision but have a low mAp of 0.592 where our model although got a low precision of 81.25% but higher mAP of 0.74.

Our model has a higher relation existence percentage of 83.33% where Feature-based relation extraction with SVM model [16] has a 70% relation existence.

Table 4. Detection Rate, Recall, Precision Rate, Mean Average Precision (MAP) of our Proposed Model and Existing Results based on the Parameters Mentioned

	Detection Rate	Recall Rate	Precision Rate	Mean Average Precision (mAP)
Our Model	84.38%	81.25%	100%	0.74
Feature-based relation extraction with SVM [16]	N/A	49.5%	63.1%	N/A
Variation-structured Reinforcement Learning[10]	N/A	13.34%	N/A	N/A
Visual appearance module combined with language module [9]	N/A	82.7%	N/A	.592

Table 5. Relationship Detection, Predication Detection, Relation Existence Percentage, Missed Detection Rate of our Proposed Model and Existing Results based on the Parameters Mentioned.

	Relationship Detection (MAP)	Predication Detection	Relation Existence Percentage	Missed Detection Rate
Our Model	0.74	0.813	83.33%	18.05%
Feature-based relation extraction with SVM [16]	N/A	N/A	70%	N/A
Visual appearance module combined with language module [9]	.592	N/A	N/A	N/A

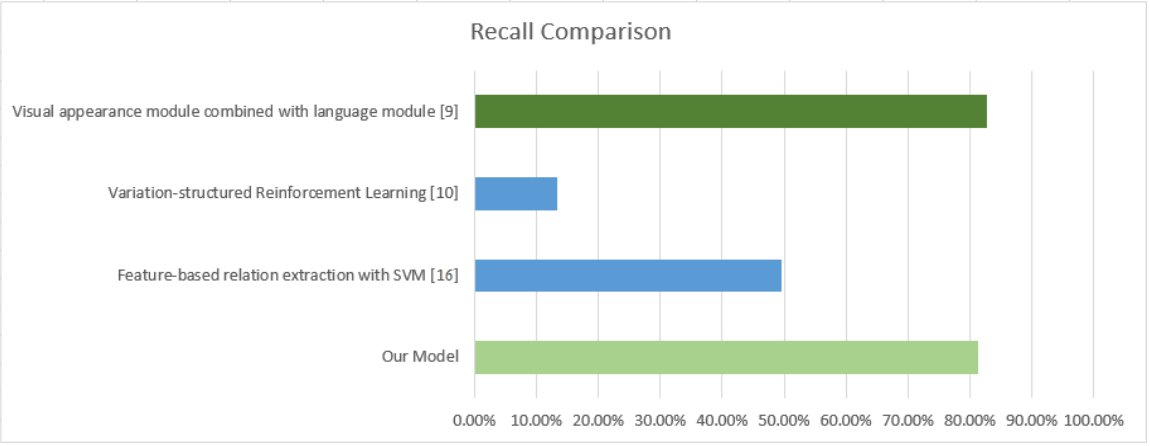

Fig.6. Recall Comparison of Some Relevant models

4.4. Discussion

From the results, we find that our divide and conquer model provides some observable benefits over previous models. Although we only test and train on spatial relations as a proof of concept, we can still see significant improvements. Using pre-trained detection and segmentation module provides state of the art results for object detection, segmentation and masking out of the box. Using these results relation category prediction can be performed. The introduction of separate modules for different categories of relations shows that for the spatial relations 84.38% accuracy can be achieved. The module especially shows robustness and does not detect relations where none exists; as such no false positives were detected. Our model (81.25%) outperforms Feature-based relation extraction with SVM (49.5%) [16] And is only slightly behind Visual appearance module combined with language module (82.7%) when recall rate is compared.

The advantage of our module is that the specialized modules have specific tasks with smaller output combinations and can be trained and tested on smaller models and smaller datasets that specialize in it. Such specialized modules are also more robust and provide better accuracy overall (e.g. our spatial module). But this approach has disadvantages as well, it cannot be one shot, and piping multiple modules together makes for a slower approach and compute heavy, especially if real-time vision applications are considered.

5. Conclusion

Visual relation information can provide important insight into an image and enable much such as “person left of dog” provides substantially more understanding than just “person, dog”. Description of “person behind the car” helps the person navigate much better than “person, counter”. Even to a visually impaired person, knowing the visual inter-relations of the objects can provide substantial insight for truly understanding their surroundings. We propose a model using recent advances in novel applications of Convolution Neural Networks (Deep Learning) combined with a divide and conquer approach to relation detection. The possible relations are broken down to categories such as spatial (left, right), vehicle-related (riding, driving), etc. Then the task is divided to segmenting the objects, estimating possible relationship category and performing recognition on modules specially built for that relation category. Our divide and conquer model provides some observable benefits over previous models. Although our model has a better score of relation detection but it was only tested as a proof of concept on a limited number of relation categories. Large scale relation detection module development and tests need be performed for further investigation. The model can also be improved by increasing the number of training instances. The advantage of our module is that the specialized modules have specific tasks with smaller output combinations and can be trained and tested on smaller models and smaller datasets. Such specialized modules are also more robust and provide better accuracy overall (e.g. our spatial module). But this approach has disadvantages as well, it cannot be one shot and piping multiple modules together makes for a compute heavy and slower approach.

5.1. Contribution of the Research

In this work, we propose a multi-module model which combines object detection, object segmentation, masking, and relation category estimation, and a collection of fully connected Conv. Nets for relationship detection of two subjects. Our proposed model first detects objects from an image, then for each pair of objects the probability of contextual relations between them is considered using a linguistic subject, predicate, object probability analysis. Then only the most probable combination of relations is considered and a segmented image containing only the two objects in consideration are put through specialized modules or networks to detect which relation actually exists in the image. By applying such a divide and conquer model, the task becomes vastly less complicated, and accuracy increases as well. Only the module for spatial relation recognition is trained and tested in this research.

5.2. Future Work

Although much work has been done in a small amount of time frame, many possibilities remain unexplored. In future work we are interested to develop a sophisticated trainable model for relation grouping and category estimation. Also develop modules for more relation category, such as person-vehicle, person-animal, person-interact-able-object etc. Our Mask RCNN trained on ms coco which can recognize only 80 types of object, this limits the capability of relation category estimation. Different object detection model that has more diverse recognizable set of objects could be used in our future work.

Acknowledgments

The authors wish to thank Dr. A.F.M. Saifuddin Saif for his guidance and supervision during the research period. This research has done in part of the completion of the BSc thesis of American International University-Bangladesh.

References

[1] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks. In International Conference on Machine Learning 2014 Jan 27 (pp. 1764-1772).

[2] Zhang H, Kyaw Z, Chang SF, Chua TS. Visual translation embedding network for visual relation detection. InCVPR 2017 Jul 1 (Vol. 1, No. 2, p. 5)

[3] Newell A, Deng J. Pixels to graphs by associative embedding. In Advances in neural information processing systems 2017 (pp. 2171-2180).

[4] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems 2015 (pp. 91-99).

[5] Sadeghi MA, Farhadi A. Recognition using visual phrases. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on 2011 Jun 20 (pp. 1745-1752). IEEE.

[6] Xu D, Zhu Y, Choy CB, Fei-Fei L. Scene graph generation by iterative message passing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017 Jul 1 (Vol. 2).

[7] Girshick R, Donahue J, Darrell T, Malik J. Region-based convolutional networks for accurate object detection and segmentation. IEEE transactions on pattern analysis and machine intelligence. 2016 Jan 1; 38(1):142-58.

[8] Baier S, Ma Y, Tresp V. Improving visual relationship detection using semantic modeling of scene descriptions. In International Semantic Web Conference 2017 Oct 21 (pp. 53-68). Springer, Cham.

[9] Lu C, Krishna R, Bernstein M, Fei-Fei L. Visual relationship detection with language priors. In European Conference on Computer Vision 2016 Oct 8 (pp. 852-869). Springer, Cham.

[10] Liang X, Lee L, Xing EP. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on 2017 Jul 21 (pp. 4408-4417). IEEE

[11] Yao B, Fei-Fei L. Modeling mutual context of object and human pose in human-object interaction activities. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on 2010 Jun 13 (pp. 17-24). IEEE.

[12] Chen T, Yu FX, Chen J, Cui Y, Chen YY, Chang SF. Object-based visual sentiment concept analysis and application. In Proceedings of the 22nd ACM international conference on Multimedia 2014 Nov 3 (pp. 367-376). ACM.

[13] Desai C, Ramanan D, Fowlkes CC. Discriminative models for multi-class object layout. International journal of computer vision. 2011 Oct 1;95(1):1-2.

[14] Ramanathan V, Li C, Deng J, Han W, Li Z, Gu K, Song Y, Bengio S, Rosenberg C, Fei-Fei L. Learning semantic relationships for better action retrieval in images. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 1100-1109).

[15] Rohrbach M, Qiu W, Titov I, Thater S, Pinkal M, Schiele B. Translating video content to natural language descriptions. In Proceedings of the IEEE International Conference on Computer Vision 2013 (pp. 433-440).

[16] GuoDong Z, Jian S, Jie Z, Min Z. Exploring various knowledge in relation extraction. In Proceedings of the 43rd annual meeting on association for computational linguistics 2005 Jun 25 (pp. 427-434). Association for Computational Linguistics

[17] He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. InComputer Vision (ICCV), 2017 IEEE International Conference on 2017 Oct 22 (pp. 2980-2988). IEEE.

Authors' Profiles

Sohan Chowdhury is an undergraduate (UG) student of Computer Science and Software Engineering under the Faculty of Science and Information Technology of American International University Bangladesh. His research interest and passion mostly focuses on but is not limited to Deep Learning, Computer Vision and Image Context Extraction. He has achieved first and second positions in multiple programming contests and hackathons worked with successful startups and worked on government projects. He is currently working as an Associate Software Engineer at Telenor Health AS.

Tanbirul Hashan is an undergraduate (UG) student of Computer Science and Engineering under the Department of Science and Information Technology of American International University Bangladesh. His research interest focuses but not limited to Image Processing, Computer Vision, Virtual Reality, Machine Learning, and Artificial Intelligence. He is currently working as a Software Engineer at Nazdaq Technologies.

Afif Abdur Rahman is an undergraduate (UG) student of Computer Science and Engineering under the Department of Science and Information Technology of American International University Bangladesh. His research interests and passion are mostly based on Computer Vision and Pattern Recognition, Image processing, Color Segmentation, Machine Learning, Object Tracking, and Detection. He is currently worked as an Intern at Robi Axiata Limited.

A.F.M. Saifuddin Saif received Ph.D. from Faculty of Information Science and Technology, University Kebangsaan Malaysia (UKM) in 2016. He received M.Sc. in Computer System Engineering (Software System) from University of East London, UK, and B.Sc. (Eng.) degree in Computer Science and Engineering from Shahjalal University of Science and Technology, Bangladesh in 2012 and 2008, respectively. Most of his contributions in Computer Vision and Artificial Intelligence Research field were published in ISI Q1 journals. He has published many papers in ISI indexed Journals; Scopus indexed Journals, Book Chapters, Conferences, and Proceedings. He served as Technical Committee Members, Reviewers, Guest Speakers, Session Chairs in many Conferences and Workshops. Currently, he is an Assistant Professor at **Faculty of Science and Information Technology**, American International University Bangladesh (AIUB). Before joining the university, he did Post Doctorate at Faculty of Information Science and Technology, University Kebangsaan Malaysia. He spent more than 6 years in IT industry such as Advanced Software Development, Web eMaze etc as IT researcher. His research interests include Image Processing, Computer Vision, Artificial Intelligence, Augmented Reality, 3D Reconstruction, and Medical Image Processing.

How to cite this paper: Sohan Chowdhury, Tanbirul Hashan, Afif Abdur Rahman, A.F.M. Saifuddin Saif,"Category Specific Prediction Modules for Visual Relation Recognition", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.5, No.2, pp.19-29, 2019.DOI: 10.5815/ijmsc.2019.02.02