# Visual Relationship Representation using GCN-LSTM

Mr. Sandip J. Murchite[1], Ms. Ashwini A. Gat[2]
[1]Assistant Professor, [2]Assistant Professor

**Abstract** This paper is about exploring visual relationship detection. This is situated as a significant task in computer vision. The goal is to explore all visual relationships in a given image between objects. Visual relationship detection targets to entitle the interactions between pairs of objects. Modeling visual relationships between objects would be needful for describing an image. This paper represents new approach to explore visual relationships between objects for labelling image using encoder and decoder attention based framework. To complement the representation ability of visual presence, we integrate Graph Convolutional Neural Networks and Long Short-Term Memory architecture that will discover both semantic and spatial visual object relationships. In this approach graphs are constructed over the detected objects based on spatial and semantic relationships. The representation of each objects are filtered by graph using graph convolutional networks(GCN). With this graph, long short-term memory(LSTM) framework is applied for image captioning together with memory attention mechanism for textual description. Experiments are conducted on Visual relationship datasets, visual genome and COCO image captioning dataset.

**Keywords:** Visual Relationship modeling, Image Captioning, Graph Convolutional network, Long short-term memory

## 1. Introduction

The eventual objective of image understanding system is to generate unambiguous and real representation like 3D of the visual environment. To achieve these goal lots of techniques has been discovered and improved in recognizing, detecting and exploring all objects within an image. The recent advances in deep learning object recognition and object detection has been dramatically improved. These advances are convincingly demonstrating high capability visual relationship models particularly for image captioning. [1,2]. As shown in Figure 1 (a), a traditional Convolutional Neural Network(CNN) is used to encrypt an image with attention mechanism to generate textual description, one word at each time interval. In Figure 1 (b) Region based Convolutional Neural Network(RCNN) is used to encrypt an image, further detected regions from image is filtered with Graph based Convolutional Neural Network(GCNN) with attention mechanism to generate sentences.

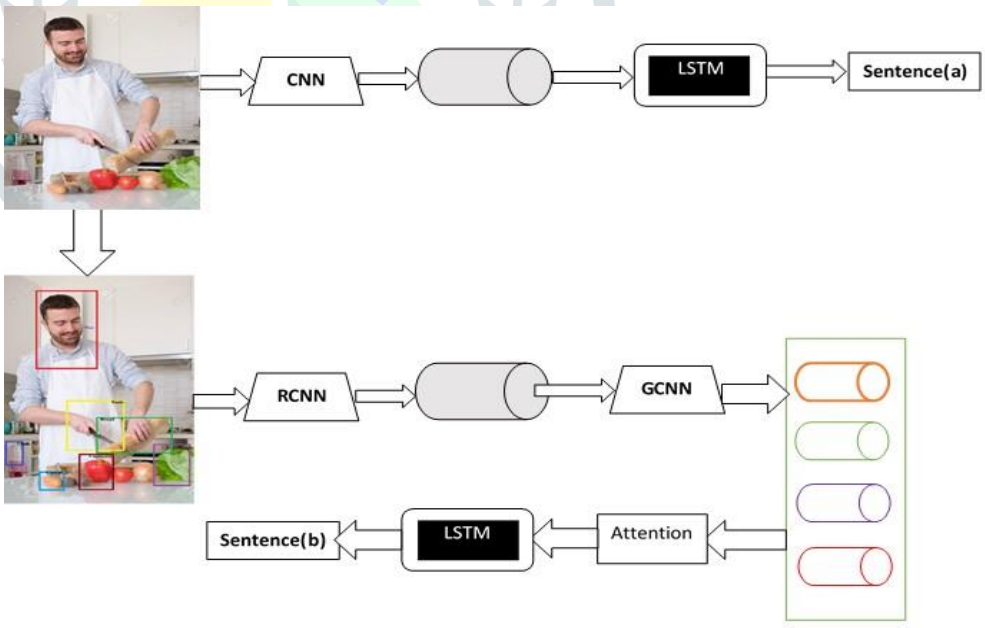


**Figure 1. Visual Representation (a) Traditional CNN plus LSTM, (b) Our GCNN plus LSTM**

Visual relationships discover the interactions between detected objects from an image. It also classifies the interactions called as predicate between each pair of objects. In particular, detected visual relationships can be represented in the form of triplets <subject; predicate; object>. Detecting visual relationship means localizing and recognizing objects, e.g. man playing football or girl eating ice-cream. These triplets are very important because it is used to build various natural language processing about image like question answering. Previous work has addressed in this problem is visual relationship is critical in understanding semantic relationship. Furthermore, the detected objects could be in large scale, and at various locations in an image from diverse categories face trouble in determine the type of relationships. In this approach basic aim is to develop relationship on both sematic and spatial category and merge them into image encrypted to generate relation based area level representations. LSTM will then use attention mechanism for sentence decoding with more power.

For modeling visual representation for image captioning, Graph based Convolutional Neural Network is used along with LSTM for attention mechanism as shown in Figure 1 (b). Faster RCNN is implemented to detected various regions present in an image. Detected regions are then feed forward to semantic graph with directed edges where vertices entitle each detected section and edge represents the semantic relationship called predicates between each pair of region which is predicted by Visual Genome. [3] Likewise, spatial graph assembled on detected areas represent geometrical relationship. After, detected relation aware region is submitted to individual attention mechanism i.e. LSTM image decoder for sentence generation. At last at each time interval sentence with highest probability is considered for the next step. The main involvement of this application is to use in an image captioning problem which are not up till now fully discovered.

## 2. Related Work

### Image Captioning

Deep learning is very extensive field now with so many applications like image captioning. Image captioning is process of generating scene information from an image based on objects and relationship between objects. [1] [2] [4] [5] [6] [7]. This problem utilizing CNN and RNN model to represent sentences both spatially and semantically. Vinyals et al. implemented a neural network framework by adding LSTM for sentence generation from an image [5], which further proceeded to attention mechanism [4] to automatically predict a word when generating a corresponding sentence. Recently, semantic relationships are playing vital role in image captioning problem when it is used in CNN plus RNN model and such semantic relationships enhancing image captioning very well. In recent times, encoder-decoder based model [2] is implemented to detect image regions along with attention mechanism via bottom up method and detected regions are then used for sentence generation with top down method.

### Visual Relationship Representation

In computer vision detection of visual relationship has attracted lot of researchers. Earlier some researcher [9] [10] effort to learn some spatial relationships like “inside”, “around”,” above” and “below” to improve sentence generation. Later, semantic relationships like actions and communications between objects are exploited in [16] [17] where each probable groupings of semantic relationship are taken as one pictorial category and representation measured as classification task. Recently, in a few works [11] [12] [13] [14] [15] deep learning based neural network frameworks are used for visual relationship representations. [15] consider visual correlation as directed edges to show relationship between the objects for processing scene graph in time interval way.

Our proposed approach belongs to sequence learning model for image captioning. Related to earlier techniques [2] [8], GCN-LSTM implement attention mechanism on detected image regions of objects for textual description. The earlier approach further exploited on both spatial and semantic relationship between objects for image captioning which is not been beforehand implemented. Thus quality of generating sentences through GCN-LSTM is increased tremendously.

## 3. Exploring visual relationship

In this approach image description is generated by Graph Convolutional Networks and Long Short-term Memory(GCN-LSTM) architecture by additionally exploring both semantic and spatial relationship between substances. This model firstly detects objects within images viz object detection module, detected objects are encoded and generate the whole image into set of image regions containing detected objects. All detected image regions of objects are feed forward to GCN to construct semantic and spatial relation graphs based on their semantic and spatial connection between objects. Furthermore, training is performed by encoding whole image region set via GCN which results in relation aware region representation. All relation aware regions are submitted to LSTM based attention mechanism framework for textual description.

### 3.1 Problem formulation

We have an image $I_m$ which is designated by sentence S, where S = $\{wd_1,wd_2,wd_3.....wd_n\}$ and wd is considered as n words. R-CNN detects image regions R which furthermore considered as vertex V. we can build semantic graph Gsem= (V, E) and spatial graph Gsp= (V, E) where E refers set of edges between detected regions. Semantic and spatial graphs are encoded into relation aware region representation and lastly decoded each target output word through LSTM attention mechanism for textual description.

### 3.2 Semantic Object Relationship

In deep learning visual relationship representation is considered as classification task [16]. Semantic relationship is represented by triplets <subject; predicate; object> between pair of substances. Semantic relationship is represented in directional way like one object linked with other object through establish which can be action or communication between pair of objects.

### 3.3 Spatial Object Relationship

Semantic object only describes action and interaction between pair of objects, spatial relations between the detected image regions are unexplored. So, spatial graph constructed over detected image regions to exploit spatial relation between image region from an image

## 4. Image Captioning

Semantic and spatial graphs of detected image regions are integrated into sequence learning to learn visual relationship using LSTM region based attention mechanism. GCN based image encoder is designed by using graph convolutional network node classification [15] and semantic classification [2] which captures semantic and spatial relations based on semantic and spatial graph shown in Figure 1. The earlier Graph convolutional networks are undirected graph and constructed by using features of its neighbors. To construct directed graph earlier GCN is updated to represent direction relation between the detected image regions. Furthermore, region level attention mechanism [2] is used to derive our attention LSTM decrypted by submitting all relative aware region features to LSTM attention mechanism.

In training phase, two types of graph constructed by exploring semantic and spatial relation from detected image regions. Each graph is the separately used to train one GCN based image encrypted and LSTM attention image decrypted. LSTM attention image decoder is utilized by using cross entropy loss.

## 5. Results and Experiments

We trained our model by using COCO captioning dataset and evaluated GCNLSTM model on COCO captioning dataset for sentence generation. Furthermore, visual genome is used for preprocessing of image encoder to detect semantic relation.

**COCO Captioning Dataset**

For object discovery, separation and image captioning most widely used dataset is COCO dataset. It is used in very large-scale. This is most popular dataset for sentence generation, which contains almost 84,678 training images and 40,890 testing images. COCO dataset also having human annotated descriptors.

**Visual Genome**

Visual Genome is image dataset, knowledge base for representing action and interaction between detected image regions. It contains almost 109K images with annotations, attributes and relationships. Visual Genome also embedding COCO dataset to train RCNN.

Figure 2 and 3 represents sentence generated using COCO dataset.

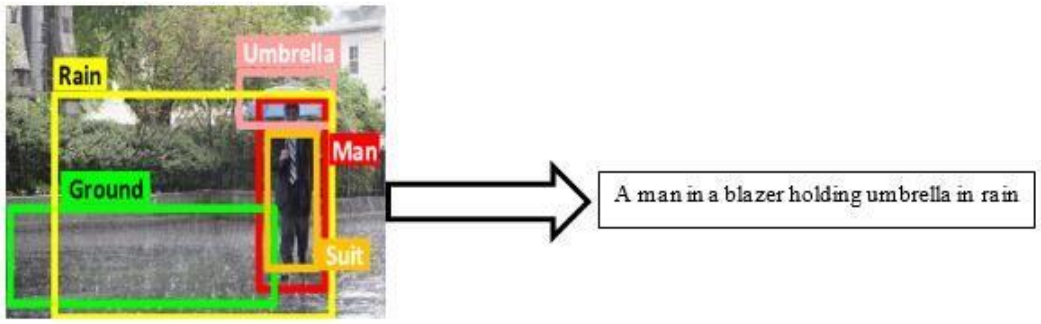


**Figure 2. Sentence generated using COCO dataset**

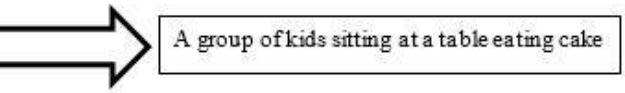


**Figure 3. Sentence generated using COCO dataset**

## 6. Conclusion

We proposed Graph Convolutional Network along with LSTM memory(GCN-
LSTM) which representing visual relationships required for image captioning. Problem is considered from interactions present between objects or regions to generate sentence. Earlier approach either considering spatial relationship only this approach exploring both semantic and spatial relationship. Experiments are directed on COCO plus Visual Genome dataset to test our proposal.

**References**

[1] D. J, "Long-term recurrent convolutional networks for visual recognition and description.," CVPR, 2015.

[2] V. O, "Show and tell: A neural image caption generator," CVPR, 2015.

[3] X. K, "Show, attend and tell: Neural image caption generation with visual attention," ICML, 2015.

[4] Y. T, "Incorporating copying mechanism in image captioning for learning novel objects.," CVPR, 2017.

[5] Y. T, "Boosting image captioning with attributes.," ICCV, 2017.

[6] Y. Q, "mage captioning with semantic attention.," CVPR, 2016.

[7] F. K, "Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts.," PAMI, 2017.

[8] G. S, "Multi-class segmentation with relative location prior.," IJCV, 2018.

[9] D. B, "Detecting visual relationships with deep relational networks.," CVPR, 2017.

[10] S. M. A., "Recognition using visual phrases," CVPR, 2011.

[11] A. P., "Bottom-up and top-down attention for image captioning and visual question answering," CVPR, 2018.

[12] L. Y, "Scene graph generation from objects, phrases and region captions.," ICCV, 2017.

[13] L. C., "Visual relationship detection with language priors.," ECCV, 2016.

[14] P. B. A., "Phrase localization and visual relationship detection with comprehensive imagelanguage cues.," ICCV, 2017.

[15] X. D., "Scene graph generation by iterative message passing.," CVPR , 2017.

[16] D. S. K., "Learning everything about anything: Webly-supervised visual concept learning.," CVPR, 2014.

[17] G. C., "Object categorization using cooccurrence, location and appearance.," CVPR, 2008.

[18] M. D., "Encoding sentences with graph convolutional networks for semantic role labeling.," EMNLP, 2017.

[19] K. T. N., "Semi-supervised classification with graph convolutional networks.," ICLR, 2017.

[20] L. T. Y., "Microsoft coco: Common objects in context.," ECCV, 2014.