Neural Motifs: Scene Graph Parsing with Global Context

Purvanshi Mehta

November 5, 2019

1 Summary

1.1 Problem

The paper analyses the importance of motifs in context to the task of scene graph generation. The question asked is whether scene graphs possess motifs and use the information to give better and more accurate scene graphs.

1.2 Innovation

The authors do an extensive data analysis and use the obtained results to design an architecture suitable for the task. The various data analysis techniques show that encoding global context is very important for a scene graph generation. Their main innovation is the **Stacked Motif Network**, where they make predictions in a hierarchical fashion - the model detects the bounding boxes, then they predict the bounding box labels and finally the relations.

1.3 Contributions

There are two major contributions of the paper -

- **Data Analysis** - An extensive study has been done on the *Visual Genome Dataset*. the two questions answered through the analysis are -
  - How different types of relations correlate with different objects
  - How higher order graph structures recur over different scenes?

  To answer the first part, the authors divide the objects and relations into high-level types. Using those high level types, the distribution of relation types between objects was visualized. Three major results were shown through this analysis - The clothing and parts are related through possessive relations, furniture and building entities are related to geometric relations and semantic relationships are headed by people. This shows that

Common Sense priors play an important role in generating accurate scene graphs

To answer the second question we need to answer how much information is gained by knowing the identity of different parts in a scene graph. This experiment shows that *Object labels are highly predictive of edge labels but not vice versa*

- **Stacked Motif Network** The data analysis shows significant reasons to include global context for scene graph generation. the model first detects the bounding box using Faster - RCNN. The results from the bounding box are linearized to output the *Object Context*. the feature from the bounding box are sent to a Bi directional LSTM for object detection. Next step is to predict the edge context by giving as input the object labels. This ask is done using another biLSTM. The edge and object labels are then combined to predict the relationship label.

1.4 Evaluation

two frequency baselines have been deployed -

- FREQ - Look up emperical distribution over relationships between objects o_i and o_j as computed in the training set
- FREQ-OVERLAP - requires two boxes intersect to count as a valid relation.

The thing to note here is that the baseline performs better than previous state of the art results. State art of the art results can be seen with MotifNet - LeftRight which shows the emperical importance of global context.

1.5 Substantiation

The authors provide a strong baseline which even beats the previous state of the art results. The authors claimed the importance of context and global features which was clearly proved by both data analysis results and the use of BiLSTMs. The authors also claim the robustness of their model towards the rol ordering scheme. They also highlight on why the network is failing in certain cases - error due to object detection and some ambiguity in the data tagging.

2 Review

- In this approach objects and relations are treated as different entities(predicted in a hierarchical manner) whereas they are related : Relations exist between objects thus the distribution of relations is not independent of the objects. We need to treat this as a multimodal classification problem where we are fusion the modalities - (the two objects and the relation)

and then finally take out individual losses so that information is shared amongst them. The proposed architecture is -

The three modalities are - Object1, object2 and the relation. After feature extraction from the three modalities, they would be fused (though point-wise multiplication, simple concatenation - different techniques could be tried. The fused information would be used to find the loss for the relation. The other two losses - for object1 and object2 would not include the fused quantity as the object detection task is independant of the relation detection task).

$$L_{final} = L_{Object1} + L_{Object2} + L_{Relation}$$

where $L_{Relation}$ comes from the fusion of the three modalities.

- **Can the spatial information of objects be leveraged** All the previous approaches detect the object and then try to predict the relationship. The question to ask here is - *Are the relations independent of the spacial/ scene or the picture.* Though the paper talks about taking global context through passing an BiLSTM, but can this global context be more explored in terms of depth relations or placement of objects rather than just the objects themselves.

  To explore this idea we need to design a **Semantic Embedding tool** which includes information like depth of the images, relative position to each other. One way of doing this is providing depth maps as another mode of input along with the objects.

- **Attention mechanism** Can we use attention mechanism over the obtained objects to determine relations between them? Given an object and a relation we can infer about the other object. Example - Given a shirt and wear the model should know there is a human involved. The key idea is to use the cyclic information to better predict the labels.