Real World Graphical Reasoning and Compositional Question Answering

Henry Friedlander, Preston Ng
{hnf035,plng}@stanford.edu

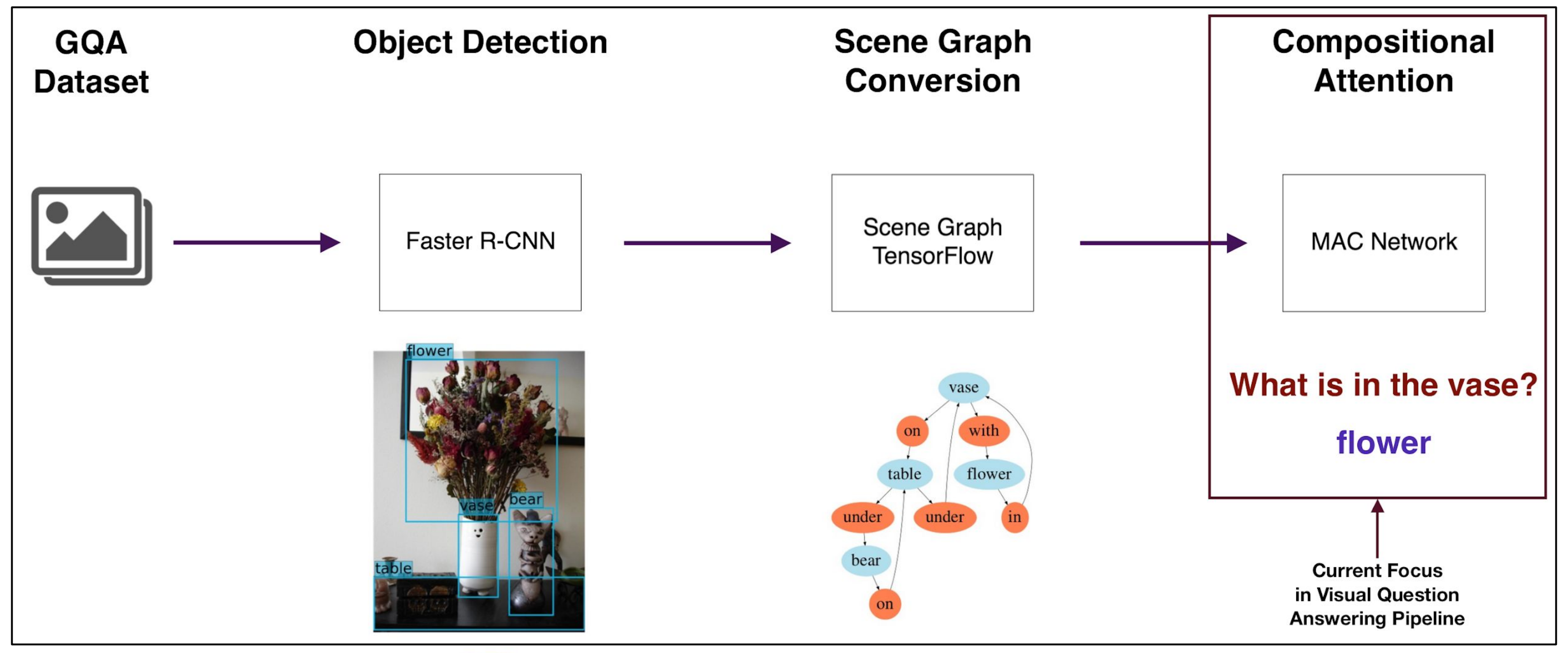
Special thanks to Dor Arad Hudson and CS 224N TAs for their advice and help on our project.

Problem

Problem: There has been much progress in recent years in the field of visual question answering. Many models are running at super-human performance, and progress appears promising. The MAC network's compositionality-inspired architecture[1] has achieved a startling 98.9% accuracy on the CLEVR[2] dataset. Unfortunately, this does not generalize to real world examples as well. Recent papers have been shown that these 'superhuman' models are not learning as much as we, human operators, would like. In the case of the MAC network on the CLEVR dataset, MAC is simply leveraging CLEVR's very small answer space.

Goal: In this research project, we propose modifications to the MAC network by utilizing scene graphs to avoid this overfitting so that it can work in a more general setting. We chose the versatile and real world GQA dataset[3] as a test set to benchmark our results. To emphasize Natural Language aspects of machine reasoning, we decided to focus on operating over the scene graph (SG) data of GQA rather than images.

Data and Task


GQA Dataset

Object Detection

Scene Graph Conversion

Compositional Attention

Faster R-CNN

Scene Graph TensorFlow

MAC Network

What is in the vase?

flower

Current Focus in Visual Question Answering Pipeline

Figure 1 (above): A Newly Proposed End to End GQA Pipeline in which we focus on the last aspect of compositional attention operating on scene graphs that are generated from images.

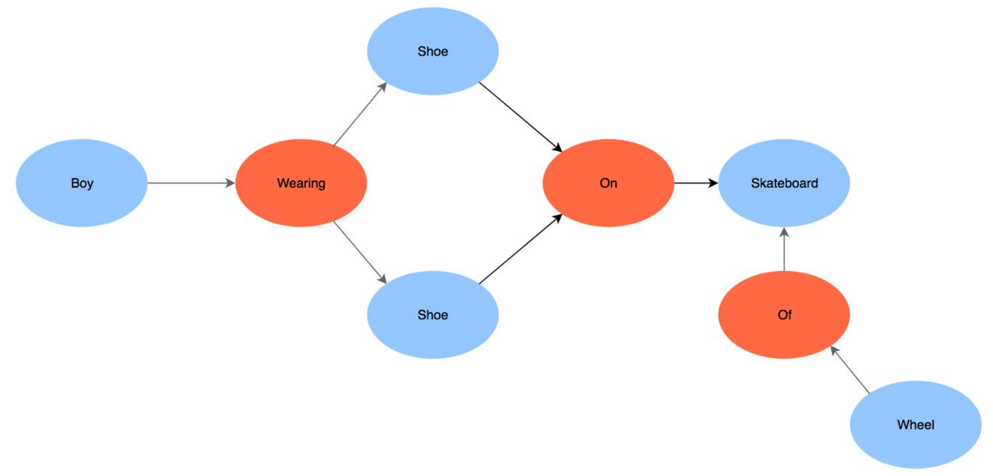
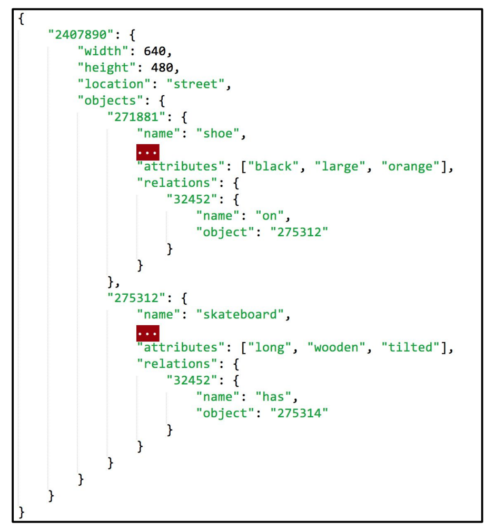


Figure 2 (above): Example scene graph representation showing the object relationships. The nodes in blue are are considered objects, whereas the nodes in orange are descriptive relationships between object nodes.

Data: Scene graphs of real world images, provided by the GQA dataset (74939 training scene graphs, 10696 validation scene graphs)

- Objects = list of objects in image's scene graph
  - Attributes = attributes of parent object
  - relations = relationships to other child object nodes
    - name = relationship type
    - object = child object node



```
{
    "2407890": {
        "width": 640,
        "height": 480,
        "location": "street",
        "objects": {
            "271881": {
                "name": "shoe",
                ...
                "attributes": ["black", "large", "orange"],
                "relations": {
                    "32452": {
                        "name": "on",
                        "object": "275312"
                    }
                }
            },
            "275312": {
                "name": "skateboard",
                ...
                "attributes": ["long", "wooden", "tilted"],
                "relations": {
                    "32452": {
                        "name": "has",
                        "object": "275314"
                    }
                }
            }
        }
    }
}
```

Approach

The base MAC architecture is comprised of a recurrent network of MAC cells, each containing a control and memory (read/write) unit. We propose the following changes:

- **Control unit**: Utilize an LSTM in order to incorporate control information from all previous timesteps rather than c_{i-1}
- **Input to Read Unit:** Convert the read unit input to operate over scene graphs (SG) instead of images that utilize a node emphasized embedding

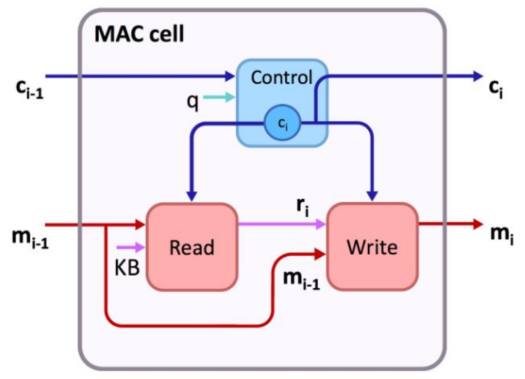


Figure 3 (above): Base MAC Architecture

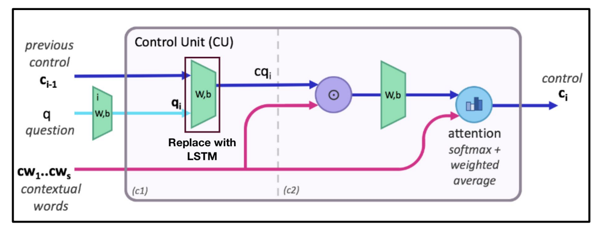


Figure 4/5 (left/right): MAC Modifications to the Control / Knowledge Base, respectively

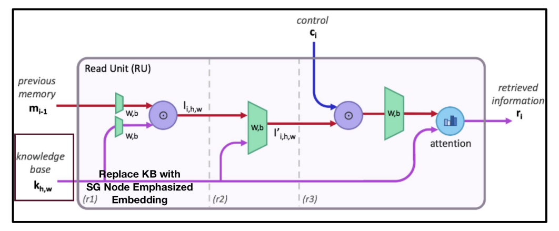
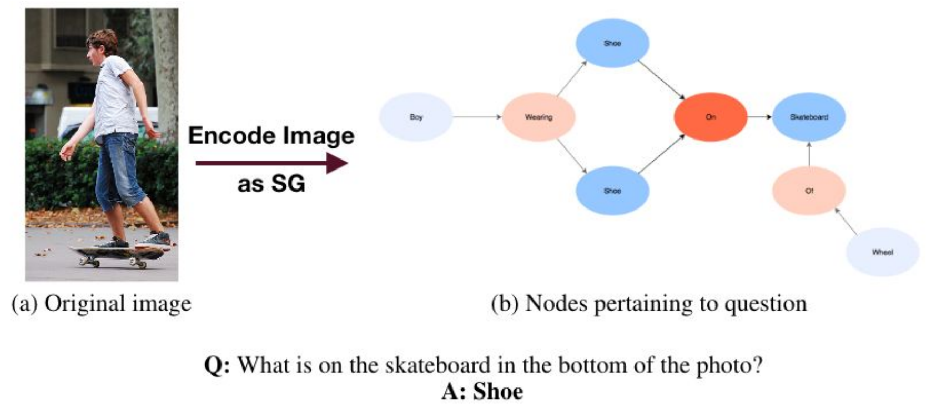


Figure 6 (right): Example of an image's scene graph attention when the new MAC architecture is asked the question "What is on the skateboard in the bottom of the photo?". The current MAC architecture now attends to object nodes rather than image regions.

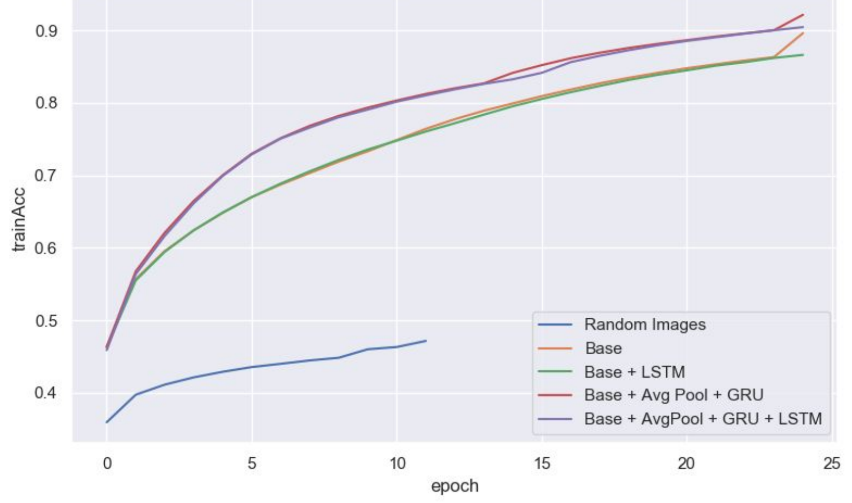


(a) Original image

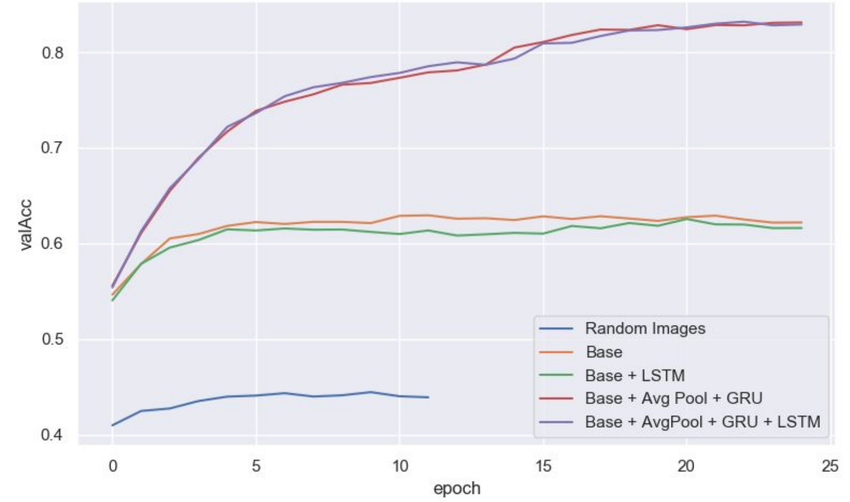
(b) Nodes pertaining to question

Q: What is on the skateboard in the bottom of the photo?
A: Shoe

Results



New Base Model (MAC + SG)	**Training Accuracy**	**Validation Accuracy**
Random Images	47.17%	43.90%
Base + Flatten	85.34%	62.90%
Base + Flatten + LSTM	83.13%	62.13%
Base + Avg Pool + GRU	**92.17%**	83.11%
Base + Avg Pool + GRU + LSTM	89.56%	**83.19%**



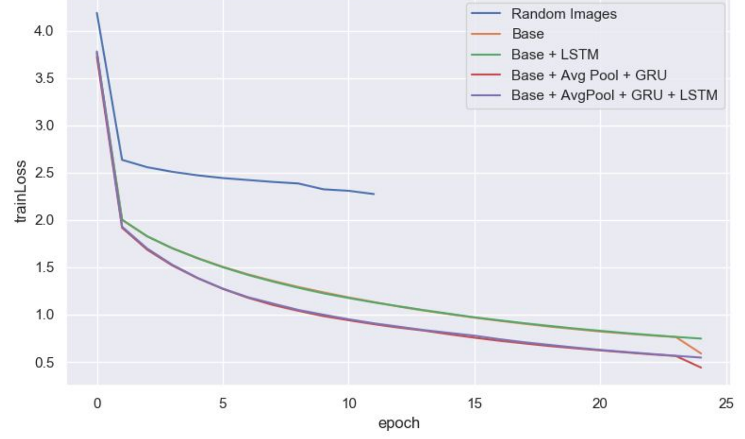
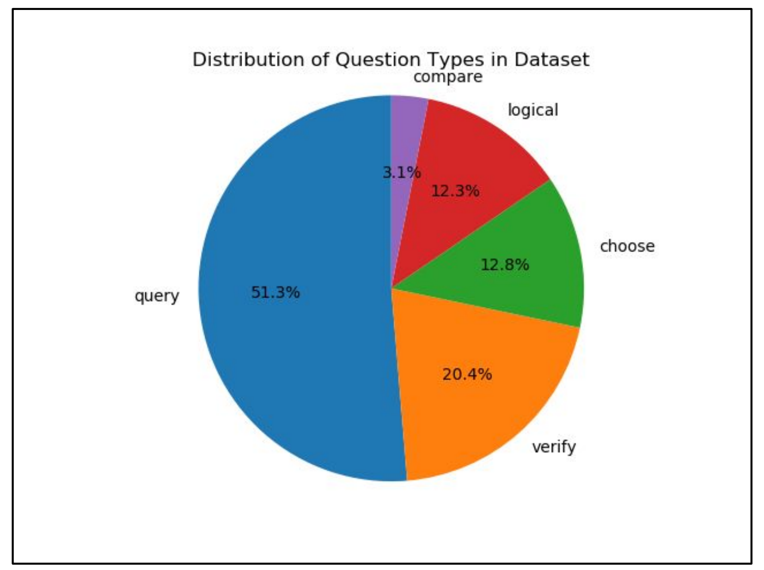


Table 1 and figure 7/8/9 (above): Accuracy and Training curves showing that converting the MAC network to operate on scene graphs enabled the system has a 21.28% increase in validation accuracy (83.19%) over the previous model that operated on images (61.91%). This is a major step towards enabling a better VQA system as a whole on datasets that have real world images, but is only part of the process. Because we are operating on fully observed scene graphs, it is expected to have a large increase in accuracies.

Analysis



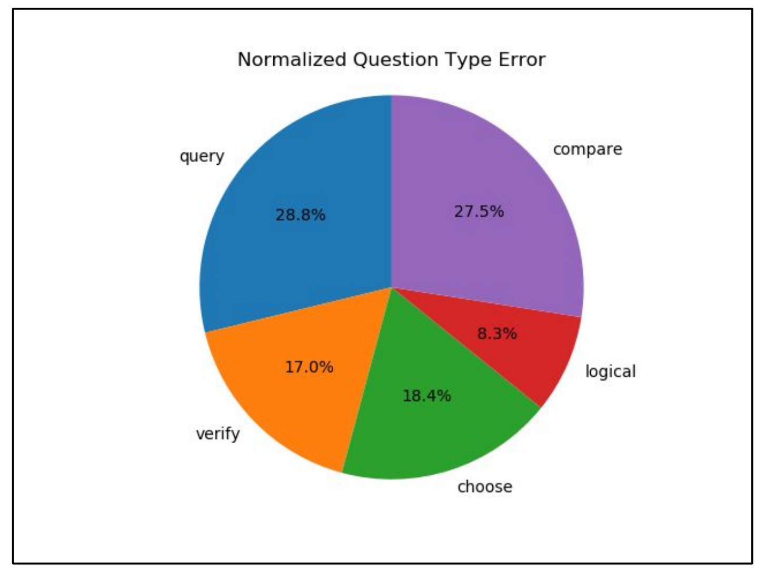


Figure 10/11 (above): Distributions of various question types and their normalized error rates (10696 validation scene graphs, 10000 questions). Various question types include *query*, *verify*, *choose*, *logical*, and *compare*. There is an overwhelmingly larger percentage of *query* questions (e.g. What is in the vase?) which require a free form response. In addition, we find that *query* type of questions have the largest error rate (normalized).

Figure 12 (left): Example of wrong answer, due to the upstream system that created the scene graph with green curtains, which may show that the MAC architecture submodule is more robust to scene graphs than it seems.

Question: What color are the curtains?
Prediction: Green
True Answer: Blue

Conclusions / Future Steps

Conclusion:
We have introduced changes to the Memory, Attention, and Composition (MAC) network that now operates on scene graphs in order to perform machine reasoning. By operating on scene graphs, the MAC network can now perform graphical reasoning on real world examples. We have replaced the image input in the read unit to encoded knowledge graphs for a particular image. By replacing images with node emphasized embeddings, individual nodes are attended to rather than image regions. Limitations to our work include not having a full end to end system, but rather a submodule to a whole system for visual reasoning.

Future Work:
For future work, we want to enable the system to work on new image input by implementing an image to scene graph system, and then passing the scene graph input to the current MAC network that operates on scene graphs. By doing so, we will complete an end to end model that can perform real world visual reasoning and compositional question answering.

References

[1] Drew A Hudson and Christopher D Manning. 2018. Compositional attention networks for machine reasoning

[2] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning (CVPR)

[3] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for compositional question answering over real-world images. Conference on Computer Vision and Pattern Recognition (CVPR)