Carnegie Mellon University

Scene Graph Generation and Reasoning for Visual Question Answering

Anirudh Kannan

Mentor : Syeda Akter

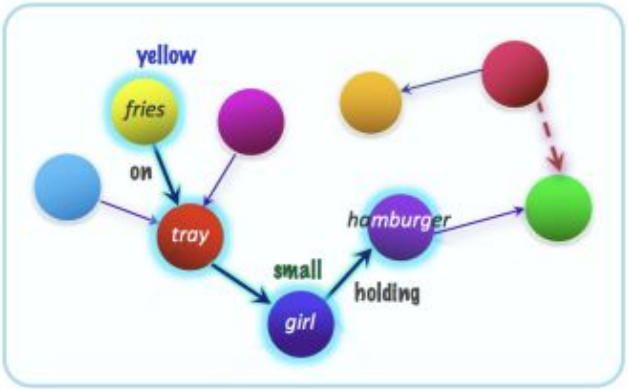
Visual Question Answering

- Visual Question Answering (VQA) is a task in computer vision and natural language processing that involves answering questions about visual information, such as images or videos, using natural language.

- The system takes an image and a question in natural language as input and produces an answer in natural language as output.

- VQA requires both visual and language understanding to reason about the relationship between the image and the question and generate an appropriate answer.

What are Scene Graphs?



Scene Graphs are structured representations of an image, where entities and associated attributes are embedded as nodes and their relationships as edges. Scene Graphs have shown great potential in enhancing the ability of VQA models to understand the complex relationships between objects in an image, which in turn helps to answer questions that demand advanced spatial reasoning and counting skills

Hudson, Drew A., and Christopher D. Manning. "Gqa: A new dataset for real-world visual reasoning and compositional question answering." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Carnegie Mellon University

Scene Graph Generation

- The goal of Scene Graph Generation (SGG) is to synthesize a graphical representation of a given input image.

- This involves identifying objects within the image (nodes) and determining relationships between these objects (predicates).

- This process enables a more comprehensive understanding of visual content and has numerous applications, including image retrieval and visual reasoning.

Motivation

- VQA models can be used to reduce visual barriers for visually impaired individuals by allowing them to get information about images from the web and the real world. However these models face the challenge of understanding the context of images in order to accurately answer questions related to them.
- Scene graphs provide a structured representation of objects, their attributes, and relationships within an image.
- By utilizing scene graphs, VQA systems might be able disentangle various visual factors and reduce the impact of irrelevant information, leading to improved generalization across different settings and greater interpretability of the system.

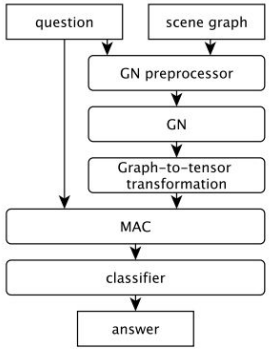
Hypothesis / Research Questions

- Can incorporation of scene graphs improve the performance of VQA systems?
- Can VQA performance benefit from utilizing question-specific scene graph information?
- Can the performance of VQA systems be enhanced by using a pipelined approach that involves scene graph generators?

Carnegie Mellon University

Literature Review

- (Lee et al., 2019) was one of the first works to explore the use of scene graphs for the Visual Question Answering task.
- The authors encoded a scene graph and a question using GNNs and then fed the encoded graph as an image to the question to the Memory, Attention, and Composition (MAC).
- Their study found that augmenting the question answering model with representations learned from the auxiliary scene graph processing task, where graph neural networks were utilized to learn condensed embeddings, led to significant improvement in performance over existing baselines.



Soohyeong Lee, Ju-Whan Kim, Youngmin Oh, and Joo Hyuk Jeon. 2019. Visual question answering over scene graph. In 2019 First International Conference on Graph Computing (GC), pages 45–50

Carnegie Mellon University

Literature Review

- Along similar lines, (Zhang et al., 2019) performed an empirical study on the use of only scene graphs as the context for question answering and concluded that "scene graphs, even automatically generated by machines, can definitively benefit Visual QA **if paired with appropriate models like GNs**."

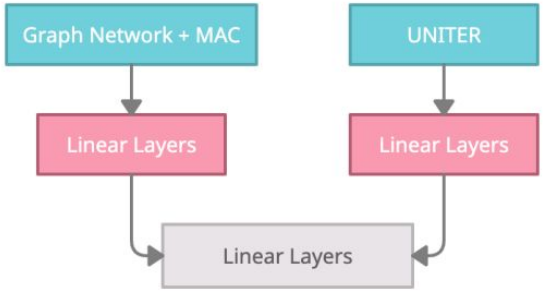
- Instead of combining image and language features, this paper proposes using scene graphs derived from images for Visual QA. They use graph networks to encode the scene graph and perform structured reasoning based on input questions. Empirical studies show that scene graphs can capture essential image information and graph networks have potential to outperform current Visual QA methods with a simpler architecture.

Cheng Zhang, Wei-Lun Chao, and Dong Xuan. 2019. An empirical study on leveraging scene graphs for visual question answering.

Literature Review

- In a study by Damodaran et al. (2021), an extensive analysis was conducted on the use of scene graphs for visual question answering (VQA). The researchers employed a late fusion approach between graph neural networks (GNNs) and the MAC model, using scene graph inputs alongside features from UNITER. Their findings highlight the potential benefits of utilizing scene graphs in VQA tasks.

- Moreover, the study investigated the impact of using ground truth versus generated scene graphs on the model's performance. The results revealed a notable difference in performance between the two, with ground truth scene graphs outperforming the generated ones. This observation bears significance as it aligns with similar patterns observed in this project.



UNITER with Scene Graphs

Vinay Damodaran, Sharanya Chakravarthy, Akshay Kumar, Anjana Umapathy, Teruko Mitamura, Yuta Nakashima, Noa Garcia, and Chenhui Chu. 2021. Understanding the role of scene graphs in visual question answering.

Carnegie Mellon University

Dataset

- GQA Dataset was used as the dataset in this project.
- 22M questions compositional questions involving a diverse set of reasoning skills
- Images and scene graphs sourced from Visual Genome.
- Each image comes with a scene graph to represent its semantics.
- Questions synthesized using a question engine leveraging the scene graph information derived from Visual Genome.
- Each question is associated with a functional program which lists the series of reasoning steps needed to be performed to arrive at the answer.

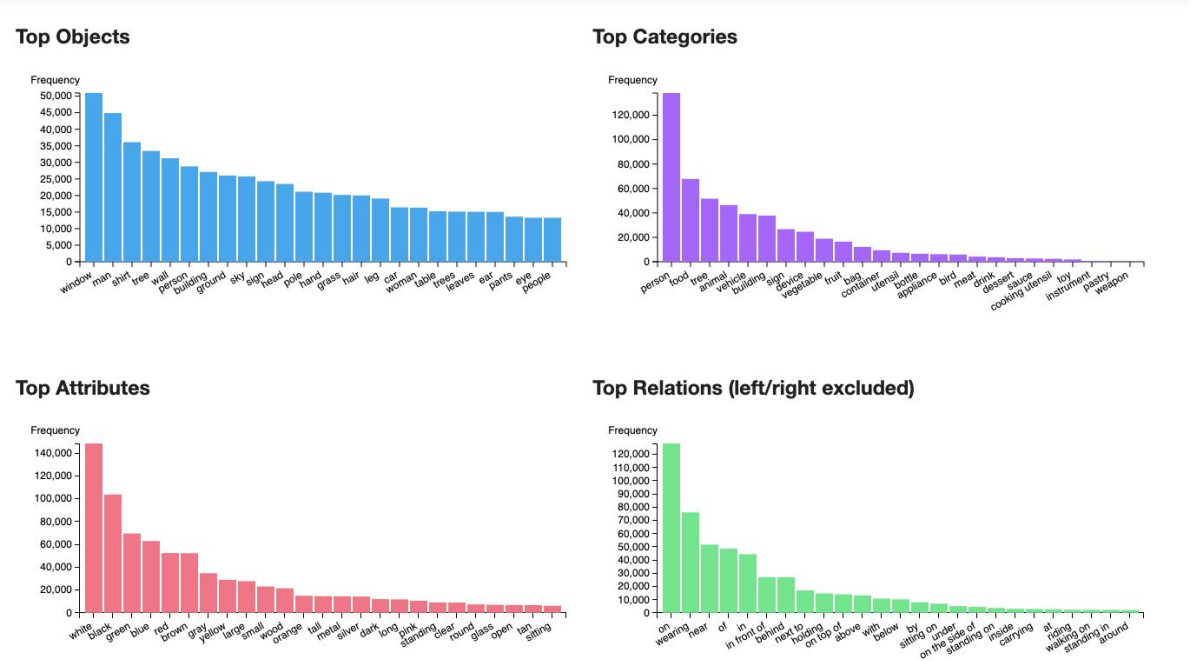
A1. Is the **tray** on top of the **table** black or light brown? light brown
A2. Are the **napkin** and the **cup** the same color? yes
A3. Is the small **table** both oval and wooden? yes
A4. Is there any **fruit** to the left of the **tray** the **cup** is on top of? yes
A5. Are there any **cups** to the left of the **tray** on top of the **table**? no
B1. What is the brown **animal** sitting inside of? **box**
B2. What is the large **container** made of? cardboard
B3. What **animal** is in the **box**? **bear**
B4. Is there a **bag** to the right of the green **door**? no
B5. Is there a **box** inside the plastic **bag**? no

Examples of questions from the GQA dataset

https://cs.stanford.edu/people/dorarad/gqa/about.html

Carnegie Mellon University

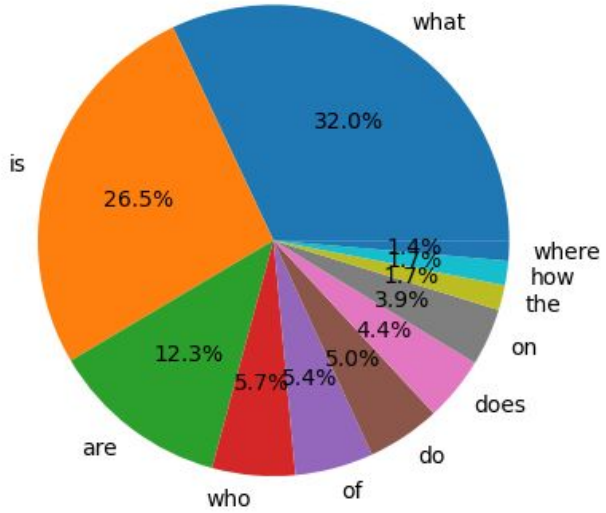
Scene Graph Statistics



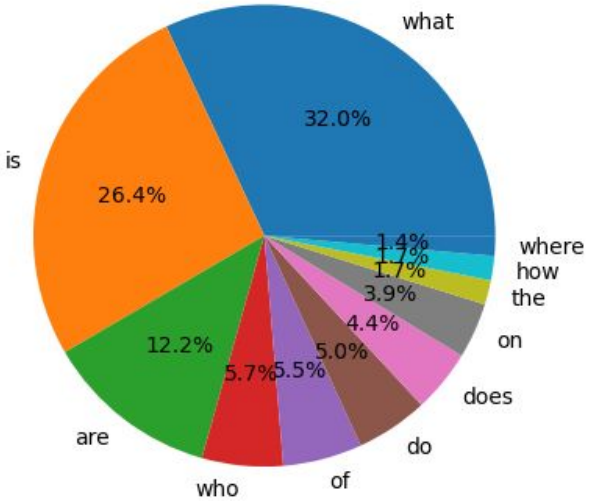
https://cs.stanford.edu/people/dorarad/gqa/vis.html

Carnegie Mellon University

Question Type Distribution



Validation Set



Training Set

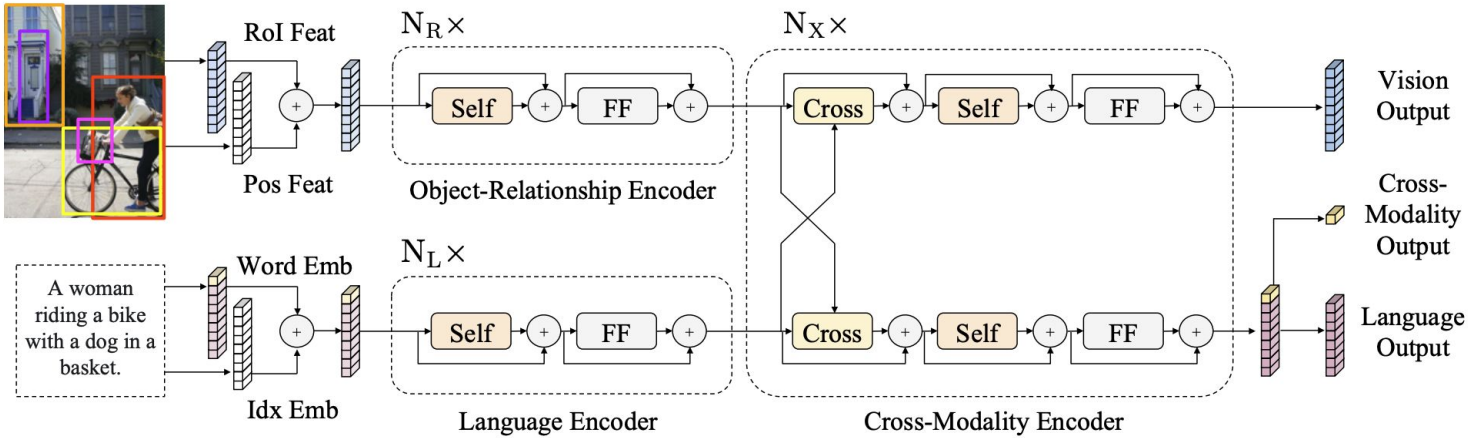
Carnegie Mellon University

QUESTION COUNTS

Question Type	Validation	Train
what	42267	301452
is	34982	249212
are	16204	115227
who	7510	53475
of	7098	52272
do	6649	47583
does	5818	41428
on	5190	36646
the	2265	16110
how	2204	15996
where	1822	13248

Carnegie Mellon University

Baseline



LXMERT: Learning Cross-Modality Encoder Representations from Transformers

Tan, Hao, and Mohit Bansal. "Lxmert: Learning cross-modality encoder representations from transformers." arXiv preprint arXiv:1908.07490 (2019).

Carnegie Mellon University

LXMERT

- The LXMERT model is constructed using a Transformer architecture, which incorporates three distinct encoders: an object relationship encoder, a language encoder, and a cross-modality encoder.

- Pretrained with diverse pre-training tasks large-scale datasets of image-and-sentence pairs.
  1. masked cross-modality language modeling
  2. masked object prediction via RoI-feature regression
  3. masked object prediction via detected-label classification,
  4. cross-modality matching
  5. image question answering

Tan, Hao, and Mohit Bansal. "Lxmert: Learning cross-modality encoder representations from transformers." arXiv preprint arXiv:1908.07490 (2019).

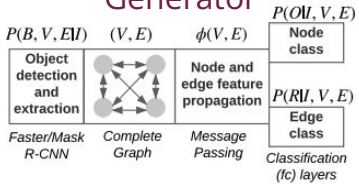
Why LXMERT ? 🤔

- **Cross-modal approach**: LXMERT is designed to process both language and visual inputs together, which makes it well-suited for tasks that require understanding the relationship between images and text.
- **Pretrained on large datasets**: LXMERT is trained on large-scale datasets such as COCO and Visual Genome, which allows it to capture a broad range of visual concepts and linguistic knowledge.
- **High accuracy**: LXMERT has achieved state-of-the-art performance on several VQA benchmarks, including VQA 2.0, GQA, and VizWiz. (Ranked #3 on GQA test-dev accuracy)
- **Well documented codebase!**

Pipeline Architecture

Input Image

Scene Graph Generator

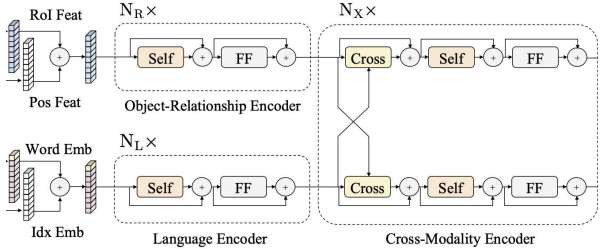


Input Question

What is the woman in front of ?

VQA Model

LXMERT



Output Answer

statue

Carnegie Mellon University

Addition of Gold Standard Scene Graphs

Results

S. No.	Model	Training Accuracy	Validation Accuracy	Dev Test Accuracy without scene graphs
V1	Finetuned on training and validation set	-	94.58 %	59.73 %
V2	Finetuned on training set	92.50 %	**83.73 %**	59.65 %
V3	Finetuned on training set + SG	92.53 %	**83.82 %**	59.37 %
V4	Finetuned on training set + SG	92.47 %	**83.78 %**	59.34 %
V5	Finetuned on SG + question of training set	23.53 %	19.91 %	52.76 %
V6	Finetuned on question + SG (context specific) of training set	95.24 %	**88.83 %**	58.89 %

Carnegie Mellon University

Results

S. No.	Model	Training Accuracy	Validation Accuracy	Dev Test Accuracy without scene graphs
V1	Finetuned on training and validation set	-	94.58 %	59.73 %
V2	Finetuned on training set	92.50 %	**83.73 %**	59.65 %
V3	Finetuned on training set + SG	92.53 %	**83.82 %**	59.37 %
V4	Finetuned on training set + SG	92.47 %	**83.78 %**	59.34 %
V5	Finetuned on SG + question of training set	23.53 %	19.91 %	52.76 %
V6	Finetuned on question + SG (context specific) of training set	95.24 %	**88.83 %**	58.89 %

Carnegie Mellon University

Scene Graph Preprocessing

4 different methods of scene graph preprocessing were explored

- Addition of entire Scene Graph as context after question (v3)
- Grouping by the object nodes and improving grammatical structure (v4)
- Addition of entire Scene Graph as context before question (v5)
- Addition of question specific context information from the scene graph (v6)

V3

Original Question

What is this bird called?

Question after adding Scene Graph context

What is this bird called? (palm tree, green leaves, tropical bushes, black eye, white ruffled feathers, silver chain-link fence, stone, stone rock wall, stone, stone, tall green palm tree, sharp curved black open pointy beak, white parrot, small white puffy clouds, blue sky, tall green palm tree, palm tree, bushes, hillside, palm yellow leaves, tan straw, rock, palm tree, rock wall, rock, rock behind parrot, rock on top of hillside, rock to the left of leaves, rock to the right of feathers, rock to the left of bushes, rock to the left of bushes, rock to the right of parrot, rock of palm tree, rock to the right of parrot, rock to the right of bushes, rock near straw, rock to the right of palm tree, rock to the right of hillside ...

Results

S. No.	Model	Training Accuracy	Validation Accuracy	Dev Test Accuracy without scene graphs
V1	Finetuned on training and validation set	-	94.58 %	59.73 %
V2	Finetuned on training set	92.50 %	**83.73 %**	59.65 %
V3	Finetuned on training set + SG	92.53 %	**83.82 %**	59.37 %
V4	Finetuned on training set + SG	92.47 %	**83.78 %**	59.34 %
V5	Finetuned on SG + question of training set	23.53 %	19.91 %	52.76 %
V6	Finetuned on question + SG (context specific) of training set	95.24 %	**88.83 %**	58.89 %

Carnegie Mellon University

V4

- Removing redundant information
- Group by object name
- Adding words like “is”, “are” to improve grammatical structure

Original Question

What is the woman in front of?

Question after adding Scene Graph context

What is the woman in front of? (talking sitting woman is to the left of white pocket, turned head is to the left of white collar, dark long hair and red hair is to the left of red mouth, white phone is to the right of arm, statue is behind talking sitting woman, arm is to the left of turned head, white phone is to the left of pink lips, arm is to the left of pink lips, talking sitting woman is to the left of white collar, turned head is to the right of arm, red mouth is to the right of turned head, talking sitting woman is by statue, talking sitting woman is wearing shirt, ...)

Results

S. No.	Model	Training Accuracy	Validation Accuracy	Dev Test Accuracy without scene graphs
V1	Finetuned on training and validation set	-	94.58 %	59.73 %
V2	Finetuned on training set	92.50 %	**83.73 %**	59.65 %
V3	Finetuned on training set + SG	92.53 %	**83.82 %**	59.37 %
V4	Finetuned on training set + SG	92.47 %	**83.78 %**	59.34 %
V5	Finetuned on SG + question of training set	23.53 %	19.91 %	52.76 %
V6	Finetuned on question + SG (context specific) of training set	95.24 %	**88.83 %**	58.89 %

Carnegie Mellon University

V5

- Append context before question

Original Question

What is this bird called?

Question after adding Scene Graph context

(palm tree, green leaves, tropical bushes, black eye, white ruffled feathers, silver chain-link fence, stone, stone rock wall, stone, stone, tall green palm tree, sharp curved black open pointy beak, white parrot, small white puffy clouds, blue sky, tall green palm tree, palm tree, bushes, hillside, palm yellow leaves, tan straw, rock, palm tree, rock wall, rock, rock behind parrot, rock on top of hillside, rock to the left of leaves, rock to the right of feathers, rock to the left of bushes, rock to the left of bushes, rock to the right of parrot, rock of palm tree, rock to the right of parrot, rock to the right of bushes, rock near straw, rock to the right of palm tree, rock to the right of hillside ...)
What is this bird called?

Results

S. No.	Model	Training Accuracy	Validation Accuracy	Dev Test Accuracy without scene graphs
V1	Finetuned on training and validation set	-	94.58 %	59.73 %
V2	Finetuned on training set	92.50 %	**83.73 %**	59.65 %
V3	Finetuned on training set + SG	92.53 %	**83.82 %**	59.37 %
V4	Finetuned on training set + SG	92.47 %	**83.78 %**	59.34 %
V5	Finetuned on SG + question of training set	23.53 %	19.91 %	52.76 %
V6	Finetuned on question + SG (context specific) of training set	95.24 %	**88.83 %**	58.89 %

Carnegie Mellon University

V6

- Extracting only question specific information from Scene Graph

Original Question

Are there napkins under the utensil to the left of the rice?

Question after adding Scene Graph context

Are there napkins under the utensil to the left of the rice? (silver utensil, utensil, rice, rice on plate, rice to the right of utensil, rice to the right of napkin, rice to the left of utensil, utensil to the left of plate, utensil to the left of garnish, utensil to the left of rice, utensil to the left of meat, utensil above napkin, utensil on top of napkin, utensil to the right of rice, utensil to the right of meat, utensil to the right of plate)

Results

S. No.	Model	Training Accuracy	Validation Accuracy	Dev Test Accuracy without scene graphs
V1	Finetuned on training and validation set	-	94.58 %	59.73 %
V2	Finetuned on training set	92.50 %	**83.73 %**	59.65 %
V3	Finetuned on training set + SG	92.53 %	**83.82 %**	59.37 %
V4	Finetuned on training set + SG	92.47 %	**83.78 %**	59.34 %
V5	Finetuned on SG + question of training set	23.53 %	19.91 %	52.76 %
V6	Finetuned on question + SG (context specific) of training set	95.24 %	**88.83 %**	58.89 %

Carnegie Mellon University

Graph Density-Aware Losses for Novel Compositions in Scene Graph Generation

- This paper highlights two main challenges that hinder the generalization ability of models in the scene graph generation task.

- Firstly, the commonly used loss function in SGG unintentionally favors dense scene graphs, resulting in the neglect of edges in sparse graphs during training. This is problematic as such edges contain important few-shot examples that are crucial for achieving good generalization performance.

- Secondly, the frequency of relationships in the dataset creates a strong bias, where a model that blindly predicts the most frequent relationship can achieve good performance. This bias is exploited by some state-of-the-art models to improve results, but it can also hinder their ability to generalize to rare compositions.

Knyazev, Boris, et al. "Graph density-aware losses for novel compositions in scene graph generation." *arXiv preprint arXiv:2005.08230* (2020).

Graph Density-Aware Losses for Novel Compositions in Scene Graph Generation

- Authors use separate loss functions for foreground edges (those that are annotated) and background edges (those that do not have any annotation) in a graph and express them as a function of graph densities.

$$\mathcal{L} = \mathcal{L}_{node} + d\mathcal{L}_{FG} + (1-d)\mathcal{L}_{BG}.$$

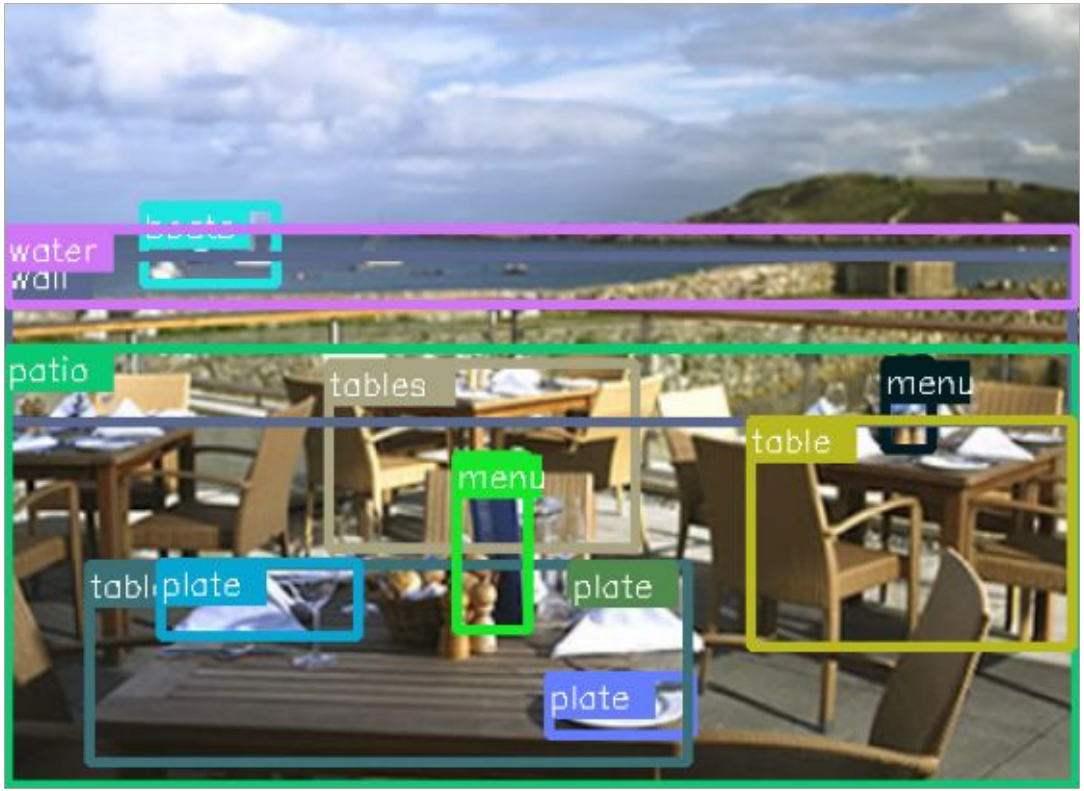
- The purpose of the BG edge type in the object detection task is similar to that of a "negative" class. The BG edges represent pairs of nodes that do not have any meaningful relationship or connection. By training the model on these BG edges, it learns to distinguish between positive and negative edges, which helps it avoid labeling all pairs of nodes as "positive" at test time.
- Used a Faster RCNN Detector and Resnet 50 Backbone for object detection and Neural Motifs as the SGG baseline.

Zellers, Rowan, et al. "Neural motifs: Scene graph parsing with global context." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

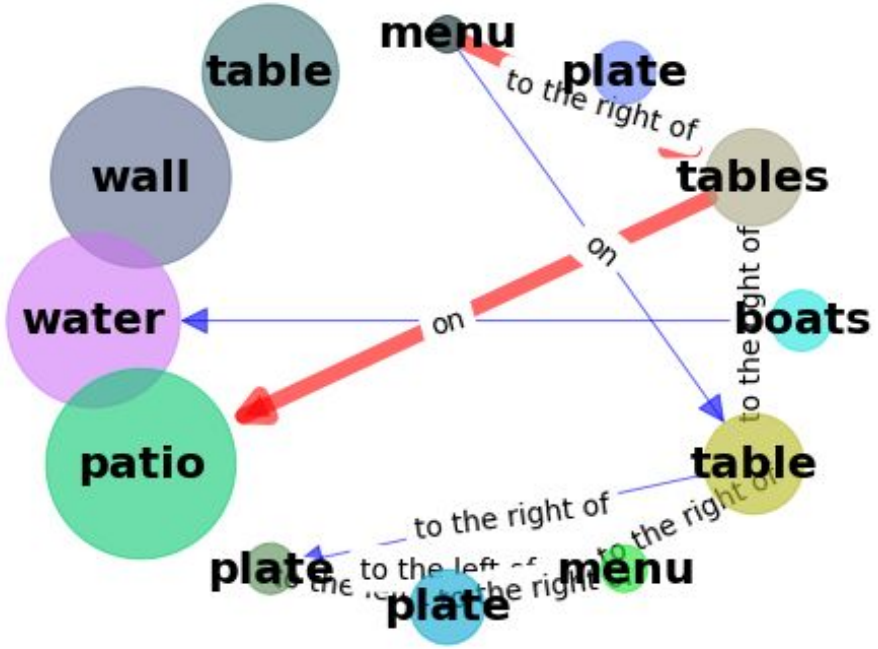
Knyazev, Boris, et al. "Graph density-aware losses for novel compositions in scene graph generation." *arXiv preprint arXiv:2005.08230* (2020).

Carnegie Mellon University

Outputs



Carnegie Mellon University



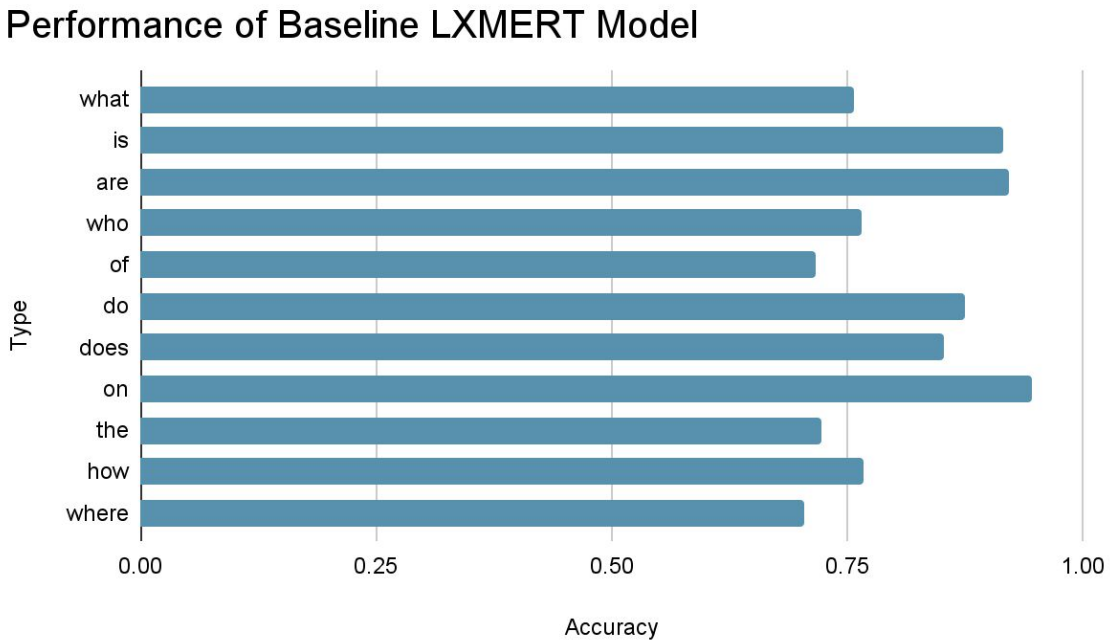
Carnegie Mellon University

Results

S. No.	Model	Training Accuracy	Validation Accuracy	Dev Test Accuracy
V1	Finetuned on training and validation set	-	94.58 %	59.73 %
V2	Finetuned on training set	92.50 %	**83.73 %**	59.65 %
V3	Finetuned on training set + SG	92.53 %	**83.66 %**	59.63 %
V6	Finetuned on question + SG (context specific) of training set	95.24 %	**88.65 %**	59.58 %

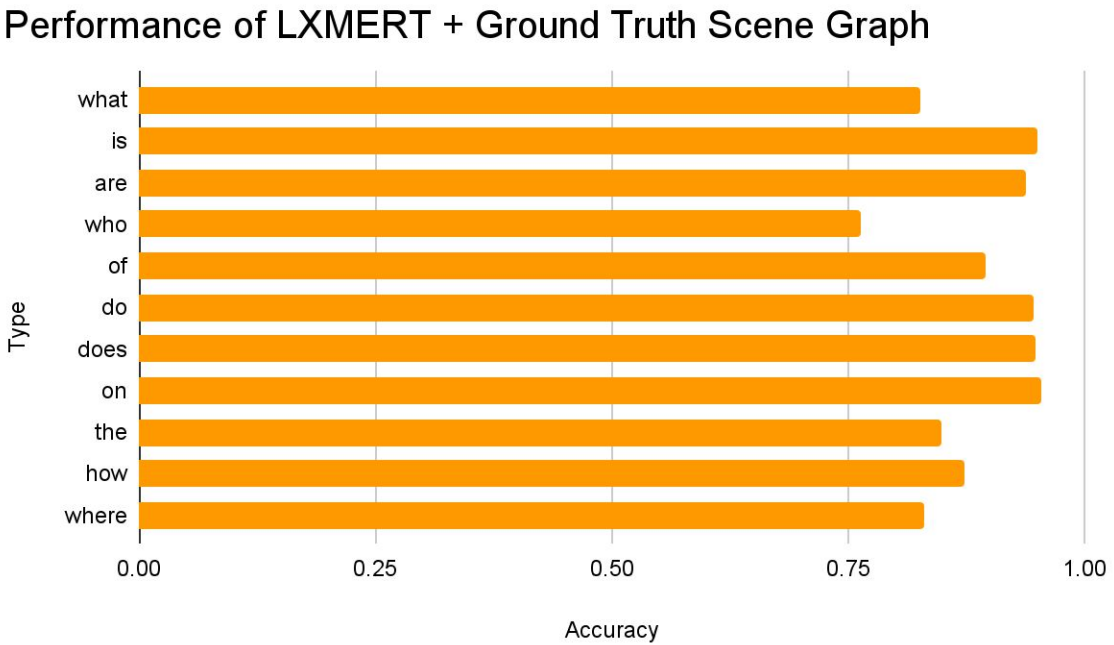
Carnegie Mellon University

Performance Comparison



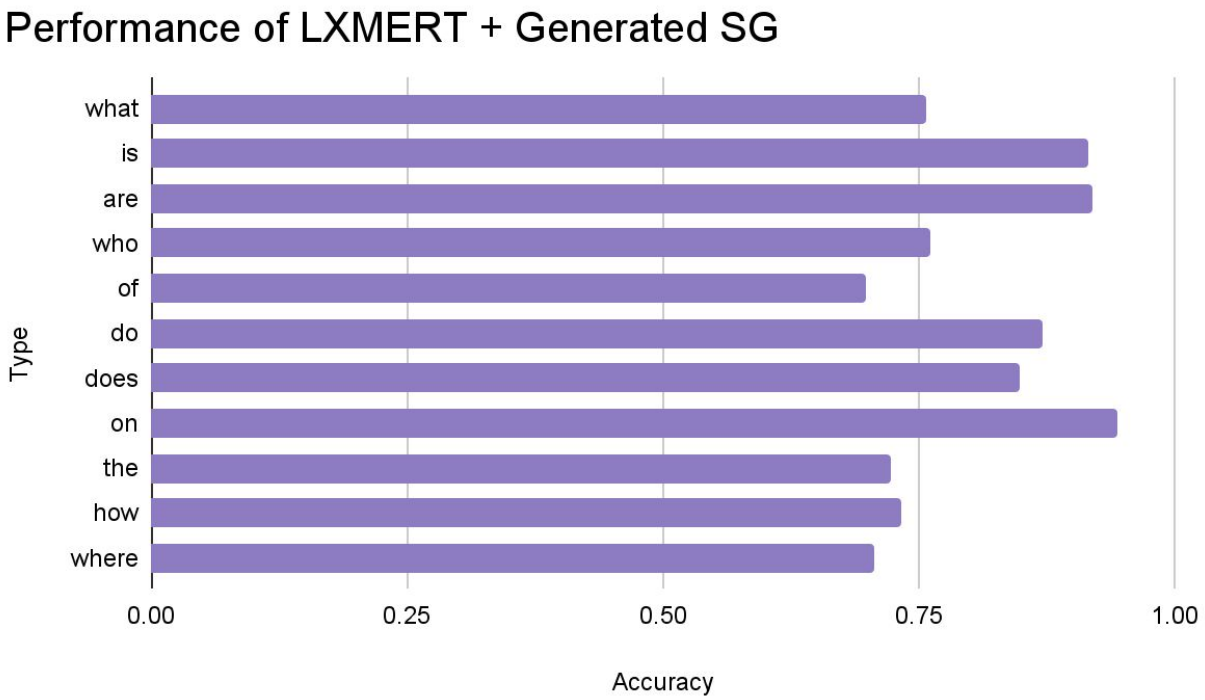
Carnegie Mellon University

Performance Comparison

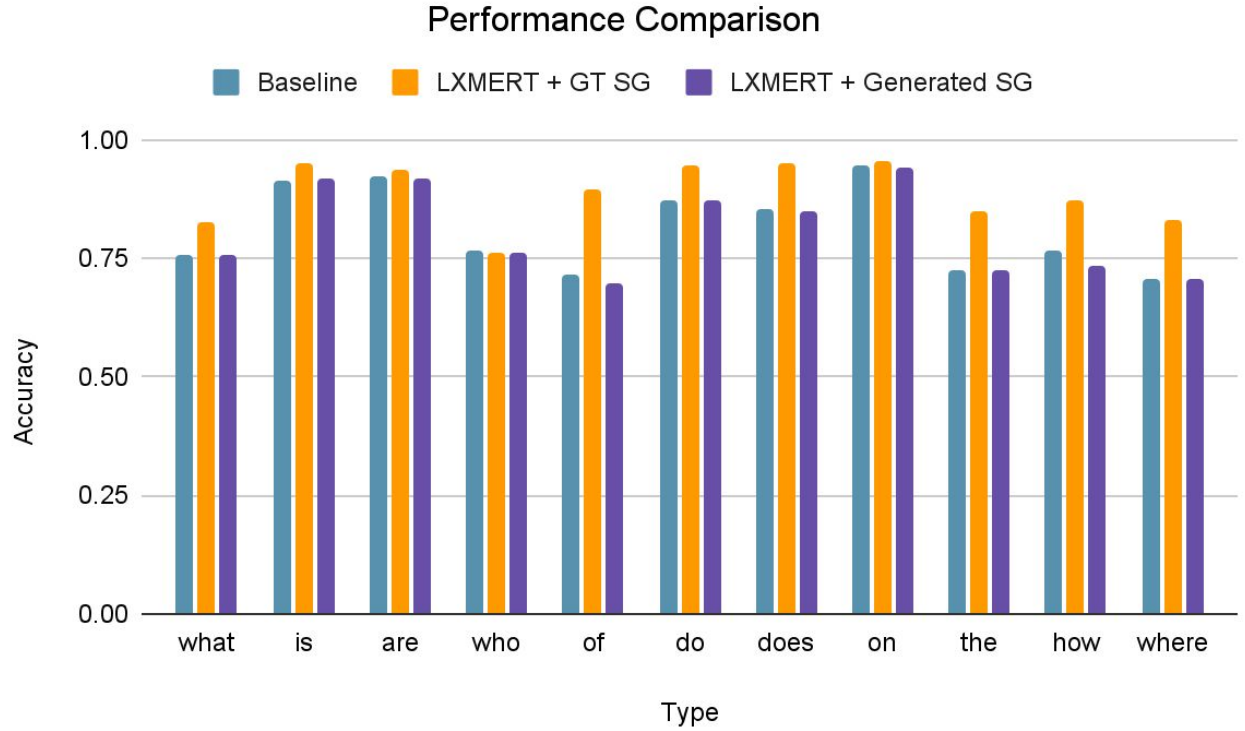


Carnegie Mellon University

Performance Comparison



Carnegie Mellon University



Carnegie Mellon University

Difference between Gold Standard and Generated Scene Graphs

white parrot, small white puffy clouds, blue sky, tall green palm tree, palm tree, bushes, hillside, palm yellow leaves, tan straw, rock, palm tree, rock wall, rock, rock behind parrot, rock on top of hillside, rock to the left of leaves, rock to the right of feathers, rock to the left of bushes, rock to the left of bushes, rock to the right of parrot, rock of palm tree, rock to the right of parrot ...

Gold Standard

tree to the left of feathers. feathers to the left of tree. feathers to the right of tree. tree to the right of feathers. tree to the left of bird. bird to the left of tree. head to the right of tree. tree to the right of head. bird to the left of wall. bird to the right of wall. tree to the right of sky. tree to the left of sky. sky to the left of tree. sky to the right of tree. beak of bird. wall behind bird.

Scene Graph Generator

Carnegie Mellon University

Case 1

What is the name of the cooking utensil that is hang from the hook?

LXMERT	LXMERT + GT SG	LXMERT + GEN SG
Pan	Spatula	Pan

Correct Answer : Pan

Carnegie Mellon University

Case 2

Is the horse next to the other horse both baby and brown?

LXMERT	LXMERT + GT SG	LXMERT + GEN SG
yes	no	no

Correct Answer : no

Carnegie Mellon University

Case 3

What is in front of the wall that is not short?

LXMERT	LXMERT + GT SG	LXMERT + GEN SG
Bookcase	Shelf	Shelf

Correct Answer : shelf

Carnegie Mellon University

Case 4

What color is the serving tray that looks rectangular?

LXMERT	LXMERT + GT SG	LXMERT + GEN SG
yellow	black	silver

Correct Answer : white

Carnegie Mellon University

Conclusion

- Can incorporation of scene graphs improve the performance of VQA systems?

  Typically, the inclusion of gold standard scene graphs results in an increase in accuracy.

- Can VQA performance benefit from utilizing question-specific scene graph information?

  Incorporating gold standard scene graphs leads to a noticeable 5.1 % enhancement in performance.

- Can the performance of VQA systems be enhanced by using a pipelined approach that involves scene graph generators?

  The scene graph generator's output is subpar, resulting in a decline in its performance, which necessitates further examination. (Similar results observed in *)

* Vinay Damodaran, Sharanya Chakravarthy, Akshay Kumar, Anjana Umapathy, Teruko Mitamura, Yuta Nakashima, Noa Garcia, and Chenhui Chu. 2021. Understanding the role of scene graphs in visual question answering.

Carnegie Mellon University

Ethical Considerations

- The Scene Graph Generator and LXMERT both utilize MSCOCO as a source of images. However, it has been reported that MSCOCO contains gender bias within its dataset, which can raise potential concerns when these systems are implemented in real-world settings.

- Many vision-related datasets exhibit inherent biases and therefore necessitate careful consideration in their utilization.

Computational Costs

- GCP Credits used ~ $350
- **Instance** : n1-highmem-16
- **CPU Platform** : Intel Skylake
- **RAM**: 112 GB
- **GPU**: Nvidia T4 / Nvidia Tesla P100
- **Time taken to finetune one LXMERT model** ~ 21 hours

Carnegie Mellon University

Limitations and Future Directions

- The current limitation lies in the Scene Graph Generation model's performance, as its output is not satisfactory, which adversely affects the overall performance of LXMERT.

- To address this, potential avenues for future research could include investigating query-specific scene graph generation techniques and exploring the feasibility of an end-to-end approach by integrating VQA models with scene graph generation.

Future Directions

Scene Graph Modification Based on Natural Language Commands

- The authors of this paper introduce the problem of Scene Graph modification where they propose a novel encoder-decoder architecture relying on graph-conditioned transformer and cross-attention to tackle the problem.
- Crowdsourced triplets of initial graph, prompt and final graph.

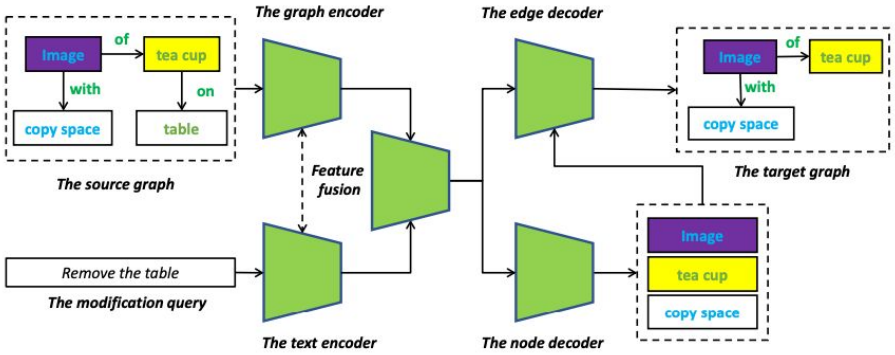


Figure 4: The information flow of our model. Green boxes denote the main computational units.

He, Xuanli, et al. "Scene graph modification based on natural language commands." arXiv preprint arXiv:2010.02591 (2020).

Carnegie Mellon University

Reflection

- Unfortunately, due to the time-consuming nature of the project and the significant effort required to establish a baselines and scene graph generators, I was unable to dedicate as much time as I would have liked to the original goal of implementing an end-to-end solution for scene graph generation and reasoning.

- I spent a considerable amount of time exploring various approaches to incorporate scene graphs as context. The primary objective was to identify the most effective configuration and determine if being context specific might help.

- To evaluate the effectiveness of the proposed solution, I conducted an in-depth error analysis using the LXMERT model both before and after integrating scene graphs. This allowed me to understand how the models were behaving.

Carnegie Mellon University

Thank You

Carnegie Mellon University