# SelfGraphVQA: A Self-Supervised Graph Neural Network for Scene-based Question Answering

**Bruno Souza**
Marius Aasan
Prof. Dr. Hélio Pedrini
Prof. Dr. Adín Ramírez Rivera

October 3rd, 2023

# Table of Contents

1. Introduction
   - Visual Question Answering
   - Motivation
2. Methodology
   - SelfGraphVQA
   - Similarity Loss
3. Results
   - GQA Results
   - Ablation
4. Conclusion
   - Contributions
   - Future Works

# Table of Contents


1. Introduction
   - Visual Question Answering
   - Motivation
2. Methodology
   - SelfGraphVQA
   - Similarity Loss
3. Results
   - GQA Results
   - Ablation
4. Conclusion
   - Contributions
   - Future Works

# Introduction

## Visual Question Answering (VQA)[1]


A testbed for the evaluation of reasoning and generalization capabilities.

[1]Anton et al. "VQA: Visual Question Answering." CVPR, 2015.

# Introduction

## Complex Reasoning task

Holistic comprehension of the scene.

Q: What is on the wall?

Ground-truth : Star
Prediction: A painting
(MCAN [Yu et al, 2019])

## Spectrum of Acceptable Answers

Broad spectrum of acceptable answers.

Q: What is happening?
Ground-truth: watching videos, showing phone
Prediction: phone
(ViLBERT [Agrawal et al, 2023])

VQA requires beyond the framework of classical statistical learning

## Great efforts towards Scene Graph for VQA.

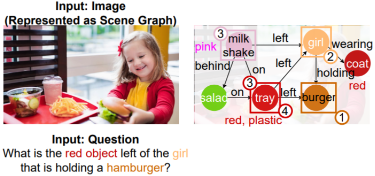



Illustration of the SG representation in the GQA Dataset[2]

Evaluation on GQA Dataset by data type and SGG usage.

| Method | Eval. Data | Acc (%) |
|---|---|---|
| Human | – | 89.30 |
| GraphVQA | Annotated/SGG | 94.78 |
| LRTA | Annotated/SGG | 93.10 |
| Lightweight | Annotated/SGG | 77.87 |
| CRF | Annotated | 72.10 |
| LXMERT | Extracted | 59.80 |
| GraphVQA | **Test Extracted/SGG** | 29.7 |

[2] Hudson and Manning. "GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering." CVPR, 2019.

# Motivation

HOWEVER, using annotated scene graphs is:

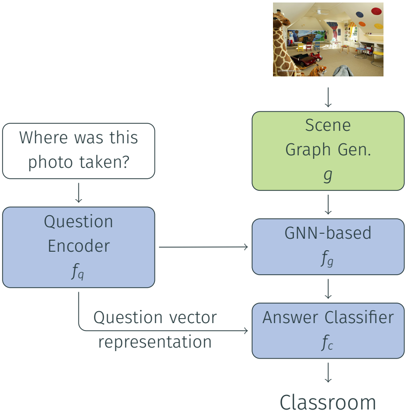
- Labor-intensive and expensive.
- Allows high spectrum of semantically correspondent scene graphs
- Potentially linguistic bias to the question.

# Motivation

More practical approach that uses a Scene graph generator model[3]

Leverage the self-supervised learning to enhance the visual information.

Baseline architecture[4]



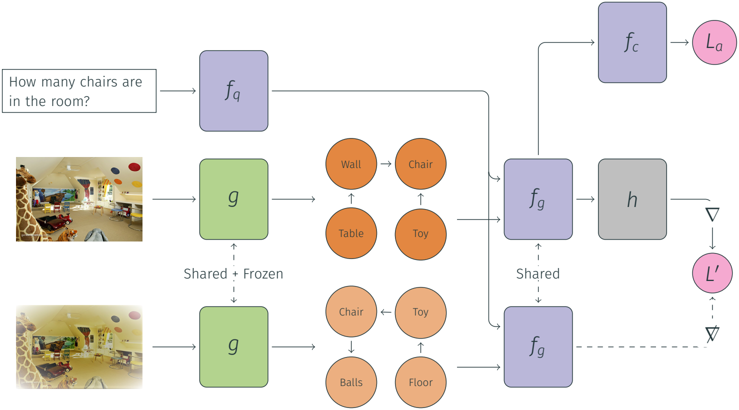

[3] Knyazev et al. "Graph Density-Aware Losses for Novel Compositions in Scene Graph Generation." BMVC, 2020.

[4] Liu et al. "GraphVQA: Language-Guided Graph Neural Networks for Scene Graph Question Answering.", 2021.

# Table of Contents

1. Introduction
    - Visual Question Answering
    - Motivation
2. Methodology
    - SelfGraphVQA
    - Similarity Loss
3. Results
    - GQA Results
    - Ablation
4. Conclusion
    - Contributions
    - Future Works

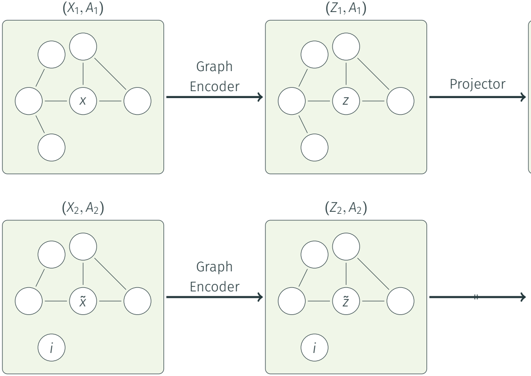
# SelfGraphVQA




We handle three distinct maximization strategies:

- Local Similarity: Node Wise
- Global Similarity: Global Wise
- SelfSim: Regularization for Permutation Equivariance

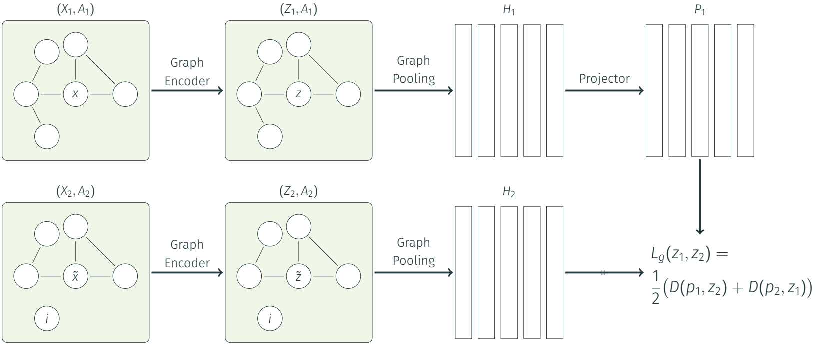
Object-wise: Similarity over object pairs from the two views.




$$L_\ell^*(p_1, z_2) = \frac{1}{O} \sum_{i}^{O} \arg\min_{z_{2,j}} D(p_{1,i}, z_{2,j})$$

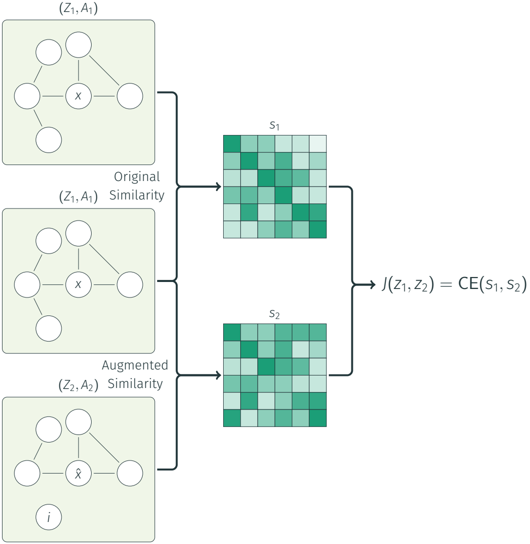
# Global similarity strategies

Global-wise: Similarity maximization for the scene representation.



# SelfSim

Aligning Comparable Nodes and Promoting Regularization: address permutation invariance in graph representations.



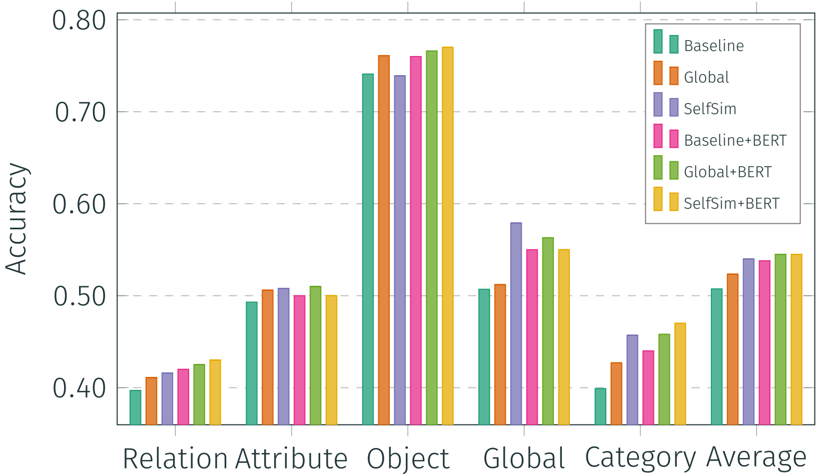
# Table of Contents

1. Introduction
   - Visual Question Answering
   - Motivation
2. Methodology
   - SelfGraphVQA
   - Similarity Loss
3. Results
   - GQA Results
   - Ablation
4. Conclusion
   - Contributions
   - Future Works

# Results (%) on GQA by standard metrics

| Method | Binary (↑) | Open (↑) | Consist. (↑) | Validity (↑) | Plausab. (↑) | Distr. (↓) | Acc (↑) |
|---|---|---|---|---|---|---|---|
| Baseline | 65.8 | 29.7 | 58.2 | 94.9 | 90.5 | 11.7 | 50.1 |
| Baseline+BERT | 68.0 | 32.2 | 62.6 | 95.0 | 90.9 | 7.7 | 53.8 |
| Local | 66.8 | 30.2 | 59.4 | 94.9 | 90.6 | 8.8 | 51.5 |
| Global | 67.7 | 30.8 | 62.5 | 94.9 | 90.6 | 6.7 | 52.3 |
| SelfSim | **68.4** | 31.3 | **65.9** | 94.9 | 90.7 | **2.1** | 54.0 |
| Global+BERT | 68.0 | **33.0** | 63.9 | 95.0 | **91.2** | 8.9 | **54.5** |
| SelfSim+BERT | 68.2 | 32.8 | 64.3 | **95.0** | 91.0 | 8.0 | **54.5** |

# Accuracy on different question types



# Does the SG really matter?

Experimental Design: unfavorable perturbation study by augmenting images based on question types.

Greater drop, better outcome.

| Question Type | Augmentation | Baseline | Global | Local | SelfSim |
|---|---|---|---|---|---|
| Relation | Flip | −1.6 | −3.4 | −3.2 | −3.9 |
| Attribute | Strong Color Jitter | +1.14 | −3.7 | −0.8 | −1.2 |
| Global | Gaussian Noise + Crop | −5.6 | −7.7 | −5.5 | −8.1 |

# Are Performance Gains Mainly Due to Augmentations?

Experimental Design: contrasted our approach with the baseline exclusively relying on data augmentation for training.

Evidence that data augmentation detrimentally affects the overall performance.

| Method | Binary | Open | Validity | Plausibility | Acc |
|---|---|---|---|---|---|
| Baseline Aug | 65.1 | 28.7 | 94.6 | 90.1 | 50.1 |
| SelfSim | 68.4 | 31.3 | 94.9 | 90.7 | 54.0 |

# Are Our Models Less Biased and More Robust?

Hypothesis: State-of-the-art models might exploit question and answer distribution bias, leading to ”clever guesses”[5]

Experimental Design: slightly perturbing node features with random noise in both the scene graph and questions.

| Setup | | Methods | | |
|---|---|---|---|---|
| Scene Graph + Question | Baseline | Local | Global | SelfSim |
| SG + Noise | 16.2 | 16.6 | 28.6 | 26.6 |
| Noise + Question | 39.9 | 38.3 | 37.4 | 39.8 |
| Noise + Noise | 12.7 | 14.6 | 18.9 | 21.0 |
| Scene Graph + Question | BERT Baseline | BERTGlobal+link | BERTSelfSim+link | |
| SG + Noise | 21.0 | 23.2 | 24.5 | |
| Noise + Question | 42.4 | 41.8 | 42.8 | |
| Noise + Noise | 19.8 | 21.7 | 21.3 | |

[5]Agrawal et al., “Don’t just assume; look and answer: Overcoming priors for visual question answering.” CVPR, 2018. Yuan et al., “Language bias in visual question answering: A survey and taxonomy.” arXiv:2111.08531.

# Table of Contents

1. Introduction
   - Visual Question Answering
   - Motivation
2. Methodology
   - SelfGraphVQA
   - Similarity Loss
3. Results
   - GQA Results
   - Ablation
4. Conclusion
   - Contributions
   - Future Works

# Conclusion

- Impact of Scene Graph Quality
- Practical SG model for VQA task
- Effective Similarity Maximization
- Consistent Visual Enhancement

# Future Works

- Extend for more datasets such as VQAv2 and VizWiz
- Investigation of Alternative Scene Graph Generator Models
- Enhancement of Encoder Architecture
- Advancement of Self-Supervised Energy-based Approaches

Thank you!