An Empirical Study on Leveraging Scene Graphs for Visual Question Answering

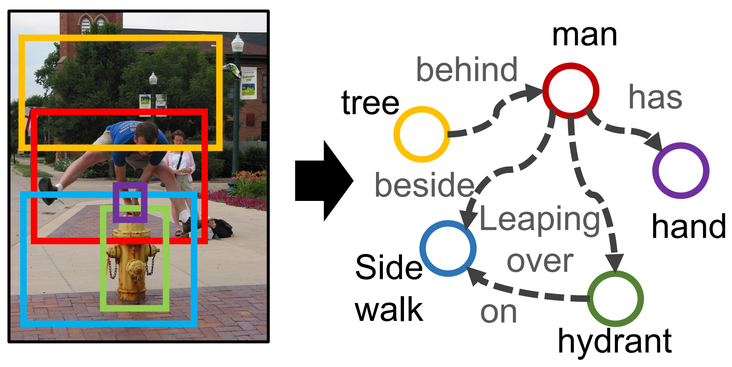
Cheng Zhang[1], Wei-Lun Chao[12], and Dong Xuan[1]
[1]The Ohio State University, [2]Cornell University

Highlights

- Investigate leveraging ***Scene Graphs (SGs)*** for visual question answering (Visual QA)
- Adapt ***Graph Networks (GNs)*** [1] to perform structured computations on SGs
- Conduct comprehensive empirical studies of GNs on the Visual Genome dataset [2,3], demonstrating that SGs can ***benefit*** Visual QA on ***various question types***
- Analyze GNs-based models to reveal the reasoning process on SGs for ***explainable*** Visual QA

Introduction

- Existing work on Visual QA with object relationships or SGs:
  - Mainly experiment on synthetic dataset → *No natural images*
  - Integrate multiple techniques → *Are SGs effective?*
- Our observation: Images can be abstractly represented by SGs
  - Nodes: object names and attributes
  - Edges: object relationships



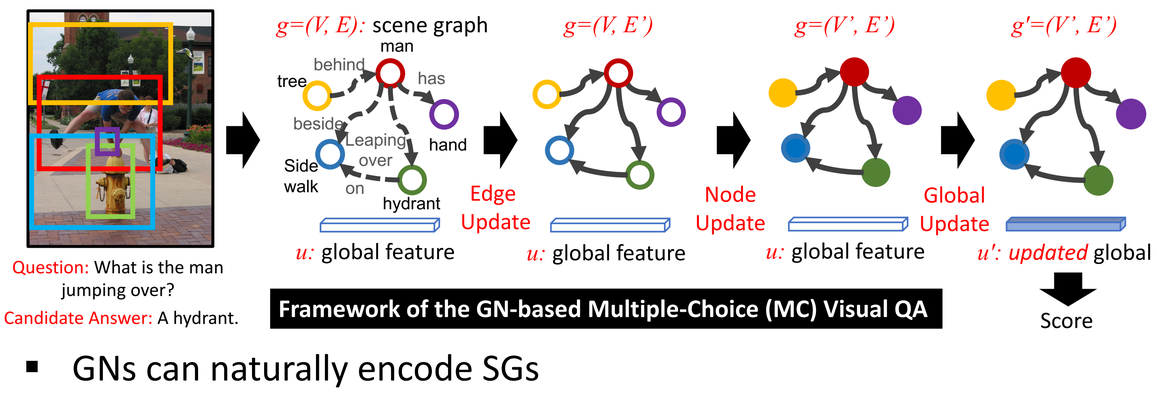
Question: What is the man jumping over?

Answer: A hydrant.

- Can we improve Visual QA via SGs?
  - How to leverage SGs for Visual QA? → *Graph Networks [1]*
  - Do we have high-quality SGs for Visual QA? → *VG [3] v.s. NM [4]*
  - How would SGs improve Visual QA? → *Empirical study on VG*

A comprehensive study on leveraging Scene Graphs for Visual QA without applying attention and multimodal fusion mechanisms

GN-based Visual QA with Scene Graphs



Question: What is the man jumping over?
Candidate Answer: A hydrant.

Framework of the GN-based Multiple-Choice (MC) Visual QA

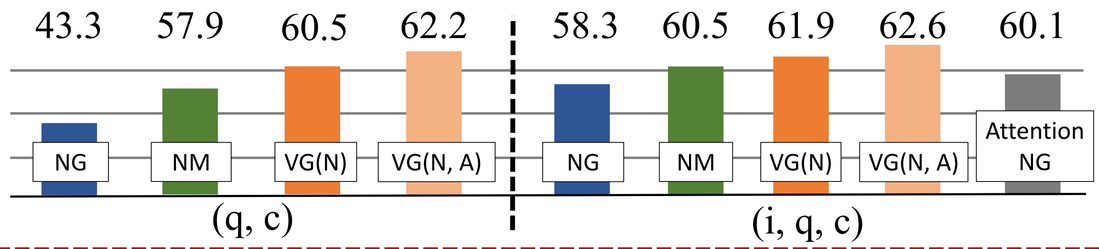
- GNs can naturally encode SGs
  - Nodes encode names (*e.g.,* man) and attributes (*e.g.,* colors)
  - Edges encode relationships (*e.g.*, behind or leaping over)
  - Global u encodes the image i, question q, and candidate answers c
- GNs perform *graph-to-graph* mapping with 3 updating procedures
  - Edge update: $E' = f^E(E, V, u)$
  - Node update: $V' = f^V(E', V, u)$
  - Global update: $u' = f^u(E', V', u)$
- The resulting (u', V', E') can be used for Visual QA, even for the open-ended setting, or serve as the input to a subsequent GN block

Selected References:
[1] Relational inductive biases, deep learning, and graph network. arXiv 2018.
[2] Being negative but constructively: Lessons learnt from creating better visual question answering datasets. In NAACL 2018.
[3] Visual Genome: Connecting language and vision using crowdsourced dense Image annotations. IJCV 2017.
[4] Neural motifs: Scene graph parsing with global context. In CVPR 2018.
[5] Revisiting visual question answering baselines. In ECCV 2016.
Acknowledgements. The computational resources are supported by the Ohio Supercomputer Center (PAS1510)

Experiments and Analysis

- **Dataset:** Enhanced Visual Genome (*qaVG*) [2,3]
- **Evaluation:** Selecting accuracy from 7 candidate answers
- **Scene Graphs:**
  - human-annotated (Visual Genome, **VG**) [3]
  - machine-generated (Neural Motifs, **NM**) [4]
- **Features [5]:**
  - Image: 2,048-dim ResNet-152 (from ImageNet)
  - Question and answers: 300-dim averaged Word2Vec (from Google News)

- **Overall accuracy (%): NG:** no graphs. **N:** node names. **A:** attributes.

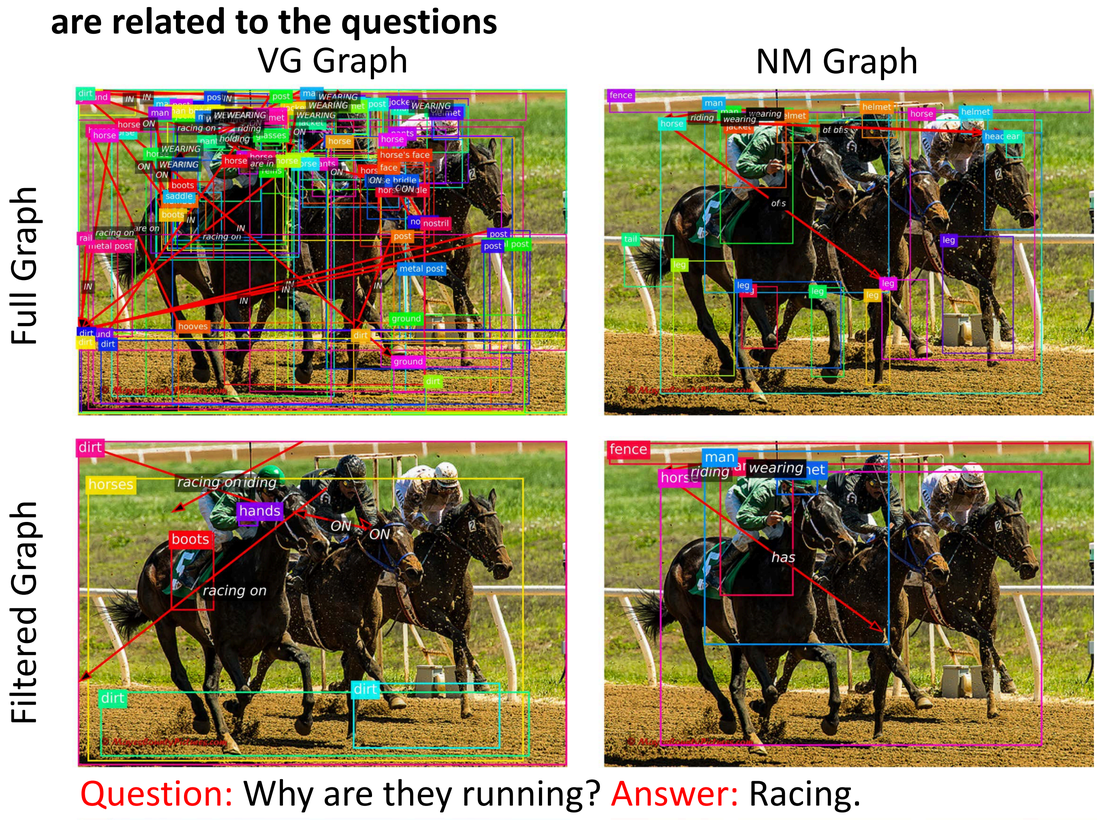


- **Scene graph can benefit Visual QA on various question types**
  - w/o image features, SGs improve all question types but ***when***
  - Even w/ image features, SGs largely improve ***what***, ***who***, and ***number***
  - Node attributes benefit ***color***
  - VG SGs are better than NM SGs except for ***number***

Question type	What	Color	Where	Number	How	Who	When	Why	Overall
Percentage	(46%)	(14%)	(17%)	(8%)	(3%)	(5%)	(4%)	(3%)	(100%)
NG + (q, c)	40.3	50.6	36.2	52.0	41.1	37.6	83.2	39.5	43.3
NG + (i, q, c)	57.8	59.5	59.1	55.5	45.4	56.6	84.6	48.3	58.3
NM(N) + (i, q, c)	59.4	58.2	60.3	63.4	54.3	66.6	85.3	48.1	60.5
VG(N) + (q, c)	61.6	54.0	62.4	58.6	45.9	63.9	83.2	50.3	60.5
VG(N) + (i, q, c)	61.1	61.4	62.3	59.4	54.3	67.5	85.3	48.9	61.9
VG(N, A) + (i, q, c)	61.4	63.8	62.6	61.5	54.8	67.5	84.8	49.6	62.6

- **GNs without edges:** reduces 1.2%↓, justifying the need to take the relationships between the nodes into account for Visual QA

- **GN-based models can implicitly attend to nodes and edges that are related to the questions**



- **Future work:** Visual features on nodes, and multimodal fusion & attention