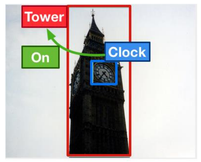
Relation-aware Graph Attention Network for Visual Question Answering

Linjie Li, Zhe Gan, Yu Cheng, Jingjing Liu

Microsoft Dynamics 365 AI Research

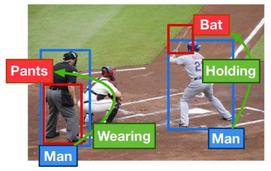
Motivation

- Most previous work on VQA focused on multimodal fusion to learn a joint representation of a sparse set of image regions and the question
- Interactions between different objects in the image (e.g., actions and relative geometrical positions) are not considered
- Visual object relations can aid the VQA task by interpreting the dynamics and interactions between different objects in an image
- Different types of relations can be formed between objects in each image, the importance of which should differ given different questions



Q: Where is the clock?
A: On the top of the tower.

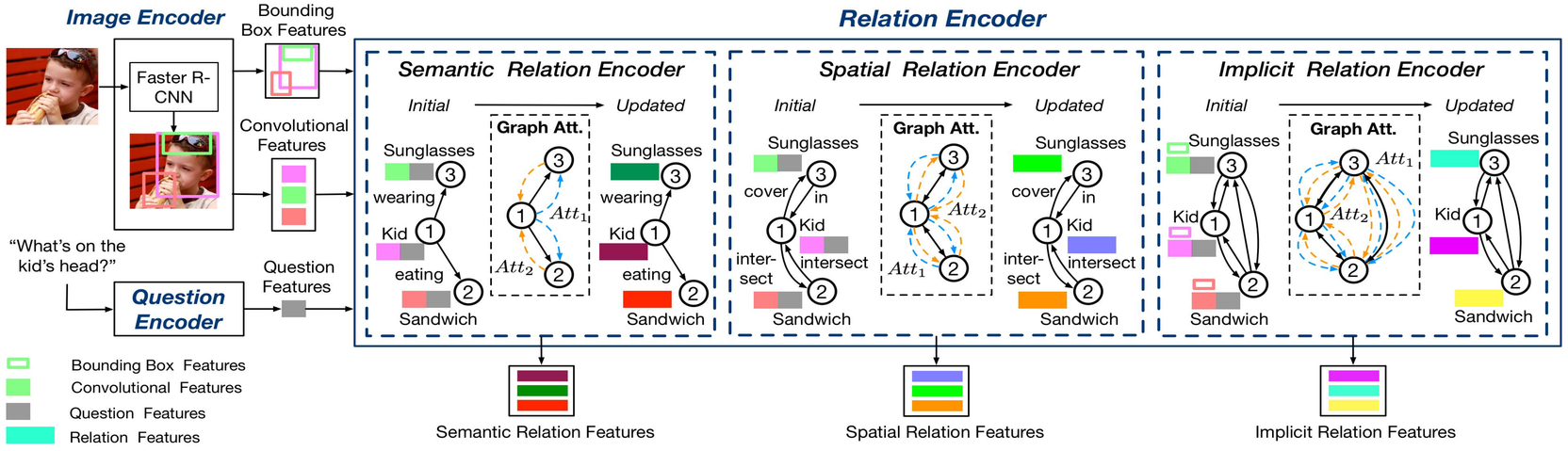
(a) Spatial Relation



Q: What is the man in blue hat holding?
A: Bat.

(b) Semantic Relation

Relation-Aware Graph Attention Network (ReGAT)



- Each question is encoded as a feature vector through a GRU
- Each image is encoded as a graph (objects as nodes and relations as edges)
- Relation-aware representations are learned via Question-adaptive Graph Attention: absorb semantic information from questions to capture relations that are most question-relevant
- Explicit relations: Spatial and Semantic relations are obtained from a rule-based classifier and a pretrained classifier on Visual Genome, respectively
- Implicit relations are learned dynamically during the training process
- The relation-aware image representations are then fused with question representation through multimodal fusion to predict an answer

Experimental Results

Quantitative Results

- Achieved state-of-the-art performance on both VQA 2.0 and VQA-CP v2, generalizable to different VQA tasks

Model	Test-dev				Test-std
	Overall	Y/N	Num	Other	
BUTD	65.32	81.82	44.21	56.05	65.67
MuRel	68.03	84.77	49.86	57.85	68.41
MFH	68.76	84.27	50.66	60.50	-
Pythia	70.01	-	-	-	70.24
BAN	70.04	85.42	54.04	**60.52**	70.35
ReGAT+BAN	**70.27**	**86.08**	**54.42**	60.33	**70.58**

Table 1. Results on VQA Benchmark

SOTA	Baseline	Semantic	Spatial	Implicit	All
39.54	39.24	39.54	40.30	39.58	**40.42**

Table 2. Results on VQA-CP Benchmark

- Consistent performance gain when combining ReGAT with different fusion methods

Model	Baseline	Semantic	Spatial	Implicit	All
BUTD	63.38	64.11	64.02	64.10	**65.30**
MUTAN	61.36	62.60	62.01	62.45	**64.37**
BAN	65.51	65.97	66.02	65.93	**67.18**

Table 3. Ablation Study on Different Relation Types Across Different Fusion Methods (VQA val)

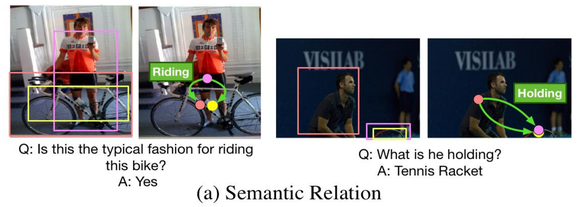
- Both graph attention (Att.) and question-adaptive (Q-ada.) mechanisms contribute to performance improvement

Att.	Q-ada.	Semantic	Spatial	Implicit
No	No	63.20	63.04	n/a
Yes	No	63.90	63.85	63.36
No	Yes	63.31	63.13	n/a
Yes	Yes	**64.11**	**64.02**	**64.10**

Table 4. Ablation Study on Question-adaptive Graph Attention (VQA val)

Visualization

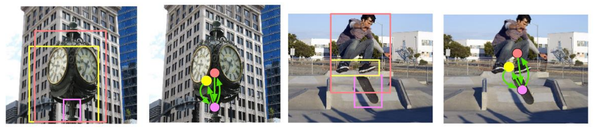
- Each relation type contributes to a better alignment between regions and questions



Q: Is this the typical fashion for riding this bike?
A: Yes

Q: What is he holding?
A: Tennis Racket

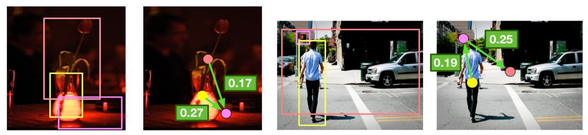
(a) Semantic Relation



Q: What's the clock attached to?
A: Pole

Q: Are his feet touching the skateboard?
A: No

(b) Spatial Relation

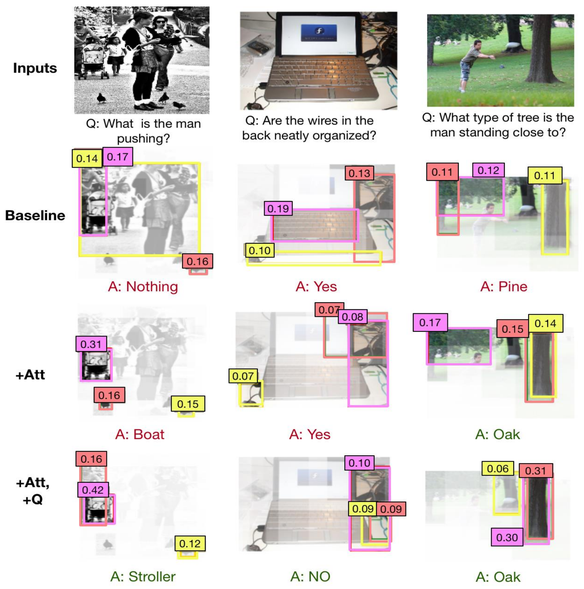


Q: Where is the vase?
A: On the table

Q: Should the people be walking according to the light?
A:No

(c) Implicit Relation

- Question-adaptive Graph Attention produces sharper attention maps and focuses on more relevant regions



Q: What is the man pushing?
A: Nothing

Q: Are the wires in the back neatly organized?
A: Yes

Q: What type of tree is the man standing close to?
A: Pine

A: Boat

A: Yes

A: Oak

A: Stroller

A: NO

A: Oak

(a) Semantic Relation (b) Spatial Relation (c) Implicit Relation