From Strings to Things: Knowledge-enabled VQA Model that can Read and Reason

Ajeet Kumar Singh[1] , Anand Mishra[2] ,
Shashank Shekhar[3], Anirban Chakraborty[3]

[1]TCS Research , [2]IIT Jodhpur , [3]IISc Bangalore

Problem

Problem

Traditional VQA

[Antol et al., ICCV'15, Zhang et al., ICLR'18]

Q: How many cars are there in this image?

A: 2

Problem

Traditional VQA
[Antol et al., ICCV'15, Zhang et al., ICLR'18]
Q: How many cars are there in this image?
A: 2

ST-VQA, Text-VQA
[Biten et al., ICCV'19, Singh et al., CVPR'19]
Q: Which restaurant brand is written on the red wall?
A: KFC

Problem

Traditional VQA
[Antol et al., ICCV'15, Zhang et al., ICLR'18]
Q: How many cars are there in this image?
A: 2

ST-VQA, Text-VQA
[Biten et al., ICCV'19, Singh et al., CVPR'19]
Q: Which restaurant brand is written on the red wall?
A: KFC

Text + Knowledge-enabled VQA [This work]
Q: Can I get chicken dish here?
A: Yes

Answering requires external knowledge

Problem

Traditional VQA
[Antol et al., ICCV'15, Zhang et al., ICLR'18]
Q: How many cars are there in this image?
A: 2

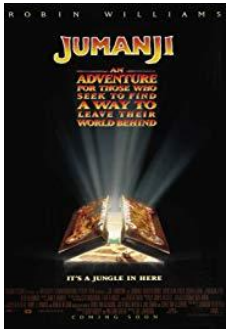
ST-VQA, Text-VQA
[Biten et al., ICCV'19, Singh et al., CVPR'19]
Q: Which restaurant brand is written on the red wall?
A: KFC

Text + Knowledge-enabled VQA [This work]
Q: Can I get chicken dish here?
A: Yes

New problem, No dataset exists!

text-KVQA: A novel dataset

Q: Is this a chinese restaurant?
A: **No**



Q: When was this movie released?
A: **1995**

Q: Can I get medicine here?
A: **Yes**

- 257K Images, 1 Million QA Pairs
- Associated knowledge base
- **First dataset:** text recognition + Knowledge graph + VQA

Proposed Solution

Question:
Is this an American brand?

Proposed Solution

Question:
Is this an American brand?

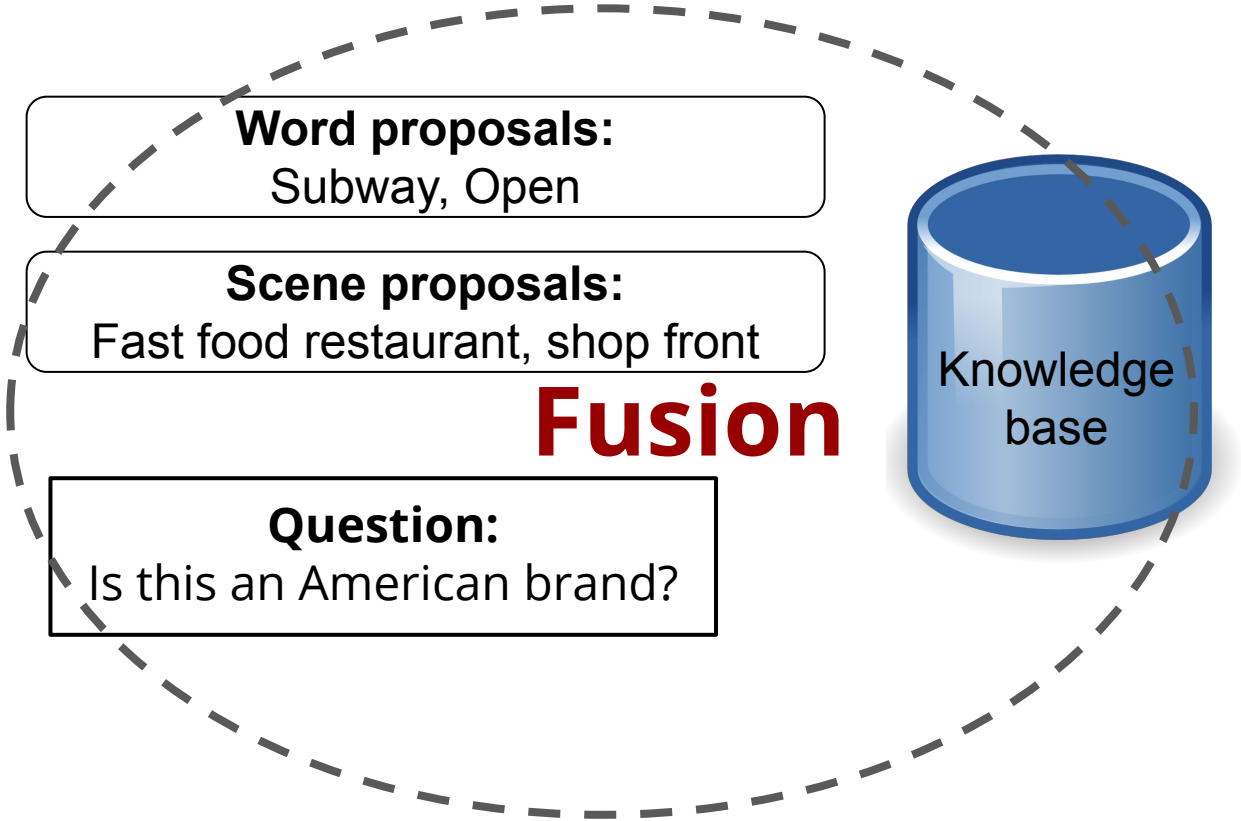
Proposal Module

Word proposals:
Subway, Open

Scene proposals:
Fast food restaurant, shop front

Word proposals
[Gupta et al., CVPR'16]
Scene proposals
[Zhou et al., TPAMI'17]

Proposed Solution



Fusion Module

Relevance score of each knowledge fact:

$$S(h_i, r_i, t_i)$$

$$= \max_{j,k} \alpha_w s_{w_j} < w_j, (h_i, r_i, t_i) >$$

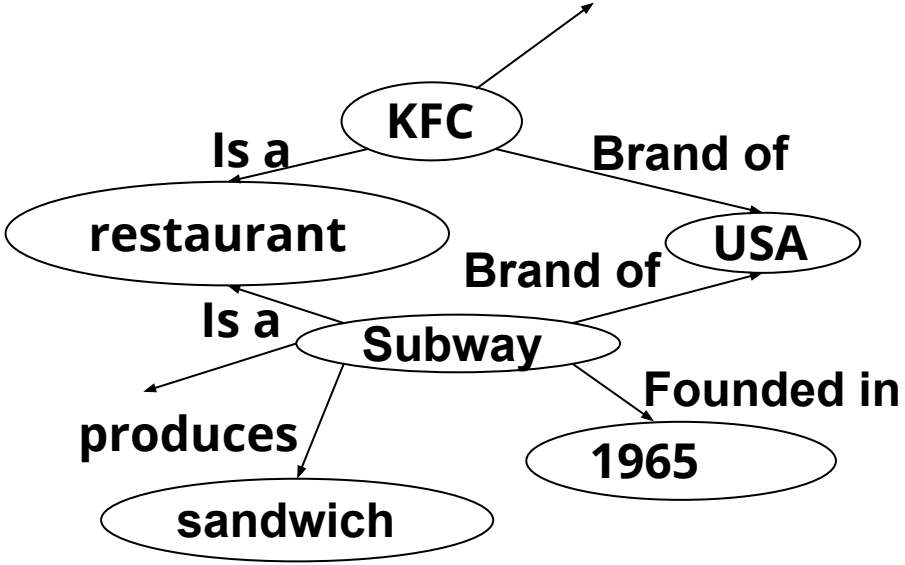
$$+ \alpha_v s_{v_k} < v_k, (h_i, r_i, t_i) >$$

$$+ \alpha_q < Q, (h_i, r_i, t_i) >.$$

Proposed Solution

Word proposals:
Subway, Open

Scene proposals:
Fast food restaurant, shop front



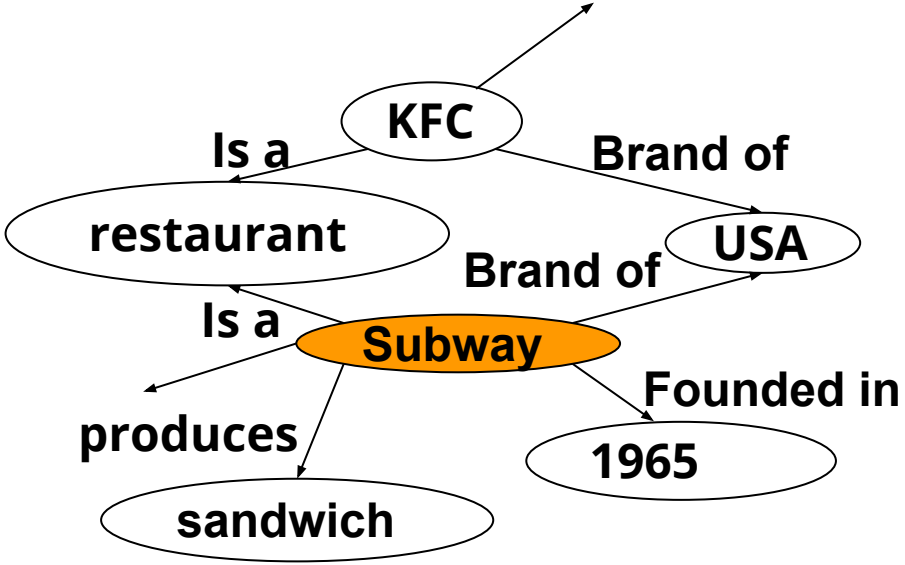
Question:
Is this an American brand?

Proposed Solution

Word proposals:
Subway, Open

Scene proposals:
Fast food restaurant, shop front

Question:
Is this an American brand?

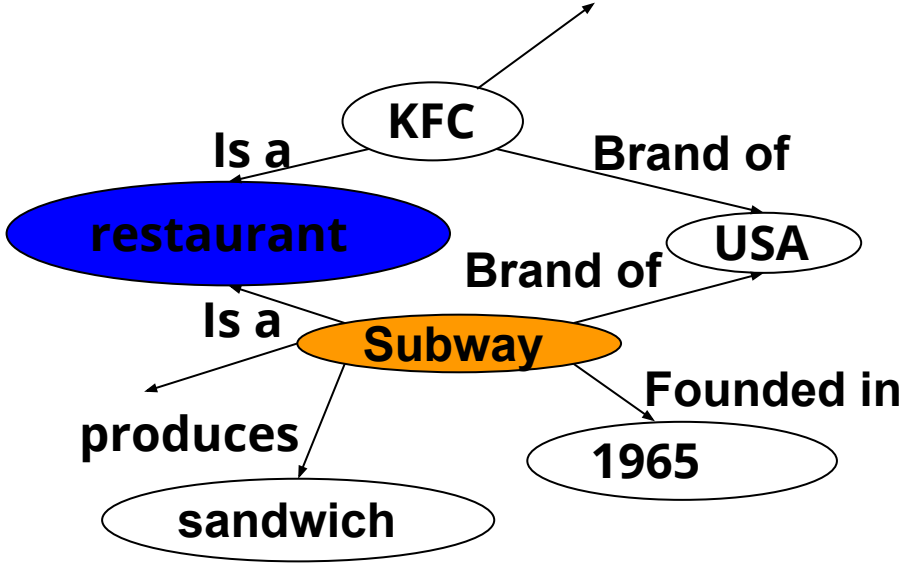


Proposed Solution

Word proposals:
Subway, Open

Scene proposals:
Fast food restaurant, shop front

Question:
Is this an American brand?

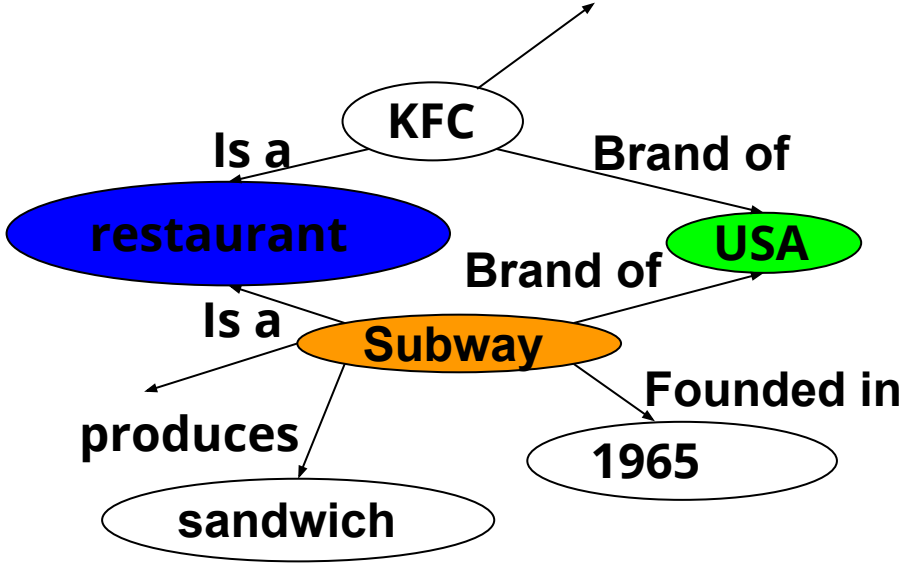


Proposed Solution

Word proposals:
Subway, Open

Scene proposals:
Fast food restaurant, shop front

Question:
Is this an American brand?

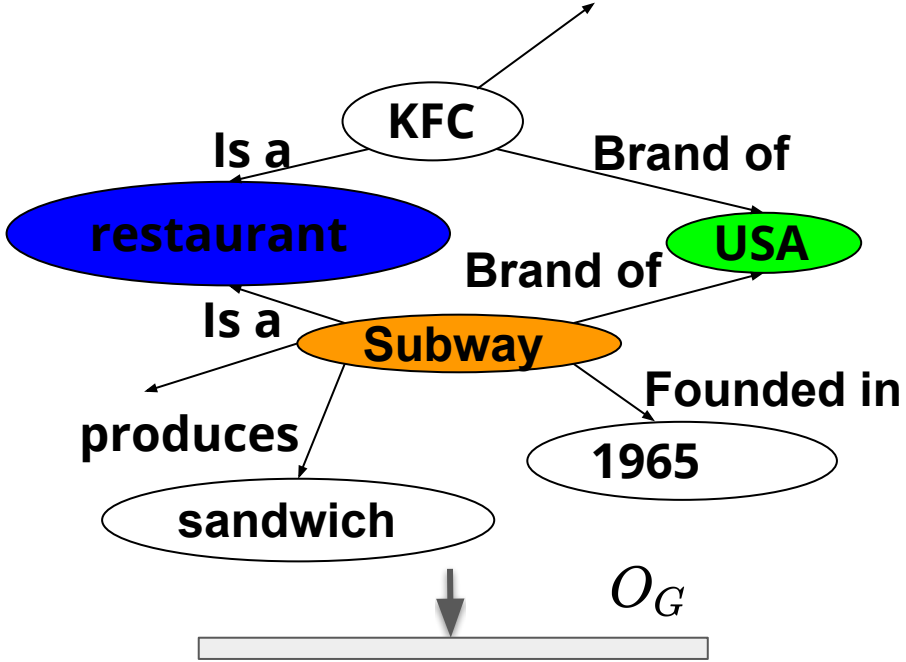


Proposed Solution

Word proposals:
Subway, Open

Scene proposals:
Fast food restaurant, shop front

Question:
Is this an American brand?



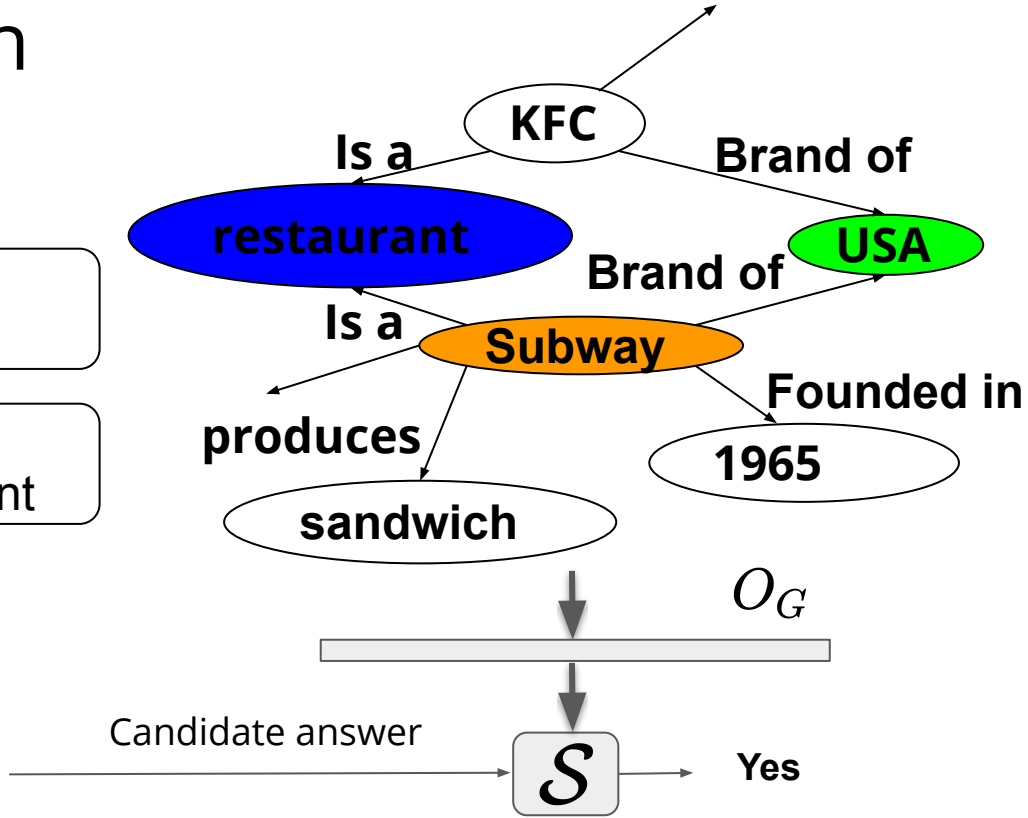
Graph representation: Gated Graph Neural Network (GGNN)
[Li et al., ICLR'15]

Proposed Solution

Word proposals:
Subway, Open

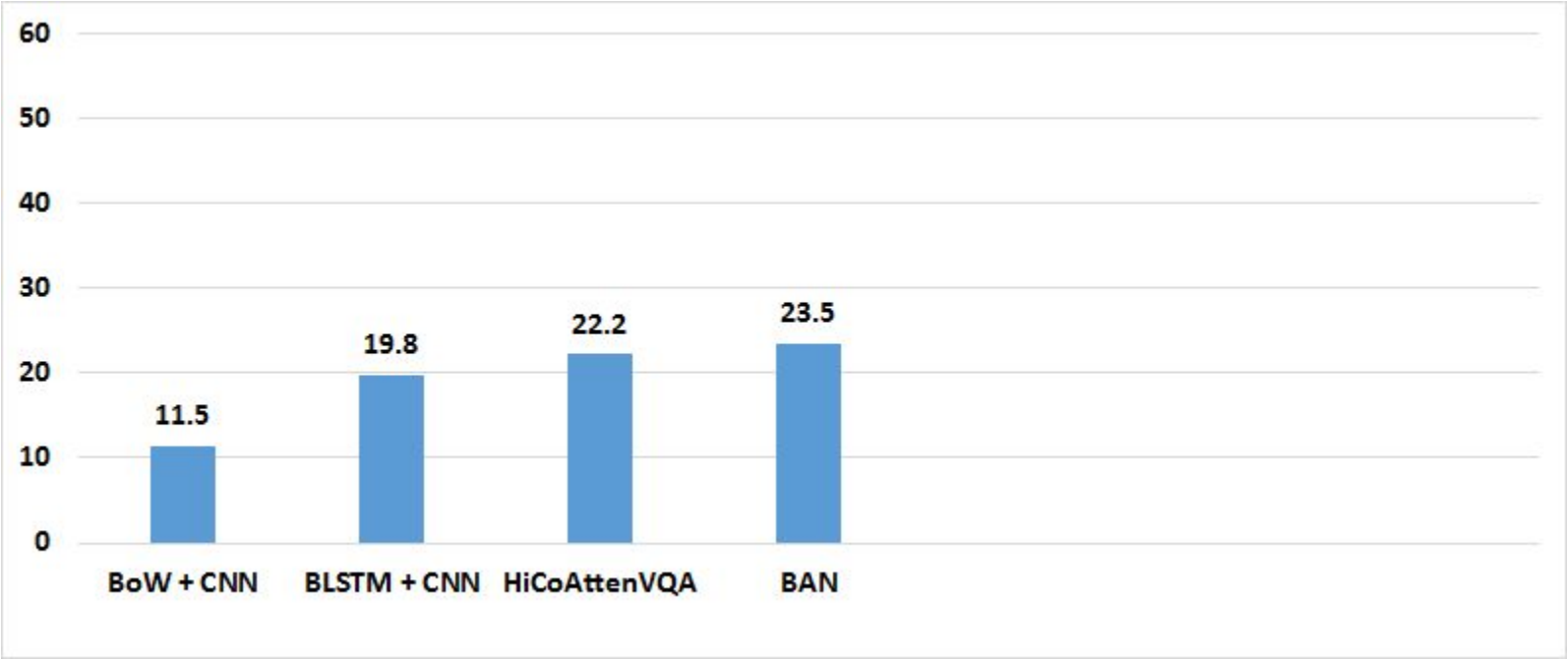
Scene proposals:
Fast food restaurant, shop front

Question:
Is this an American brand?



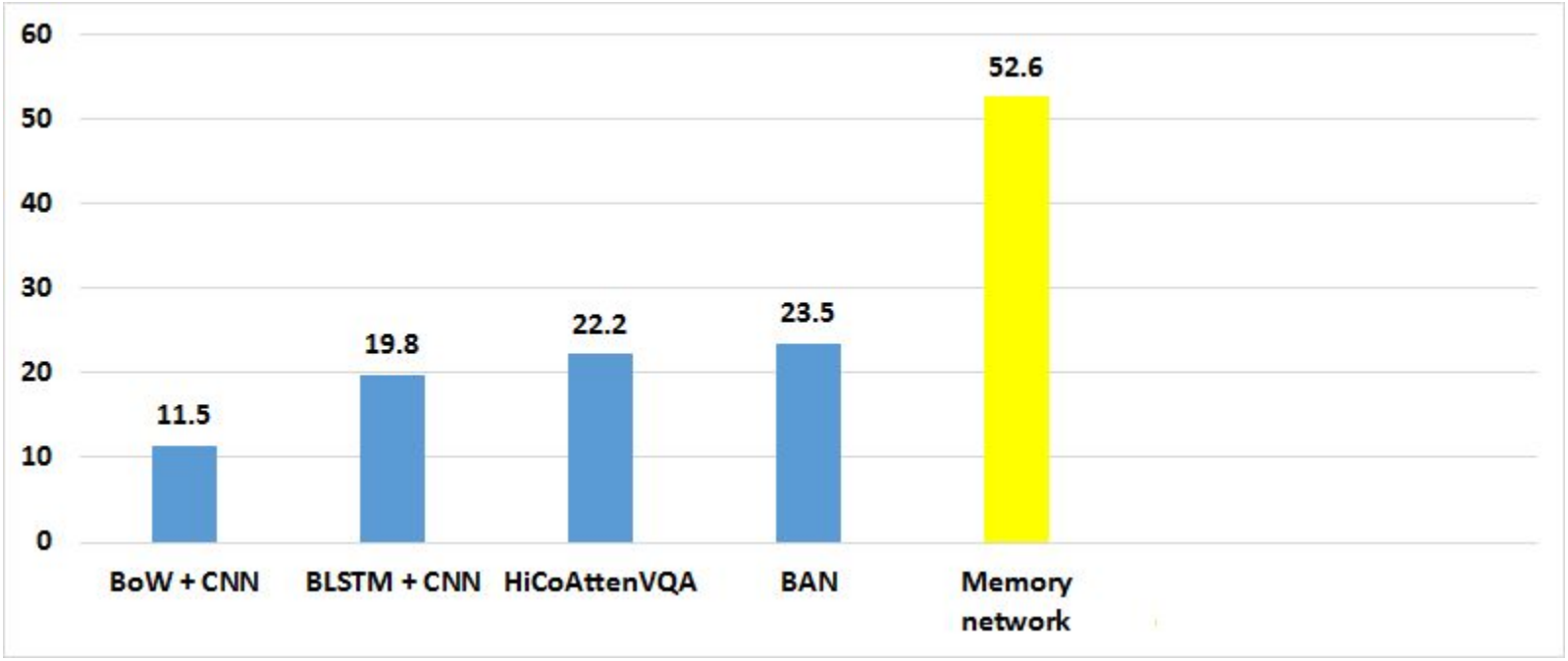
Graph representation: Gated Graph Neural Network (GGNN)
[Li et al., ICLR'15]

text-KVQA accuracy (%)



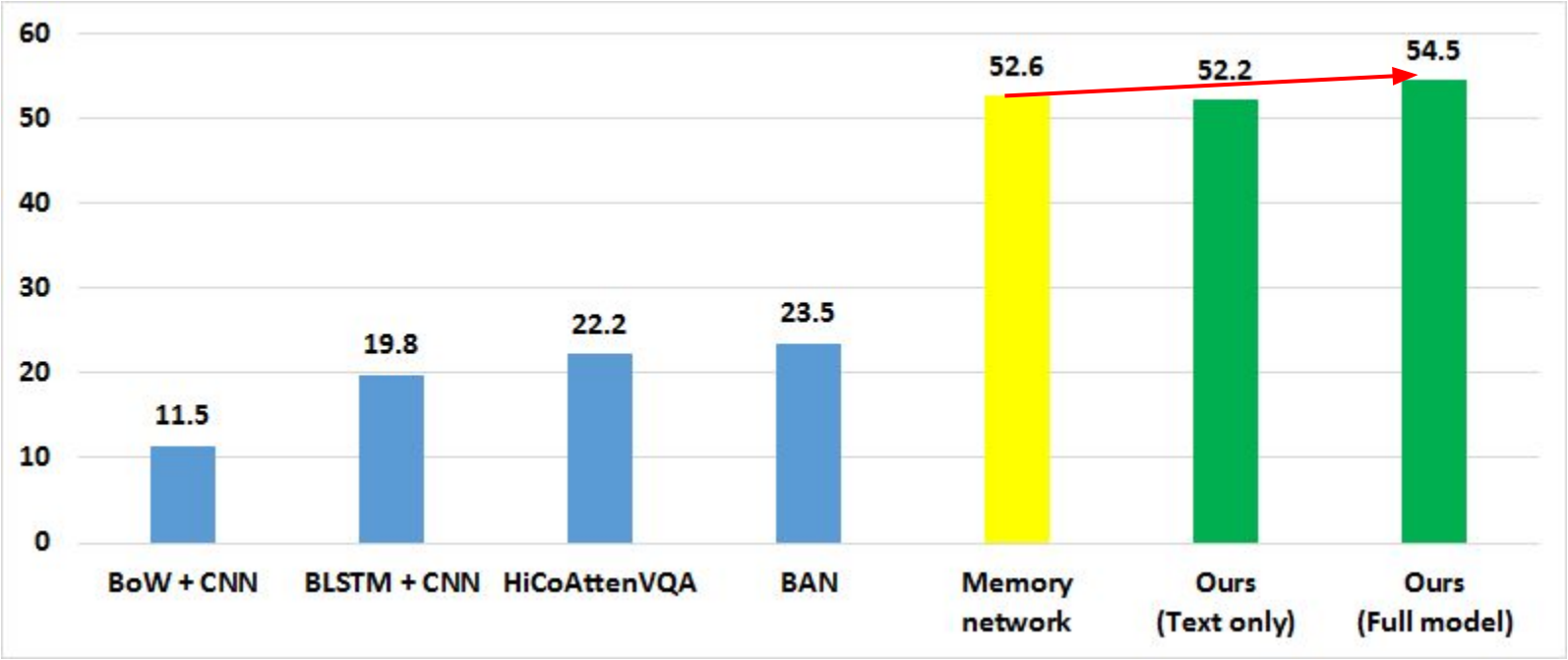
Traditional VQA methods are not successful

text-KVQA accuracy (%)



A popular QA over KB method improves the performance

text-KVQA accuracy (%)



Our GGNN-based full model (text + vision) further improves the performance

Summary

1. **text-KVQA**: first dataset for **knowledge-enabled** VQA by reading text in image
2. **Novel GGNN formulation**

Dataset available at https://textkvqa.github.io/

Please visit us at poster number: 18

Acknowledgements: Anirban Chakraborty is supported by Tata Trusts Travel Grant.