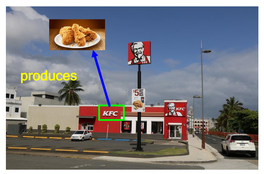
From Strings To Things: Knowledge-Enabled VQA Model That Can Read And Reason

Ajeet Kumar Singh[1], Anand Mishra[2], Shashank Shekhar[3], and Anirban Chakraborty[3]
[1]TCS Research, [2]Indian Institute Of Technology Jodhpur, [3]Indian Institute Of Science, Bangalore

ICCV 2019 Seoul, Korea

https://textkvqa.github.io

Problem

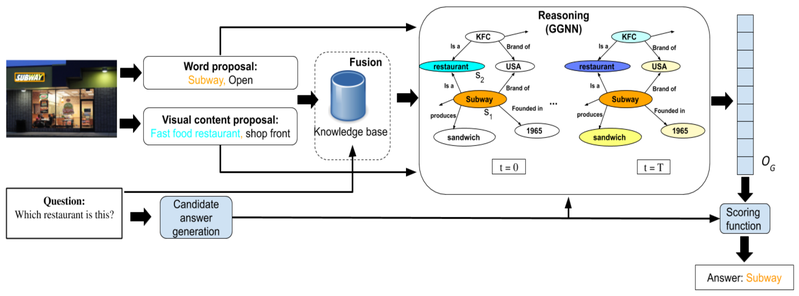


Traditional VQA
[Antol et al., ICCV'15, Zhang et al., ICLR'18]
Q: How many cars are there in this image?
A: 2
ST-VQA, text-VQA
[Biten et al., ICCV'19, Singh et al., CVPR'19]
Q: Which restaurant brand is written on the red wall?
A: KFC
Text + Knowledge-enabled VQA [This work]
Q: Can I get chicken dish here?
A: Yes

Contributions

- **text-KVQA:** first dataset to combine Text Recognition, Knowledge Graph Reasoning and VQA
- Novel GGNN formulation for knowledge-enabled VQA

Proposed Knowledge-enabled VQA Model



Question:
Is this an American brand?

Word proposals
[Gupta et al., CVPR'16]
: Subway, Open

Scene proposals
[Zhou et al., TPAMI'17]
: Fast food restaurant, shop front

Proposal Module

Fusion Module

Fuse proposals, knowledge base facts, and question to find relevant facts and construct graph

GGNN Reasoning

Perform gated graph neural network based reasoning to arrive at answer

text-KVQA – Proposed Dataset

- Large scale knowledge-enabled VQA dataset
- 257K images containing text in scene images, movie posters and book covers
- 1.3M question-answer pairs
- Associated web-scale knowledge bases

Dataset	Number of Images	Number of QA Pairs	Knowledge Enabled
text-KVQA (This work)	**257,380**	**1,322,272**	**Yes**
ST-VQA [ICCV'19]	23,038	31,791	No
OCR-VQA--200K [ICDAR'19]	207,572	1,002,146	No
text-VQA [CVPR'19]	28,408	45,336	No

Sample Images and QA Pairs

Q. Which mobile store is this?
A. Airtel.
SF. Airtel is a telecommunication company.

Q. Can I fill fuel in my care here?
A. Yes.
SF. HP is a petroleum industry.

Q. Does this showroom sells car?
A. Yes.
SF. Hyundai produces car.

Q. Is this an American brand?
A. No.
SF. Adidas is brand of Germany.

Q. Which restaurant is This?
A. IHOP.
SF. IHOP is a restaurant.

Q. Who is the director of this movie?
A. Joe Johnston.
SF. Jumanji is directed by Joe Johnston.

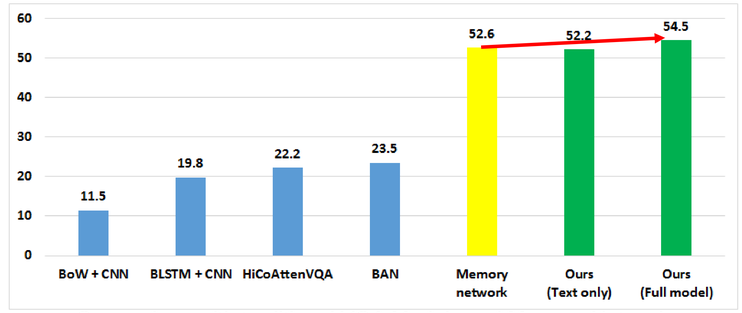
Q. Which bank is this?
A. Bank of America.
SF. Bank of America is a bank.

Q. Can I fill fuel in my car here?
A. No.
SF. Nissan is an auto-showroom.

Q. Can I get medicine here?
A. Yes.
SF. CVS sells medicines.

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

Experimental Results



Comparison with traditional VQA Models and Memory Networks

Word Recognition: {GALP}
Word Proposals: {GALP, GAP}
Scene Proposals: {clothing store, department store, gift shop}
Q. Can I get clothes here?
A (text only): No
A (full model): Yes

Fusions	Fact Recall (%)
W (photoOCR-1)	55.8
W (photoOCR-2)	59.9
V	20.8
q	5.3
W (photoOCR-1)+V+**q**	**58.9**
W (photoOCR-2)+V+**q**	**60.7**

Fact Recall (in %) at top-100 retrieval for text-KVQA

Some Qualitative Results

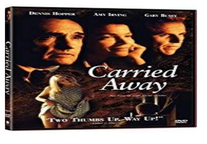
Words: {Ferrari}
Visuals: {auto showroom, garage}
Q. What is this?
A. Auto showroom.
SF. Ferrari is an auto showroom.

Words: {KFC}
Visuals: {fastfood restaurant, food court}
Q. Is this an American brand?
A. Yes
SF. KFC is a brand of USA.

Words: {Amul, Aral}
Visuals: {gas station, highway}
Q. Can I fill fuel in my car?
A. Yes
SF. Aral produces gas..

Words: {Saxon Geometry}
Visuals: {Travel, Maths}
Q. Who published this book?
A. Saxon Publishers
SF. Saxon Geometry is published by Saxon publishers.

Words: {bligh, CVS}
Visuals: {motel, pharmacy}
Q. What is this?
A. Pharmacy
SF. CVS is pharmacy.



Words: {carried, away}
Visuals: {Drama, Thriller}
Q. In which year this movie was released?
A. 1996
SF. Carried Away was released in 1996.

॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥