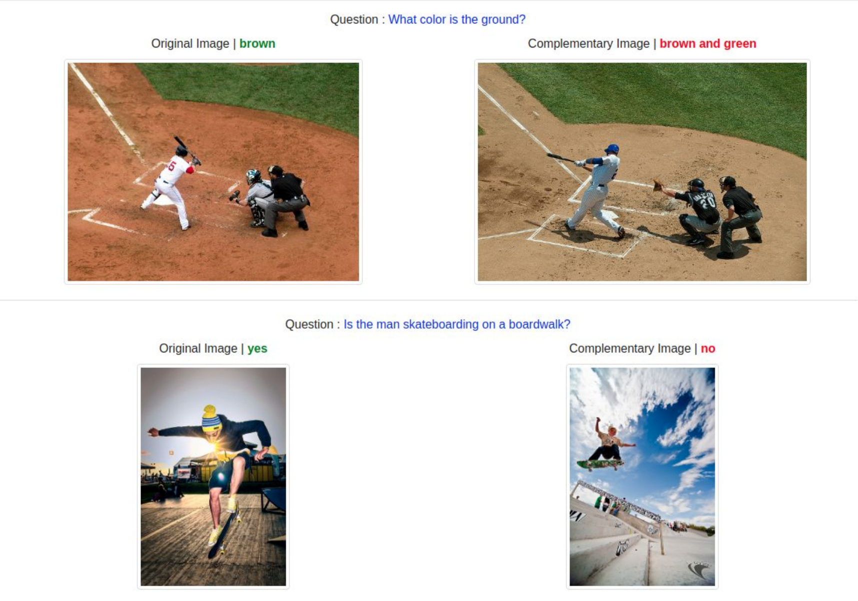
Knowledge-Enabled Visual Question Answering Model That Can Read And Reason

Anand Mishra[1] , Shashank Shekhar[2], Ajeet Kumar Singh[3] , Anirban Chakraborty[2]
[1]IIT Jodhpur , [2]IISc Bangalore, [3]TCS Research

Outline

- Problem & Background
- Related Work
- Text-KVQA Dataset
- Approach
    - Images + Text (ICDAR 2019)
    - Image + Scene-Text + Knowledge Graph (ICCV 2019) - *(covered in this talk)*
- Results
- Conclusions

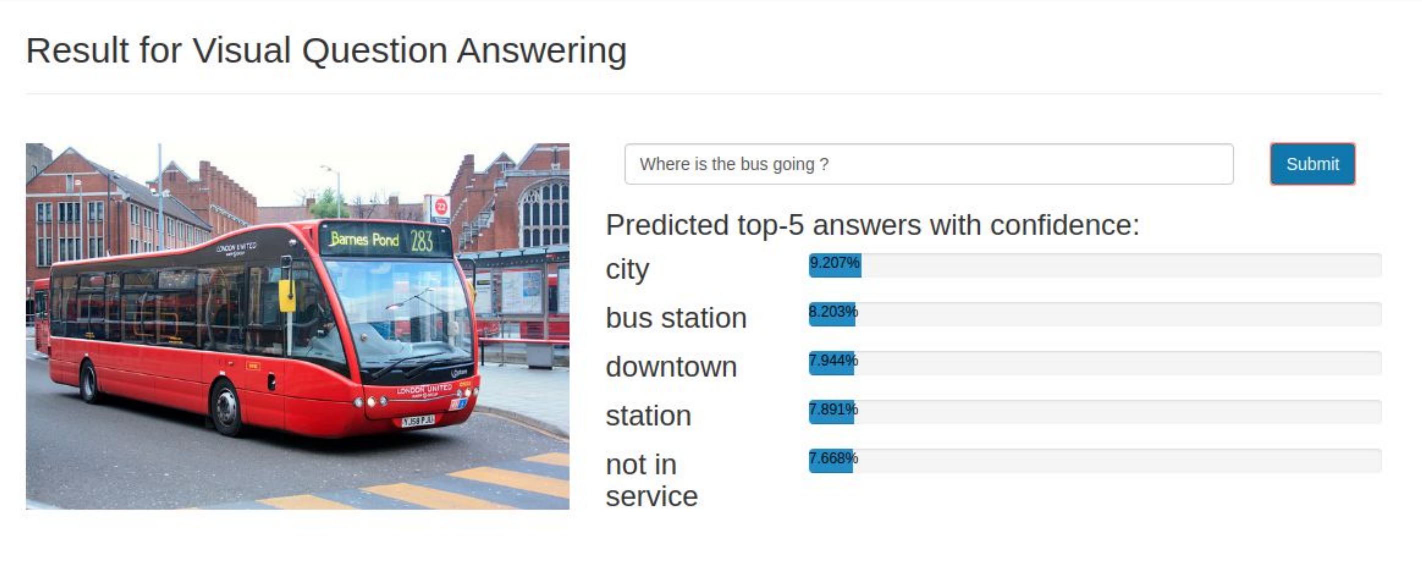
Problem



Visual Question Answering (Agrawal et al, ICCV 2015) is the problem of answering **natural language questions** based on **images**

Problem: VQA Limitations

- Present VQA approaches do not utilise scene text present in images



Problem: VQA Limitations

VQA methods are also limited by the visual knowledge present in images and cannot answer questions that require external knowledge (shown next)

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 name 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 name is written on the red wall?
A: KFC

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

Problem

Challenges:

- Scene understanding (traditional VQA)
- Scene-text recognition
- World knowledge

- New Problem: No existing datasets

Related Work

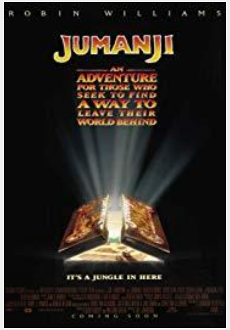
- **KVQA: Knowledge-Aware Visual Question Answering [Shah et al., AAAI'19]** Introduce Knowledge-Enabled QA in visual domain. Use FaceNet to get visual features, Bi-LSTM to get text features and use memory network for QA over KG upto 3 hops

- **Towards VQA Models That Can Read [Singh et al., CVPR'19]** Introduce text-based VQA problem. Use GloVE for questions embeddings, and CNN for visual features and OCR. Approach limited to answer from vocabulary + OCR detected words.

- **Scene Text Visual Question Answering [Biten et al., ICCV'19]** Alternate work which introduced text-based VQA simultaneously. Uses traditional VQA approach of CNN based visual and RNN based question embeddings. Performs retrieval over fixed vocabulary.

text-KVQA: A novel dataset

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

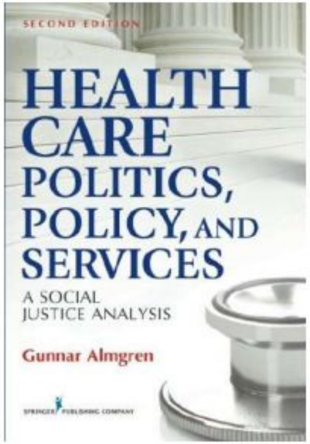


Q: When was this movie released?
A: **1995**

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

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

text-KVQA: A novel dataset



(a)
Q: What is the genre of this book?
A: **Medical Books**

(b)
Q: Does it sell Pizza?
A: **Yes**

(c)
Q: Which restaurant is this?
A: **Cafe Coffee Day**

(d)
Q: What is this?
A: **Book store**

(e)
Q: What does this store sell?
A: **Watches**

(f)
Q: Is this a Dutch brand?
A: **Yes**

Dataset available@

https://textkvqa.github.io

Proposed Solution

Proposed solution made of three separate modules:

- **Proposal Module**
  Generate word proposals from scene-text, visual scene content

- **Fusion Module**
  Combine representations from image (text and scene proposals) and questions

- **Reasoning Module**
  Use image + question representations to perform reasoning over external knowledge graph

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

Proposal Module

- **Scene proposals [Zhou et al., TPAMI'17]**

  Use VGGNet trained on MIT Places dataset to get scene proposals from the images.

- **Word proposals [Gupta et al., CVPR'16]**

  Used different text detection and recognition models pre-trained on MS COCO-Text including Contextual Text Proposal Network, EAST, TextSpotter (best) and PixelLink

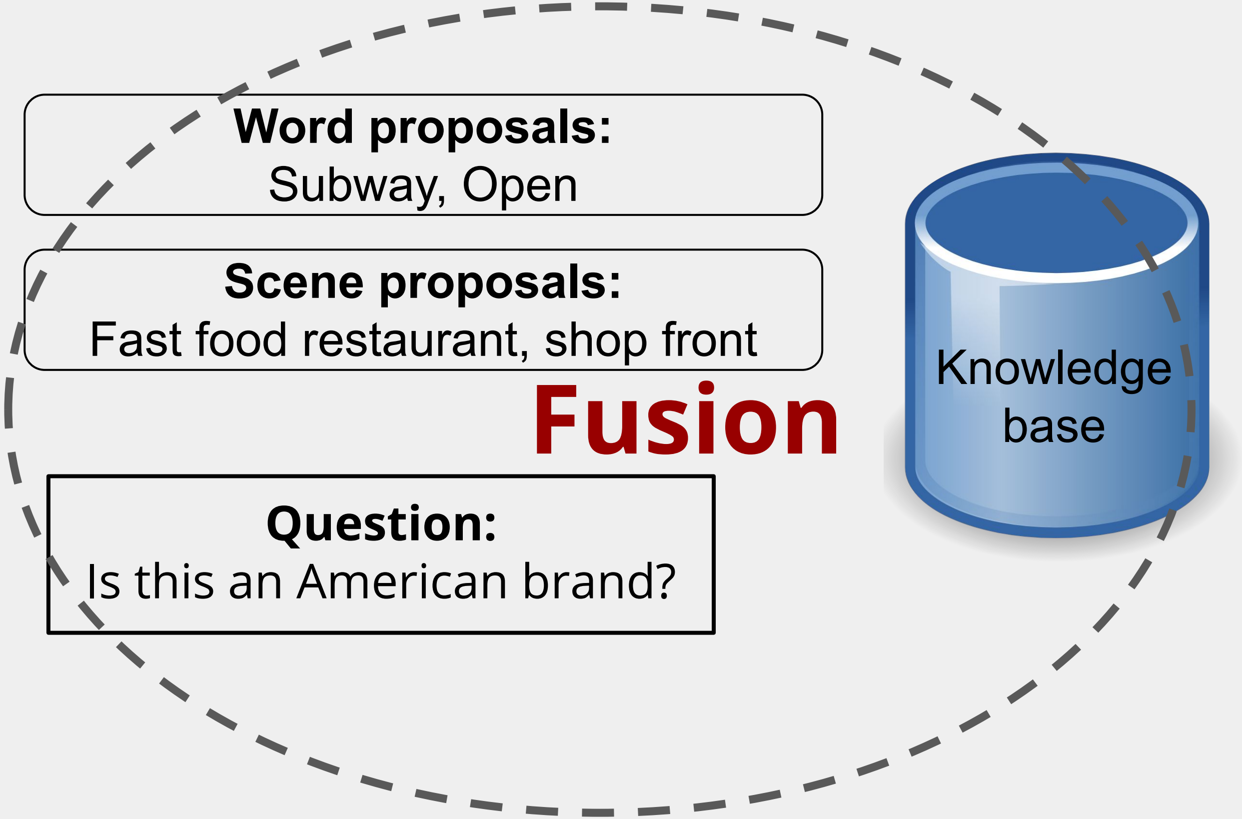
Proposed Solution

Proposal Module

- Visual content proposals V={v1, v2, ..., vm} along with their confidence scores are obtained using a VGGNet trained on MIT Places.

- Word proposals are generated using a pre-trained scene text detection and recognition model. A set of n words W={w1, w2, ..., wn} & their respective confidence scores is obtained by considering all words which are within normalized edit distance (NED) = 0.5 of a Knowledge Graph entity.

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

Fusion Module

- Let i_{th} fact of knowledge base be $f_i = (h_i, r_i, t_i)$ where h_i, r_i, t_i denote head entity, relation and tail entity, respectively.

- Given a set of word proposals W, visual content proposal V and question Q, fusion score for ith knowledge fact is computed as:

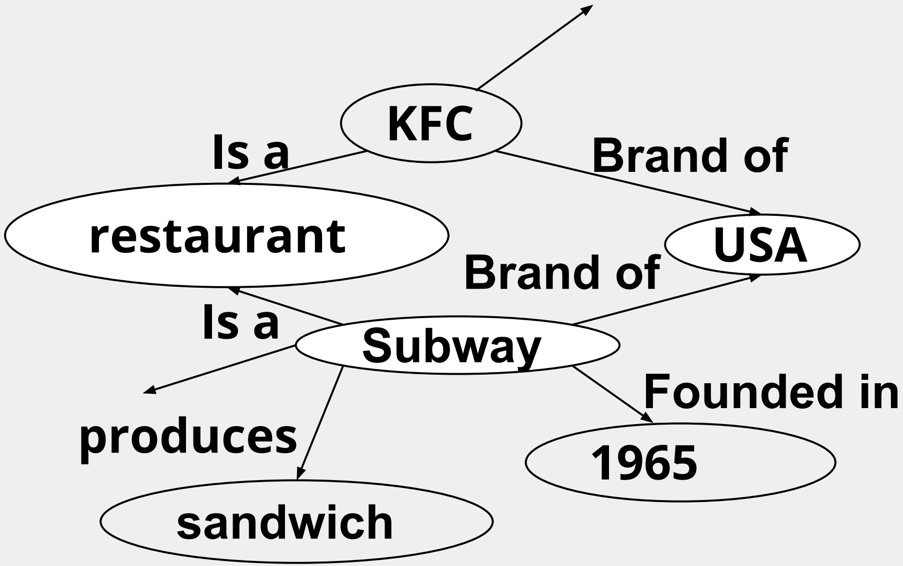
$$S(h_i, r_i, t_i) = \max_{j,k} \alpha_w s_{w_j} < w_j, (h_i, r_i, t_i) > + \alpha_v s_{v_k} < v_k, (h_i, r_i, t_i) >$$
$$+ \alpha_q < Q, (h_i, r_i, t_i) >.$$

(where all of w,v,Q,h,r,t are represented by their Word2Vec vector)

Proposed Solution

Word proposals:
Subway, Open

Scene proposals:
Fast food restaurant, shop front



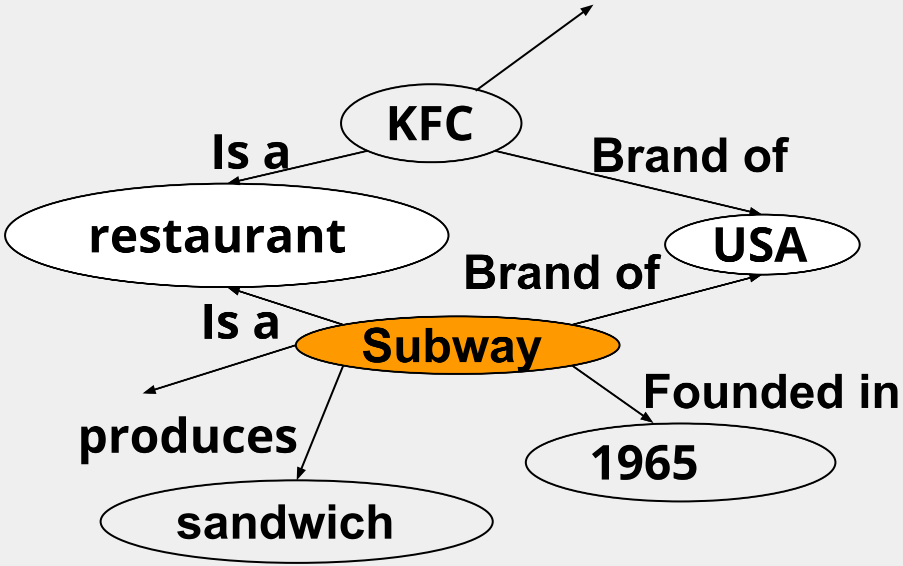
Question:
Is this an American brand?

Proposed Solution

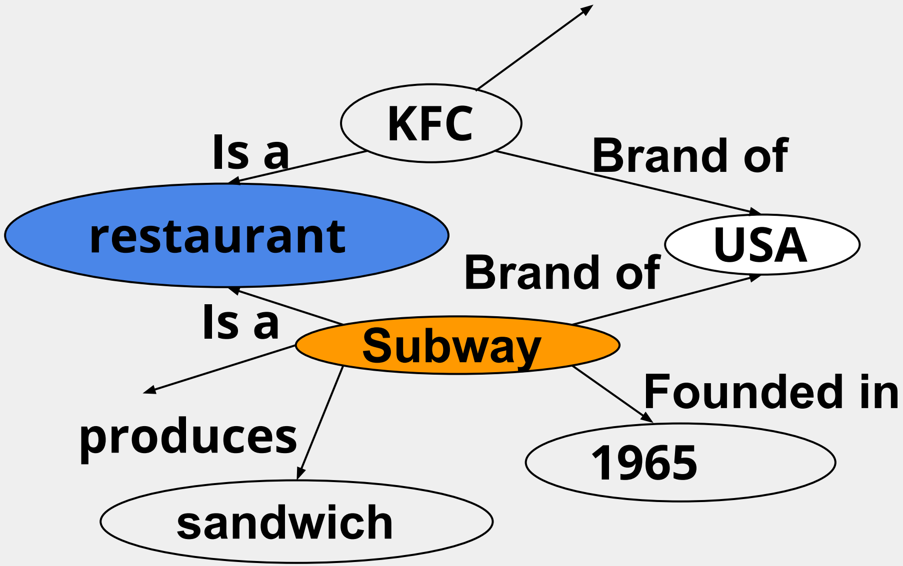
Word proposals:
Subway, Open

Scene proposals:
Fast food restaurant, shop front

Question:
Is this an American brand?



Proposed Solution

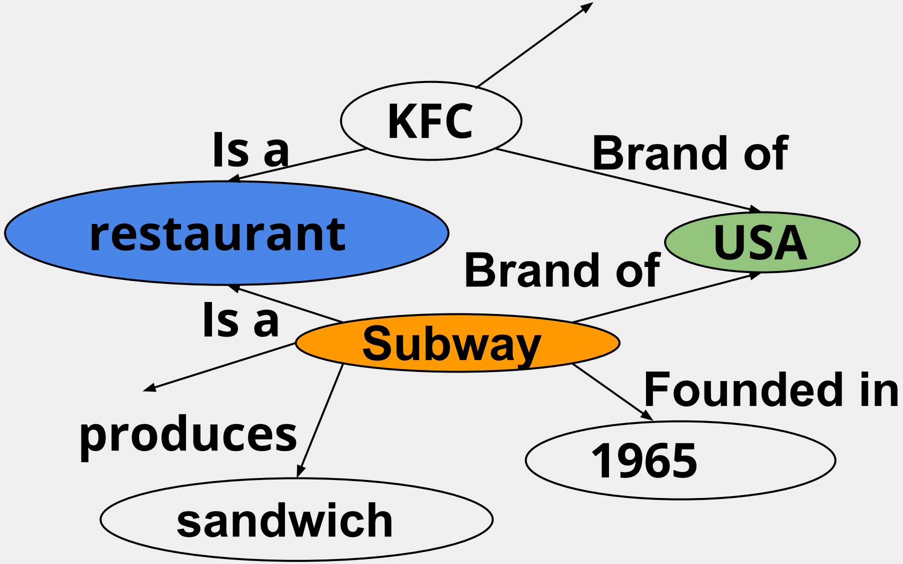


Word proposals:
Subway, Open

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

Reasoning Module

- Fusion score is used to retrieve top-K knowledge facts for each Question-Image pair and construct a multi-relational weighted graph.
- Candidate answers set A = $\{e_1, e_2, \ldots, e_m\}$ in one hop of anchor entity is generated by predicting answer type using a simple Bi-LSTM.
- For graph with N nodes, a task specific embedding of nodes x_u for node u, word proposals W & visual content proposals V, a graph-level embedding O_G is produced using a Gated Graph Neural Net **[Li et al., ICLR'15]**

Proposed Solution

Reasoning Module

- The node embeddings x_u for node u are initialised as:

$$\mathbf{x}_u = \begin{cases} [\mathbf{n}_u,\ 0,\ 1,\ c_u]; & \text{if node } u \text{ is a word proposal,} \\ [\mathbf{n}_u,\ 1,\ 0,\ c_u]; & \text{if node } u \text{ is an answer candidate,} \\ [\mathbf{n}_u,\ 1,\ 0,\ c_u]; & \text{if node } u \text{ has highest embedding} \\ & \text{similarity with the question,} \\ [\mathbf{n}_u,\ 0,\ 0,\ c_u]; & \quad \text{Otherwise.} \end{cases}$$

  Where n_u is the word2Vec embedding of the node and c_u is the confidence score when node represents word/scene proposal or 0 otherwise.

- The final graph level embedding O_g is obtained as (formulation next):

$$O_G = \tanh(\Sigma_{u \in \mathcal{U}} \ \ \sigma(f_\theta(\mathbf{h}_u^{(T)}, \mathbf{x}_u)) \ \odot \ \tanh(f_\phi(\mathbf{h}_u^{(T)}, \mathbf{x}_u)))$$

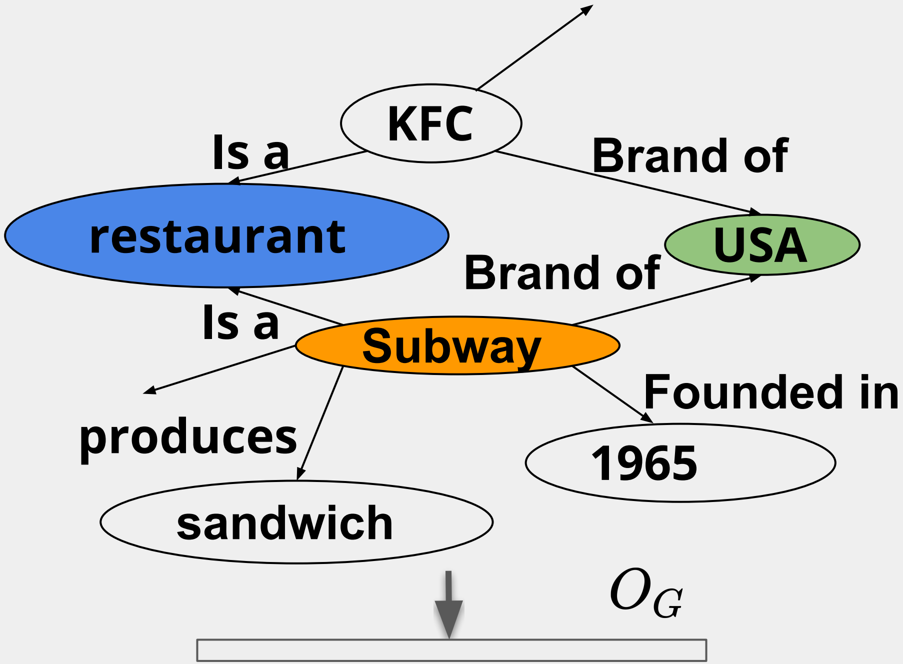
Proposed Solution

GGNN training formulation

- Hidden state dimension = 110
- Number of time steps = 5

- Output network: 2-layer fully connected network. In this, first layer activation is set to Sigmoid and second layer to tanh. The initial learning rate, momentum, batch size and maximum number of epochs is set to 0.1, 0.9, 16 and 100 respectively. Learning rate is decreased by a factor of 0.1 at every 10 epochs.

Proposed Solution



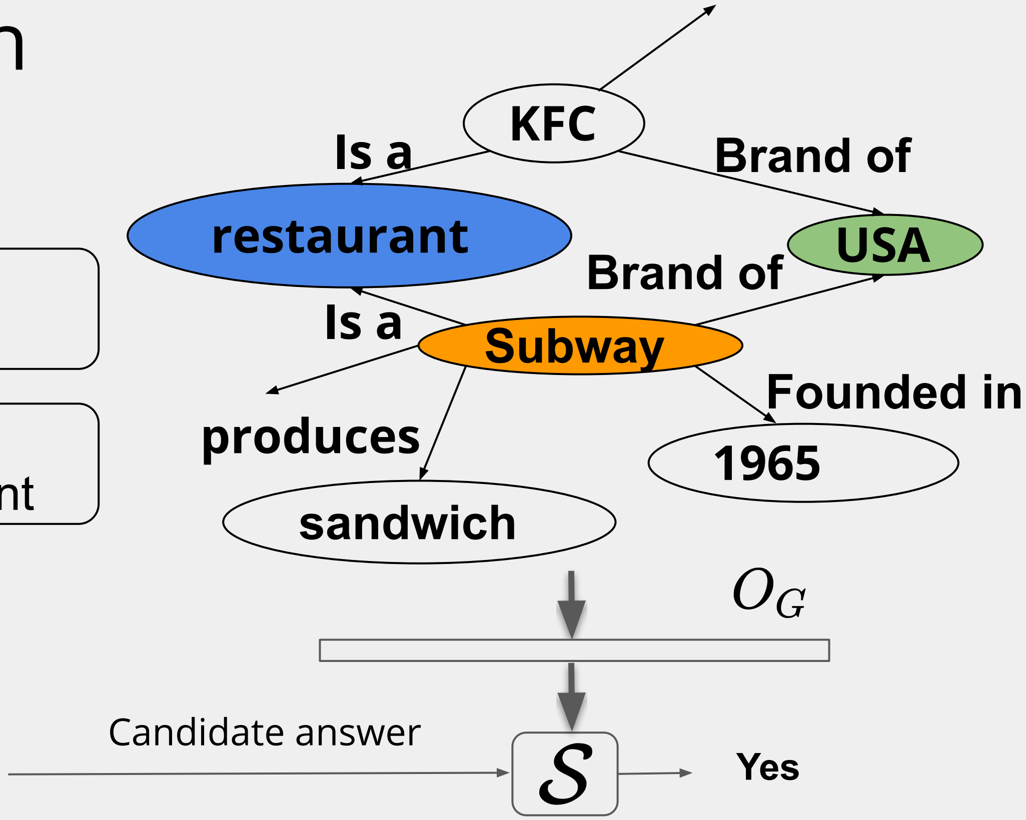
Word proposals:
Subway, Open

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]

Proposed Solution

Reasoning Module

- O_G and answer candidates e_i, are passed through a simple MLP classifier S to predict whether e_i is the correct answer. S is trained using binary cross-entropy loss.

Results

(a) **Detected words:** {GAP}
Word Proposals: {GAP, GALP}
Visual Content Proposal: {Clothing store, department store, gift shop}
Q: What is this store?
A: Clothing store
Supporting fact: GAP is a clothing store.
Observation: GALP is a petroleum brand, visual contents helps here to recover from lower precision in word proposals.

(b)
Detected words: {Baja, c, d}
Word Proposals: {Bata, c, d}
Visual Content Proposal: {Clothing store, Shoe shop, Gift shop}
Q: Which shoe shop is this?
A: Bata
Supporting fact: Bata is a shoe brand.
Observation: recovers from wrong recognition: Baja.

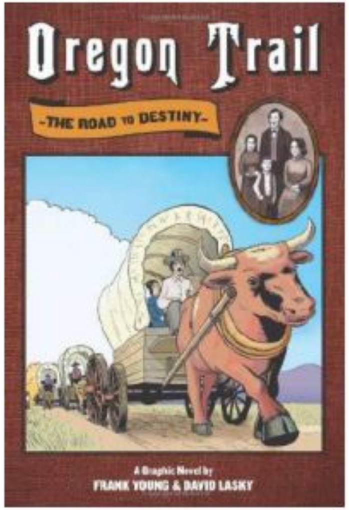
(c)
Detected words: {Arai, Arai, Arai, 11}
Word Proposals: {Aral, 11}
Visual Content Proposal: {Fastfood restaurant, Gas station, Industrial area}
Q: Is this a German brand?
A: Yes
Supporting fact: Aral is brand of Germany.
Observation: Top-1 place recognition goes wrong here, but word proposal helps.

Some examples where our model succeeds

Results

(a)
Detected words: {Lears}
Word Proposals: {Sears}
Visual Content Proposal: {Clothing store, Fastfood restaurant, Jewelery shop}
Q: Does it sell cloths?
A: Yes
Supporting fact: Sears is a clothing brand. **Observation:** Both word proposal and visual content mislead.

(b)
Detected words: {Luminosity, Shill}
Word Proposals: {Luminosity, Shell}
Visual Content Proposal: {Gas station, Fire station, General store}
Q: Can I fill fuel in my car here?
A: Yes
Supporting fact: Shell sells gas.
Observation: Word and visual content both misleads.



(c)
Detected words: {Oregon, Trail, The, Road, to, Destiny, Frank, Young, David, Laski}
Word Proposals: {Oregon Trail: The Road to Destiny, Young, David, Laski}
Visual Content Proposal: {Children's Books, Arts and Photography, Travel}
Q: What is the title of this book?
A: Oregon Trail: The Road to Destiny
Supporting fact: Oregon Trail: The Road to Destiny is a Children's Books.
Observation: Works even for long answers.

Some examples where our model fails

Results

Method	text-KVQA (scene)		text-KVQA (book)		text-KVQA (movie)	
	Original	NED=0.5	Original	NED=0.5	Original	NED=0.5
CTPN + CRNN	0.16	0.38	0.15	0.27	0.22	0.37
EAST + CRNN	0.36	0.60	0.43	0.66	0.24	0.42
Text Spotter	0.38	0.58	**0.53**	**0.70**	**0.35**	**0.48**
PixelLink + CRNN	**0.43**	**0.64**	0.38	0.56	0.14	0.27

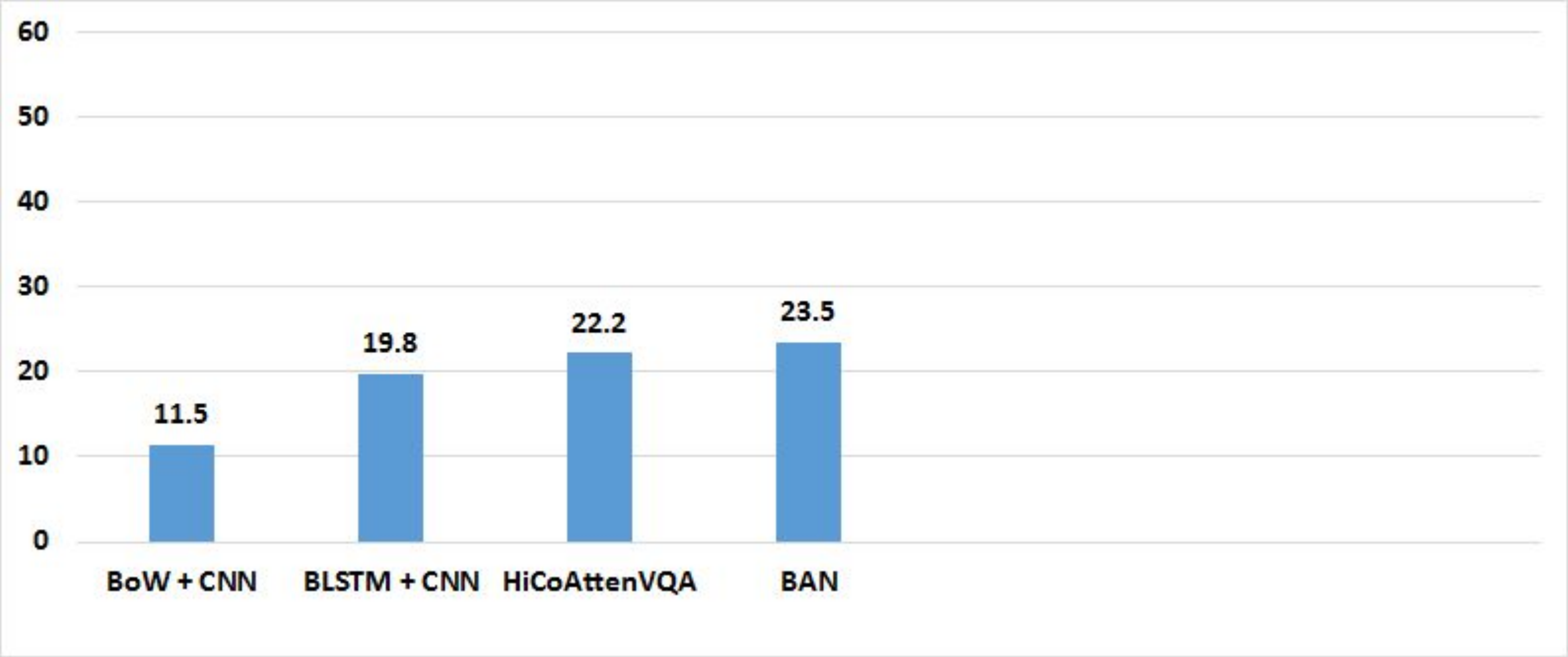
Performance of state-of-the-art text recognition models on our dataset. The two best performing methods are taken for word proposal generation. (photoOCR1 and 2)

Results

Fusions	Fact recall (in %)
W (*photoOCR1*)	55.8
W (*photoOCR2*)	59.9
V	20.8
q	5.3
W (*photoOCR1*)+V+**q**	58.9
W (*photoOCR2*)+V+**q**	**60.7**

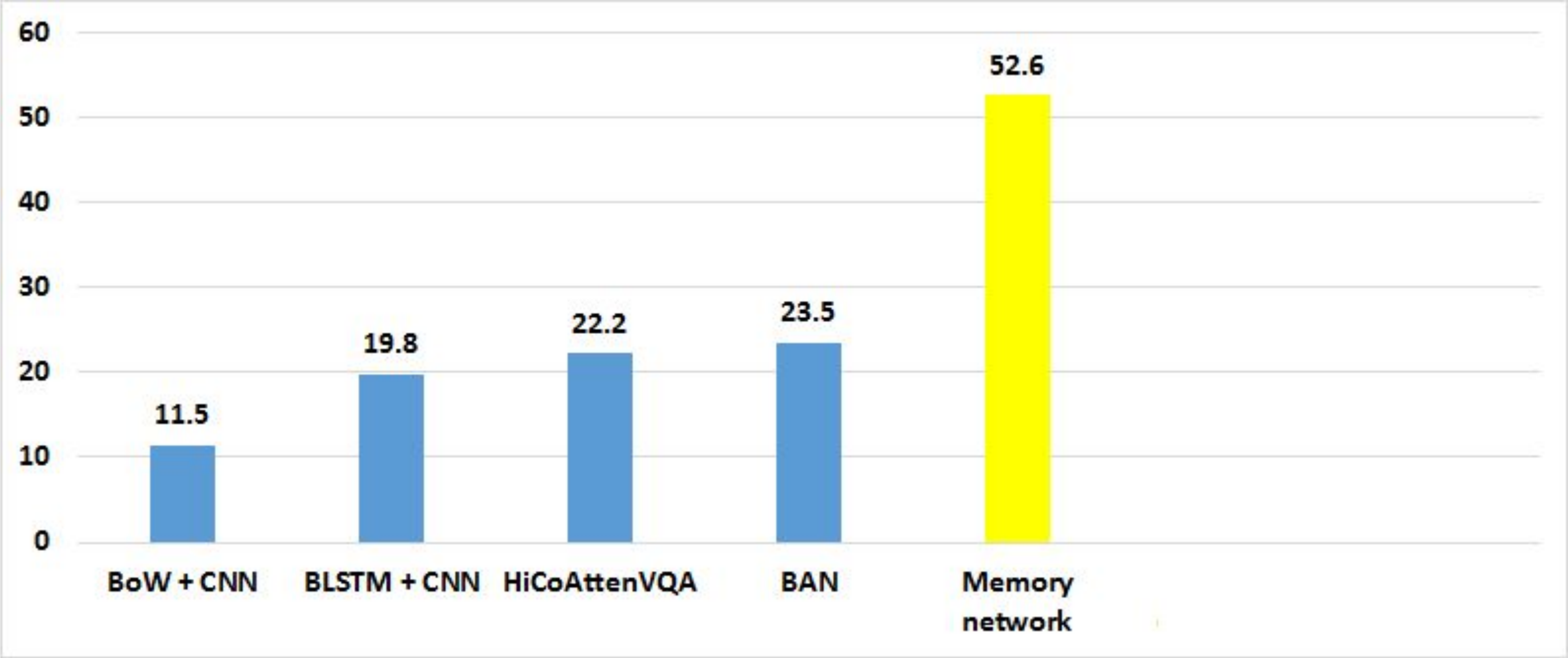
Performance of our fusion module. Also shows importance of word, visual content and question representations individually.

Results



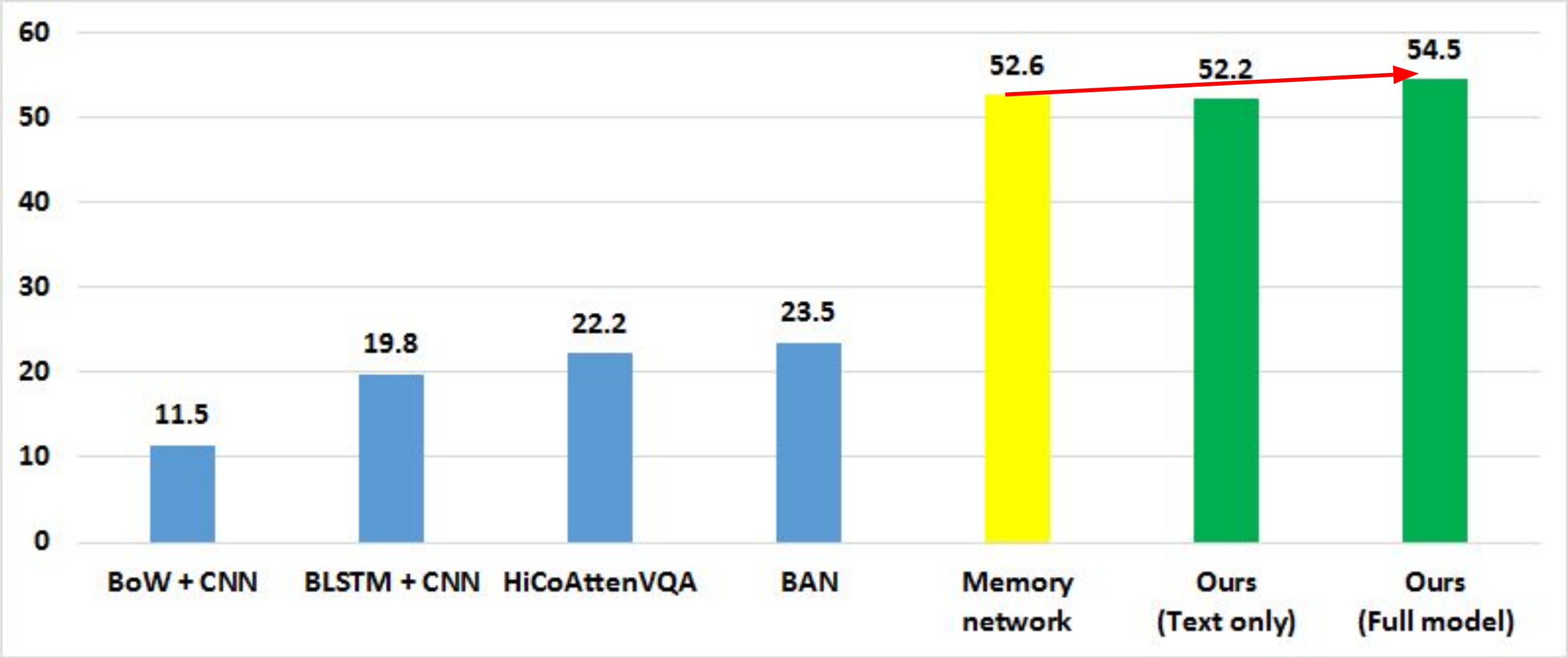
Traditional VQA methods are not successful

Results



A popular QA over KB method improves the performance (once we have the word embeddings)

Results



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

Results

Method	text-KVQA (scene)	text-KVQA (book)	text-KVQA (movie)
Traditional VQA methods			
BoW + CNN	11.5	8.7	7.0
BLSTM (language Only)	17.0	12.4	11.3
BLSTM + CNN	19.8	17.3	15.7
HiCoAttenVQA	22.2	20.2	18.4
BAN	23.5	22.3	20.3
QA over KB based method			
Memory network(with *photoOCR-1*)	49.0	57.2	42.0
Memory network(with *photoOCR-2*)	52.6	47.8	22.2
Our variants			
Vision only	21.8	19.8	18.2
Text only (with *photoOCR-1*)	48.9	55.0	41.4
Text only (with *photoOCR-2*)	52.2	48.6	20.5
Full model (with *photoOCR-1*)	52.2	**62.7**	**45.2**
Full model (with *photoOCR-2*)	**54.5**	49.8	23.0
Oracle (ideal text recognition)	80.1	71.3	76.2

Our model improves over other VQA models since they don’t utilise scene text or address zero-shot VQA. We also show that if we are able to get perfect word proposals the performance could still improve by a lot.

Conclusions

1. Introduce the problem of world knowledge based visual question answering which utilizes scene text.
2. **text-KVQA**: first dataset for **knowledge-enabled** VQA by reading text in image
3. **Novel GGNN formulation** for reasoning
4. Show the usefulness of each information source (scene, text, question) in our pipeline

OCR-VQA: Visual Question Answering by Reading Text in Images, Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, Anirban Chakraborty. International Conference on Document Analysis and Recognition 2019

From Strings to Things: Knowledge-enabled Visual Question Answering Model that can Read and Reason, Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, Anirban Chakraborty. International Conference on Computer Vision 2019 (oral)

Thank You