Visual Question Answering

Liyang Zhang

未来媒体研究中心
CENTER FOR FUTURE MEDIA
电子科技大学
University of Electronic Science and Technology of China

Outline

- Introduction
- Methods for VQA
- Dataset and Evaluation
- CVPR 2017

Outline

- Introduction
- Methods for VQA
- Dataset and Evaluation
- CVPR 2017

Introduction

- *What is VQA?*

Visual Question Answering. Given an image and a natural language question about the image, the task is to provide an accurate natural language answer.

What color are her eyes?
What is the mustache made of?

Free-form & Open-ended task

Q: Where is the kid pointing?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** green
(k) park	**(l)** up	**(m)** floor mat	**(n)** so people don't get wet
(o) down	**(p)** mom	**(q)** pharos	**(r)** ketchup pickle relish mustard

Q: How many people are in the picture on side of refrigerator?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** green
(k) 108 mph	**(l)** banana, apple	**(m)** 7	**(n)** 10 many
(o) fruit salad	**(p)** full swing	**(q)** 5	**(r)** vattenfall strom fur gewinner

Multiple-choice task

Introduction

Blank Sentence : He slows down in front of one _____ with a triple garage and box tree on the front lawn and pulls up onto the driveway.
Answer : house **Our result** : house

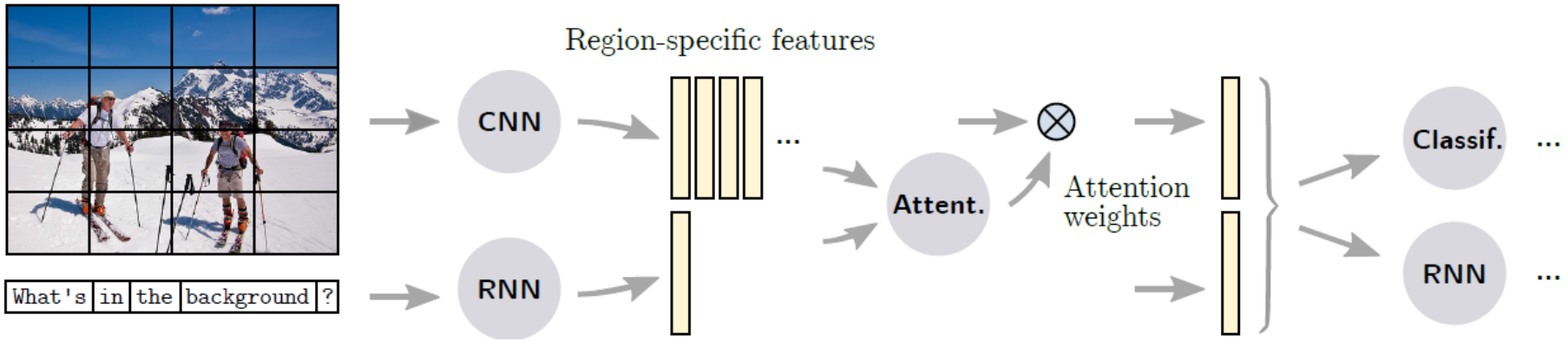
Fill-in-the-blank task

Outline

- Introduction
- Methods for VQA
- Dataset and Evaluation
- CVPR 2017

Methods for VQA

Attention



motivation

- Global features to represent the visual input may feed irrelevant or noisy information to the prediction stage.

goal

- Use local image features and assign different importance to features from different regions.

Outline

- Introduction
- Methods for VQA
- Dataset and Evaluation
- CVPR 2017

Dataset and Evaluation

Dataset

Dataset	Source of images	Number of images	Number of questions	Num. questions / num. images	Num. question categories	Question collection	Average quest. length	Average ans. length	Evaluation metrics
DAQUAR [51]	NYU-Depth V2	1,449	12,468	8.6	4	Human	11.5	1.2	Acc. & WUPS
COCO-QA [63]	COCO	117,684	117,684	1.0	4	Automatic	8.6	1.0	Acc. & WUPS
FM-IQA [22]	COCO	120,360	-	-	-	Human	-	-	Human
VQA-real [3]	COCO	204,721	614,163	3.0	20+	Human	6.2	1.1	Acc. against 10 humans
Visual Genome [41]	COCO	108,000	1,445,322	13.4	7	Human	5.7	1.8	Acc.
Visual7W [100]	COCO	47,300	327,939	6.9	7	Human	6.9	1.1	Acc.
Visual Madlibs [95]	COCO	10,738	360,001	33.5	12	Human	6.9	2.0	Acc.
VQA-abstract [3]	Clipart	50,000	150,000	3.0	20+	Human	6.2	1.1	Acc.
VQA-balanced [98]	Clipart	15,623	33,379	2.1	1	Human	6.2	1.0	Acc.
KB-VQA [78]	COCO	700	2,402	3.4	23	Human	6.8	2.0	Human
FVQA [80]	COCO & ImageNet	1,906	4,608	2.5	12	Human	9.7	1.2	Acc.

- COCO-QA: Automatic question collection means turning the **image descriptions** part of the original COCO dataset into question/answer form. **High repetition** rate of the questions with 23.29%.
- VQA-real: It allows evaluation of **multiple-choice** setting, by providing 17 additional(incorrect) answers. VQA-abstract: The aim is to remove the need to parse real images. Lower ambiguity.
- Visual Genome/Visual7W: the largest dataset with **1.7 million** question/answer pairs. 7W means who, what, where, when, why, how and which.

Dataset and Evaluation

COCO-QA [63]

Q: What is the color of the bus ?

A: yellow

VQA-real [3]

Q: What shape is the bench seat ?

A: oval, semi circle, curved, curved, double curve, banana, curved, wavy, twisting, curved

VQA-abstract [3]

Q: Who looks happier ?.

A: old person, man, man, man, old man, man, man, man, man, grandpa

Visual Genome [41]

Q: How is the ground ?

A: dry

Dataset and Evaluation

Evaluation metrics

Toronto COCO-QA	Acc. (%)	WUPS @0.9	WUPS @0.0
GUESS[63]	6.65	17.42	73.44
VIS+LSTM[63]	53.31	63.91	88.25
Multimodal-CNN [49]	54.95	65.36	88.58
2-VIS+BLSTM[63]	55.09	65.34	88.64
VIS+BOW[63]	55.92	66.78	88.99
QAM [11]	58.10	68.44	89.85
DPPnet [58]	61.19	70.84	90.61
Attributes-LSTM [85]	61.38	71.15	91.58
SAN [92]	61.60	71.60	90.90
Bayesian [34]	63.18	73.14	91.32
HieCoAtt [48]	65.40	75.10	92.00
ACK [87]	69.73	77.14	92.50
ACK-S [86]	70.98	78.35	92.87

Results on the COCO-QA

- Accuracy
  - the ratio of exact matches between predictions and answers.
  - the ratio of matches between predictions and 10 ground truth answers:

$$\text{accuracy} = \min\left(\frac{\#\text{ humans provided that answer}}{3}, 1\right)$$

- WUPS
  - Evaluate the **similarity** between common subsequence of predictions and answers, against two thresholds 0.9 and 0.0.

Dataset and Evaluation

Method	Test-dev					Test-standard				
	Open-ended				M.C.	Open-ended				M.C.
	Y/N	Num.	Other	All	All	Y/N	Num.	Other	All	All
Com-Mem [32]	78.3	35.9	34.5	52.6	-	-	-	-	-	-
Attributes-LSTM [85]	79.8	36.1	43.1	55.6	-	78.7	36.0	43.4	55.8	-
iBOWING [99]	76.5	35.0	42.6	55.7	-	76.8	35.0	42.6	55.9	62.0
Region-Sel [66]	-	-	-	-	62.4	-	-	-	-	62.4
DPPnet [58]	80.7	37.2	41.7	57.2	-	80.3	36.9	42.2	57.4	-
VQA team [3]	80.5	36.8	43.1	57.8	62.7	80.6	36.5	43.7	58.2	63.1
MLP-AQI [31]	-	-	-	-	-	-	-	-	-	65.2
SMem [90]	80.9	37.3	43.1	58.0	-	80.9	37.5	43.5	58.2	-
Neural-Image-QA [52]	78.4	36.4	46.3	58.4	-	78.2	36.3	46.3	58.4	-
NMN [2]	81.2	38.0	44.0	58.6	-	81.2	37.7	44.0	58.7	-
SAN [92]	79.3	36.6	46.1	58.7	-	-	-	-	58.9	-
ACK [87]	81.0	38.4	45.2	59.2	-	81.1	37.1	45.8	59.4	-
DNMN [1]	81.1	38.6	45.5	59.4	-	-	-	-	59.4	-
FDA [29]	81.1	36.2	45.8	59.2	-	-	-	-	59.5	-
ACK-S [86]	81.0	38.5	45.3	59.2		81.1	37.2	45.9	59.5	-
Bayesian [34]	80.5	37.5	46.7	59.6		80.3	37.8	47.6	60.1	-
DMN+ [89]	80.5	36.8	48.3	60.3	-	-	-	-	60.4	-
MCB [21]	81.7	36.9	49.0	61.1	-	-	-	-	-	-
DualNet [65]	82.0	37.9	49.2	61.5	66.7	81.9	37.8	49.7	61.7	66.7
MRN [38]	82.3	39.1	48.8	61.5	66.3	82.4	38.2	49.4	61.8	66.3
HieCoAtt [48]	79.7	38.7	51.7	61.8	65.8	-	-	-	62.1	66.1
MCB-Att [21]	82.7	37.7	54.8	64.2	-	-	-	-	-	-
Joint-Loss [57]	81.9	39.0	53.0	63.3	67.7	81.7	38.2	52.8	63.2	67.3
Ensemble of 7 models [21]	83.4	39.8	58.5	66.7	70.2	83.2	39.5	58.0	66.5	70.1

Results on the VQA-real test set

Outline

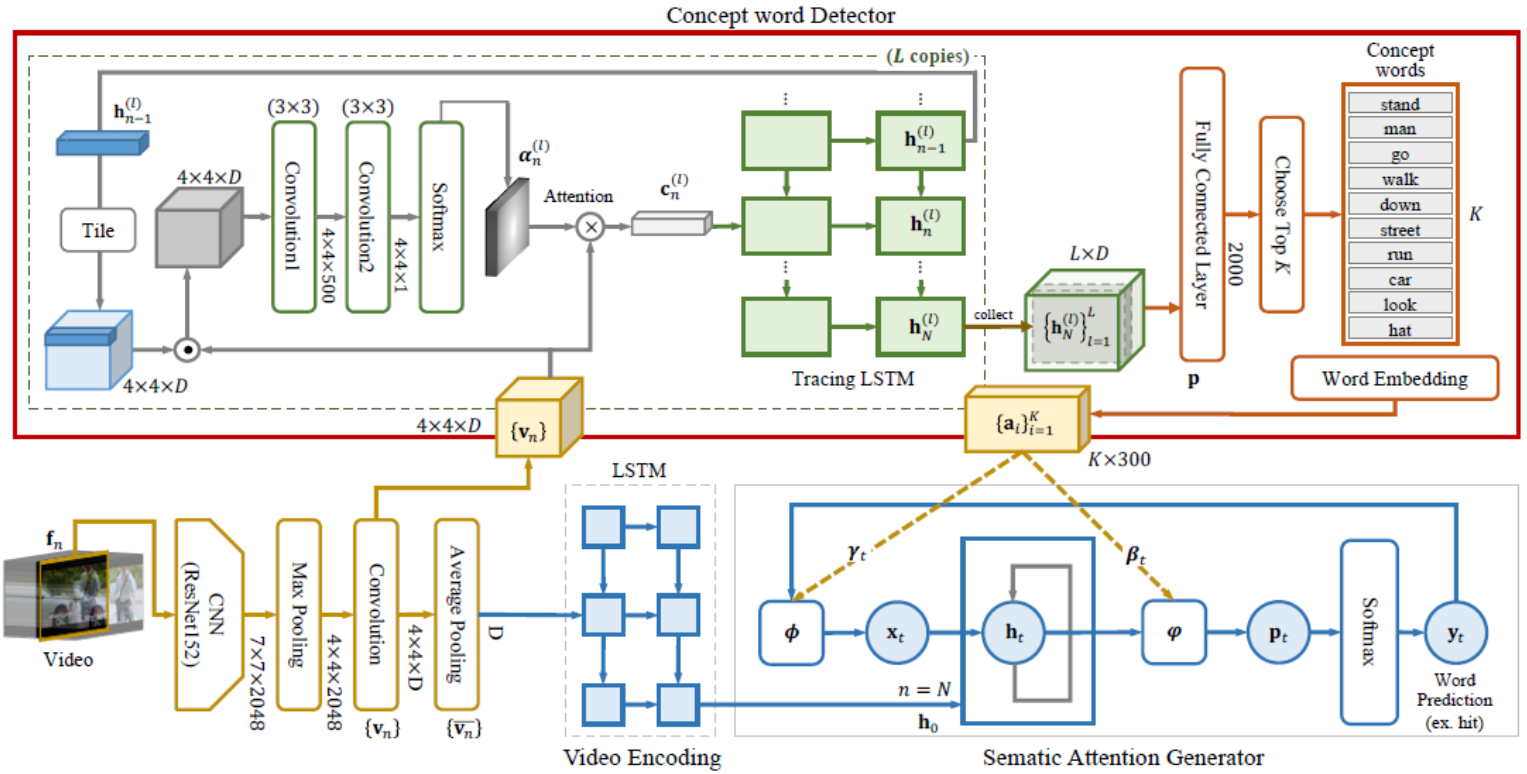
- Introduction
- Methods for VQA
- Dataset and Evaluation
- CVPR 2017

CVPR 2017

- End-to-end Concept Word Detection for Video Captioning, Retrieval, and Question Answering
- Graph-Structured Representations for Visual Question Answering
- Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering

CVPR 2017

- End-to-end Concept Word Detection for Video Captioning, Retrieval, and Question Answering



CVPR 2017

Concept words:

Blank Sentence : He slows down in front of one _____ with a triple garage and box tree on the front lawn and pulls up onto the driveway.

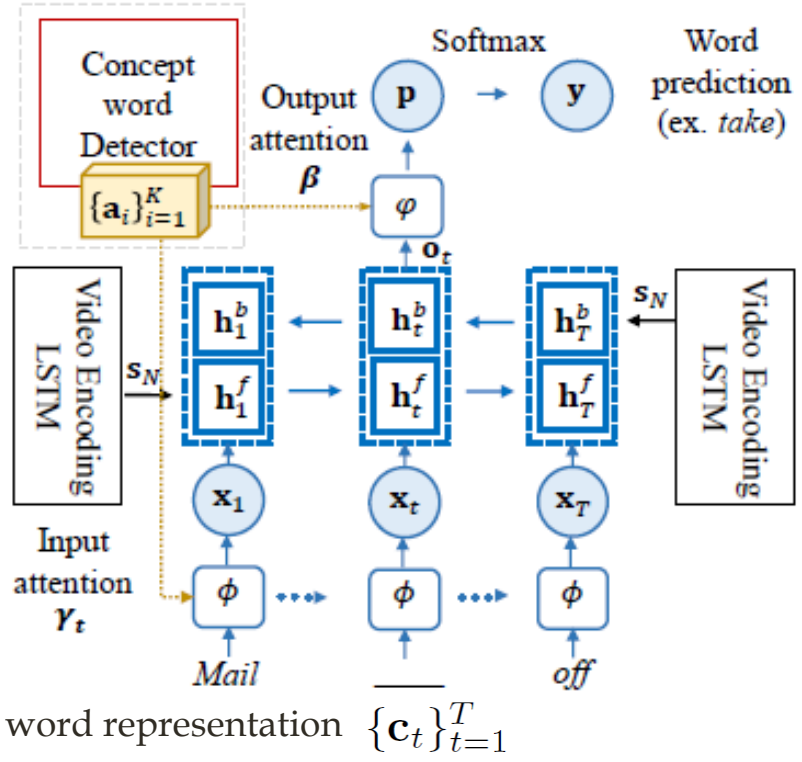
Answer : house **Our result** : house

Concepts : *drive, car, pull, down, front, outside, house, street, get, road*

(c)

CVPR 2017

Bi-directional LSTM



- **Task:** fill-in-the-blank
- **Dataset:** LSMDC (118,081 video clips)
- **Metric:** prediction accuracy

Fill-in-the-Blank	
Methods	Accuracy
Simple-LSTM	30.9
Simple-BLSTM	31.6
Base-SAN	34.5
amirmazaheri	34.2
SNUVL (Single)	38.0
SNUVL (Ensemble)	40.7
CT-SAN (Single)	41.9
CT-SAN (Ensemble)	**42.7**

CVPR 2017

- Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering

motivation

- Existing VQA models exploit language priors without truly understanding the visual content.

CVPR 2017

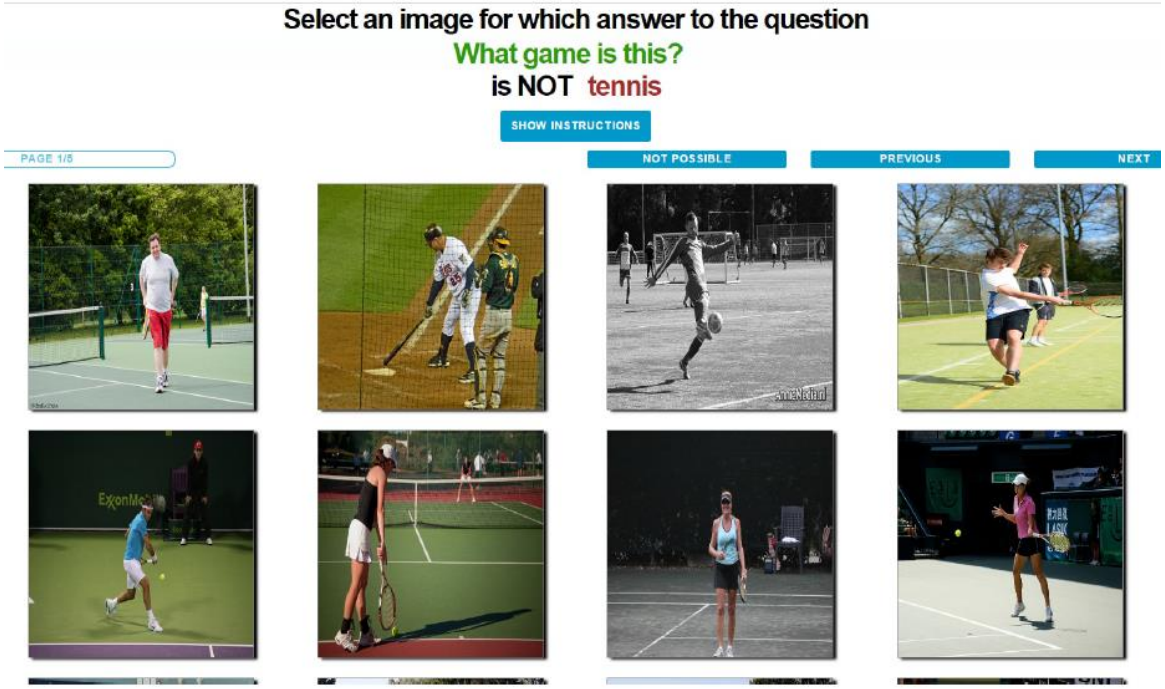
- Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering

Works

- Collect images to construct a balanced dataset. Every question with a pair of similar images have two different answers.
- Benchmark existing VQA models
- Propose a new explanation modality called counter-example

CVPR 2017

Collect images to construct a balanced dataset



Our goal

- to identify an image I′ similar to original image I, but its answer is different from A.

Two-stage data collection

- AMT workers collect targeted image from 24 nearest neighbors.
- 10 new AMT workers give answers on these new images.

CVPR 2017

- **Graph-Structured Representations for Visual Question Answering**

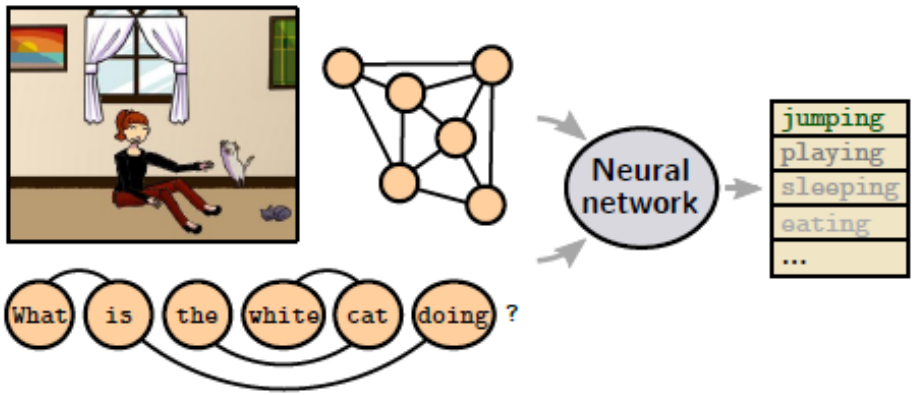
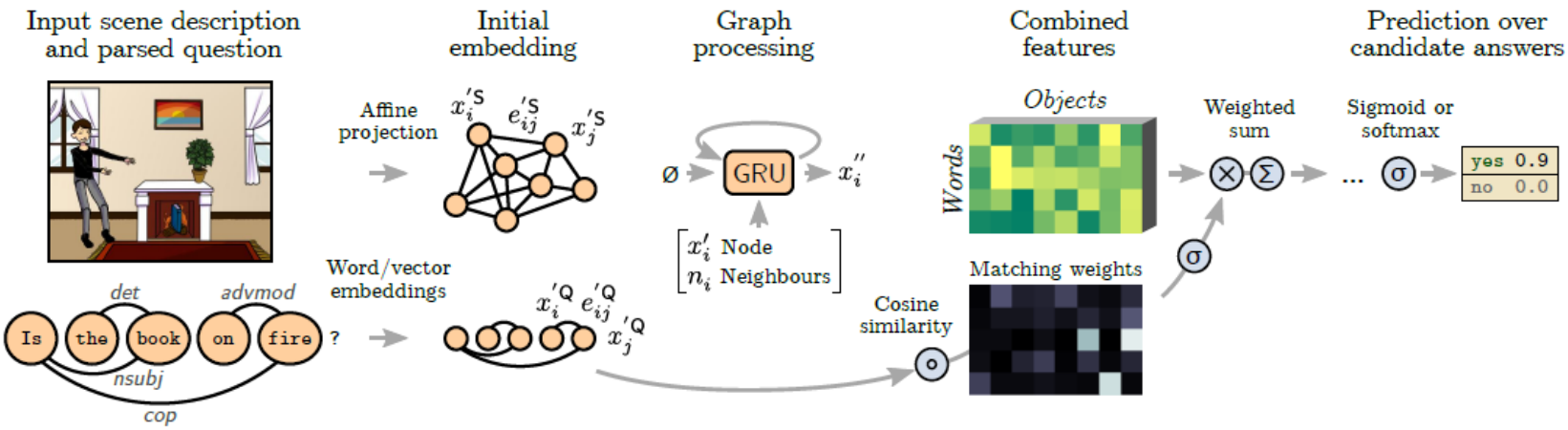


Figure 1. We encode the input scene as a graph representing the objects and their spatial arrangement, and the input question as a graph representing words and their syntactic dependencies. A neu-

motivation

- to improve VQA with structured representations of both scene contents and questions.
- CNN/LSTM-based approach to VQA largely ignores structure in the scene and in the question.

CVPR 2017



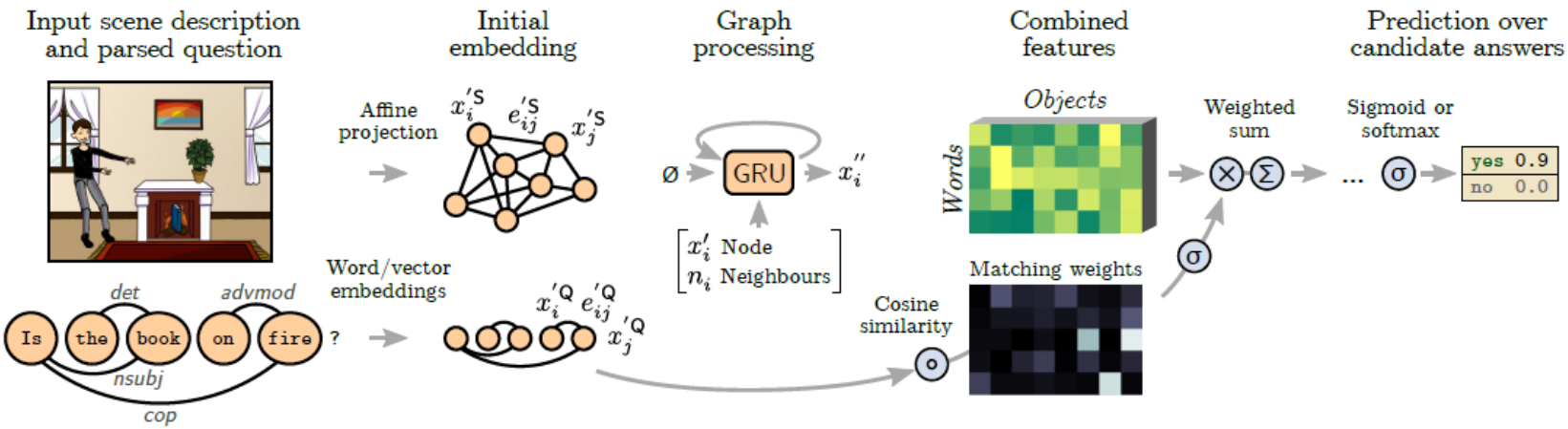
- Affine projection:

$$x_i'^{S} = W_3 x_i^{S} + b_3 \qquad e_{ij}'^{S} = W_4 e_{ij}^{S} + b_4$$

- Word/vector embedding:

$$x_i'^{Q} = W_1 [x_i^{Q}] \qquad e_{ij}'^{Q} = W_2 [e_{ij}^{Q}]$$

CVPR 2017



- GRU

$$h_i^0 = 0$$
$$n_i = \text{pool}_j(e'_{ij} \circ x'_j)$$
$$h_i^t = \text{GRU}(h_i^{t-1}, [x'_i; n_i]) \qquad t \in [1, T].$$

- Weights

$$a_{ij} = \sigma\left(W_5\left(\frac{x_i'^{Q}}{\|x_i'^{Q}\|} \circ \frac{x_j'^{S}}{\|x_j'^{S}\|}\right) + b_5\right)$$

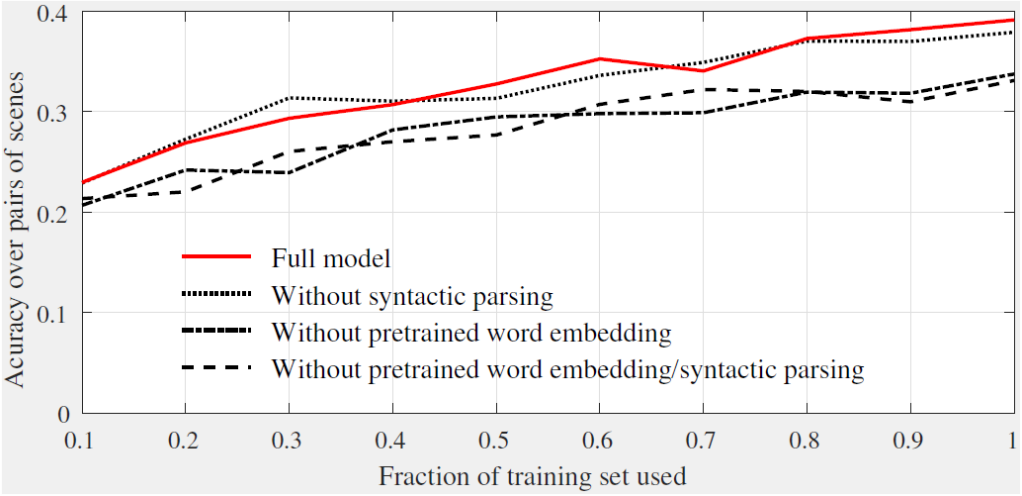
$$y_{ij} = a_{ij} \cdot [x_i''^{Q}; x_j''^{S}]$$

$$y'_i = f\left(W_6 \sum_j^{N^S} y_{ij} + b_6\right)$$
$$y'' = f'\left(W_7 \sum_i^{N^Q} y'_i + b_7\right)$$

CVPR 2017

experiments

- **Dataset**
  "balanced" dataset
  answer: yes/no
  for evaluating visual understanding



Method	Avg. score over scenes	Avg. accuracy over pairs
Zhang *et al.* [31] blind	63.33	0.00
with global image features	71.03	23.13
with attention-based image features	74.65	34.73
Graph VQA (full model)	**74.94**	**39.1**
(1) Question: no parsing (graph with previous/next edges)		37.9
(2) Question: word embedding not pretrained		33.8
(3) Scene: no edge features ($e_{ij}^{'S}$=1)		36.8
(4) Graph processing: disabled for question ($x_i^{''Q}$=$x_i^{'S}$)		37.1
(5) Graph processing: disabled for scene ($x_i^{''S}$=$x_i^{'Q}$)		37.0
(6) Graph processing: disabled for question/scene		35.7
(7) Graph processing: only 1 iteration for question (T^Q=1)		39.0
(8) Graph processing: only 1 iteration for scene (T^S=1)		37.9
(9) Graph processing: only 1 iteration for question/scene		39.1
(10) Uniform matching weights (a_{ij}=1)		24.4

Results on the test set of the "balanced" dataset

Thank you!