ICPR-2020 T3.6: Computer Vision, Robotics and Intelligent Systems

Dual Path Multi-Modal High-Order Features for Textual Content based Visual Question Answering

Yanan Li, Yuetan Lin, Hongrui Zhao, Donghui Wang
Institute of Artificial Intelligence, Zhejiang Lab
Tencent Youtu Lab
Institute of Artificial Intelligence, Zhejiang University

Virtual, 1/15, 2021

TextVQA Problem

- Environmental images contain rich textual contents.
- VizWiz study shows that up to 21% of question asked by visually-impaired people are related to the text in the environmental images.
- Current VQA models are incapable of reading and then reasoning about text.

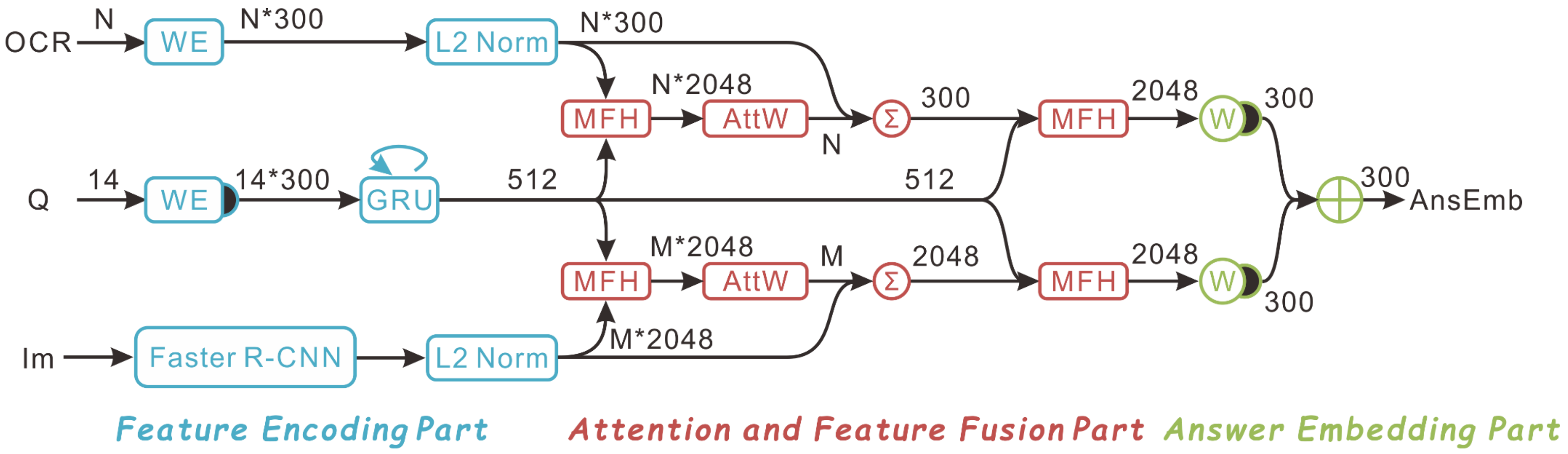
Q: What direction is shown?
A: west

Q: What is the score of the game?
A: 39-17

Q: What is the word above “music” on the top right corner?
A: activities

The Proposed Method

- Fuse question-image and question-OCR pairs by multi-modal high-order modules and attention mechanism to get the answer embedding.



M=100, N=50; WE: word embedding;

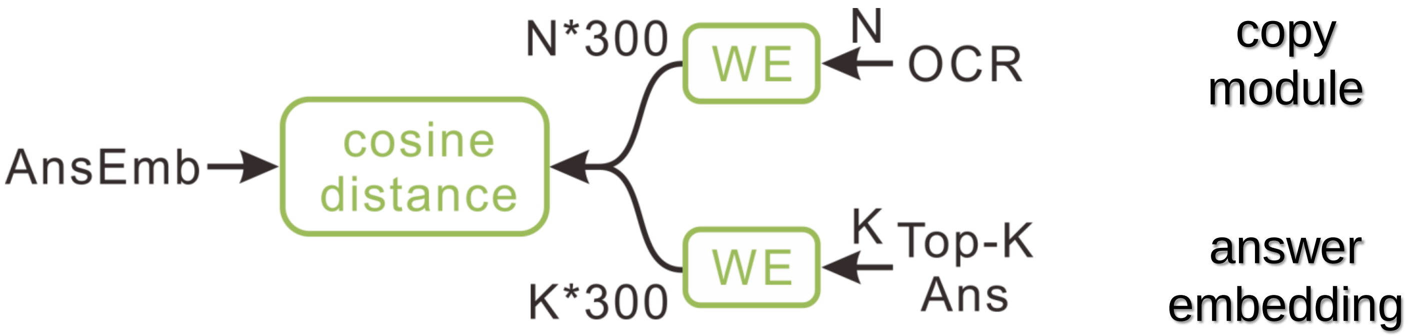
- Google BERT, 768d (BERT-Base, Uncased, 12-layer, 768-hidden, 12-heads)
- 0 initialization for blank OCR box
- PCA - 300d

MFH: MFH module

Black solid circle: hyperbolic tangent

* Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, “Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering,” IEEE Transactions on Neural Networks and Learning Systems, 2018.

The Proposed Method



- **Answer Prediction**
  - We use semantic word vectors to represent each answer, instead of one-hot vectors.
  - We map the fused text-question feature and image-question feature into the word embedding space, where we select the nearest answers as the prediction.

$$y = \arg\max_i s(\tilde{a}, a_i), i \in \{1, ..., K + N\}$$

Experimental Results

- TextVQA dataset:
  - It includes 28,408 images coming from Open Images and 45,336 text-related questions.
  - It also provides OCR information of each image recognized by Rosetta system.

Table 1. Ablation studies on TextVQA validation set. I, Q, O denotes image, question and OCR respectively.

Model	Ours	LoRRA
I+Q	**15.14%**	13.04%
I+Q(g)+O(g)	**21.43%**	18.35%
I+Q(g)+O(b)	**21.65%**	18.35%
I+Q(b)+O(b)	**22.10%**	18.35%
AnsEmb	**24.39%**	18.35%
OCR ans only	**26.87%**	20.06%
AnsEmb+OCR	**28.96%**	26.56%
ensemble11	31.18%	-
ensemble16	31.48%	-
ensemble23	**31.50%**	-

- BERT embeddings is slightly better than GloVe embeddings.
- Visual content and textual content provide complementary information for question anwering.
- OCR tokens are of high importance for answer prediction

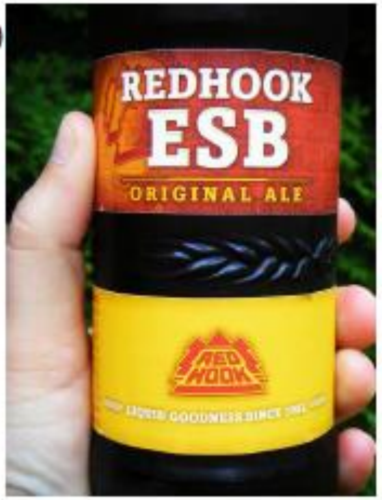
Experimental Results

Table II. Performance comparison with other models.

Model	val	test
Question Only Baseline	8.09%	8.70%
Image Only Baseline	6.29%	5.88%
Pythia Baseline	13.04%	14.0%
Pythia + LoRRA	26.56%	27.63%
Schwail	-	30.54%
Human	85.01%	86.79%
Ours	**31.50%**	**31.44%**

- We achieve the highest 31.44% with 11 models.
- There is still a significant performance gap between our method and humans.

(a)



Q: what type of drink is this?
O: [redhook, esb, original, ale, red, -hook, liquid, goodness, since, 1982]
A: ale
P: ale

(b)

Q: what does the last word on the label say?
O: [bambergs, spezto, heller, bamberg]
A: bamberg
P: bemberg

Thanks for your time!

END