Stacked Self-Attention Networks for Visual Question Answering

Qiang Sun, Yanwei Fu

Academy for Engineering & Technology, Fudan University
School of Data Science, Fudan University
Fudan-Xinzailing joint research centre for big data
ZheJiang Xin ZaiLing Technology Co. LTD

ICMR 2019 ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA RETRIEVAL
June 10-13 2019, Ottawa, Canada

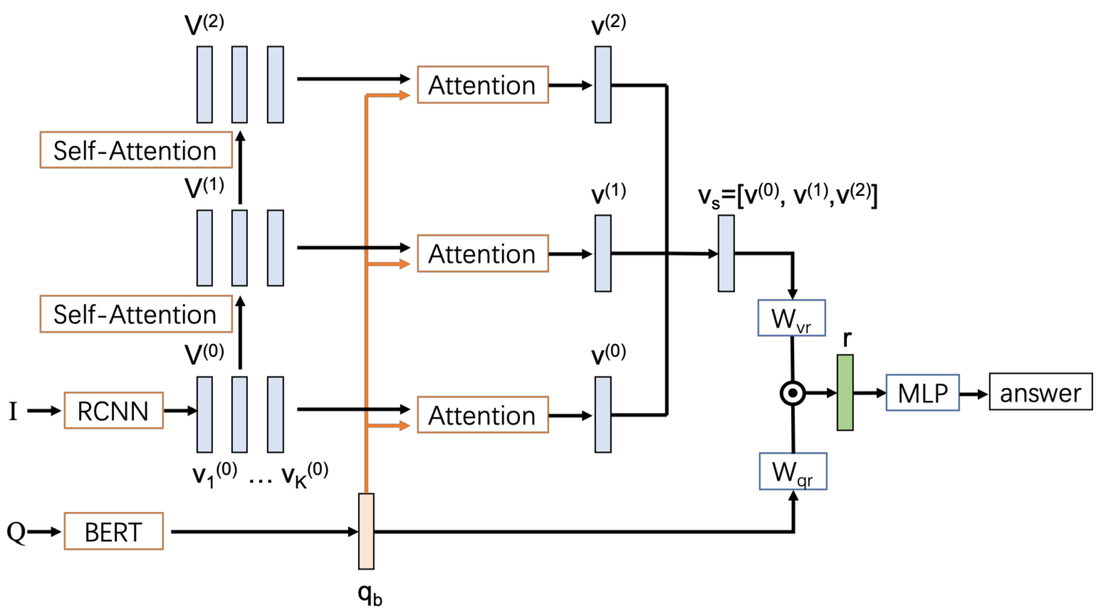
1. Introduction

☐ Motivation

– Stacked self-attention can be applied to learn the relations between objects.
– Bert-based question embedding model can be effective in VQA.

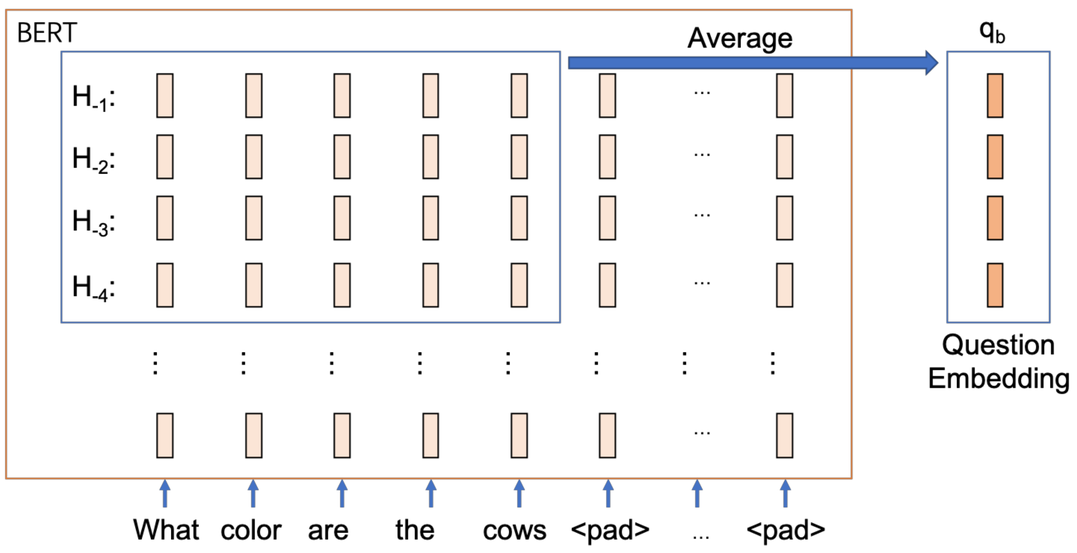
2. Methodology

Overview of stacked self-attention network



– The image is embedded by Faster R-CNN.
– The question is embedded by Bert-based model.
– Self-attention is applied to generate relations between objects.
– Question-guided attention is used to extract the visual features.

Question Embedding by BERT



The question embedding is calculated as the average of the last four hidden layers in BERT.

$$\mathrm{H} = BERT(Q, attention\_mask)$$

$$q_b = Avg([H_{-1}; H_{-2}; H_{-3}; H_{-4}])$$

3. Experimental Results

☐ Compared Methods:

– MCB
– MF-SIG-VG
– Bottom-up

☐ Results on VQA2.0:

Average Precision

Models	Test-dev Yes/No	Number	Other	Overall	Test-std Yes/No	Number	Other	Overall
MCB [5]	-	-	-	-	78.82	38.28	53.36	62.27
MF-SIG-VG [24]	81.29	42.99	55.55	64.73	-	-	-	-
bottom-up [15]	81.82	**44.21**	56.05	65.32	82.20	**43.90**	56.26	65.67
our full model	**82.81**	42.60	**56.58**	**65.80**	**83.24**	42.01	**56.78**	**66.14**

Table 1: Results on the test-dev and test-std splits of VQA v2.0 dataset.

☐ Ablation Study:

Average Precision

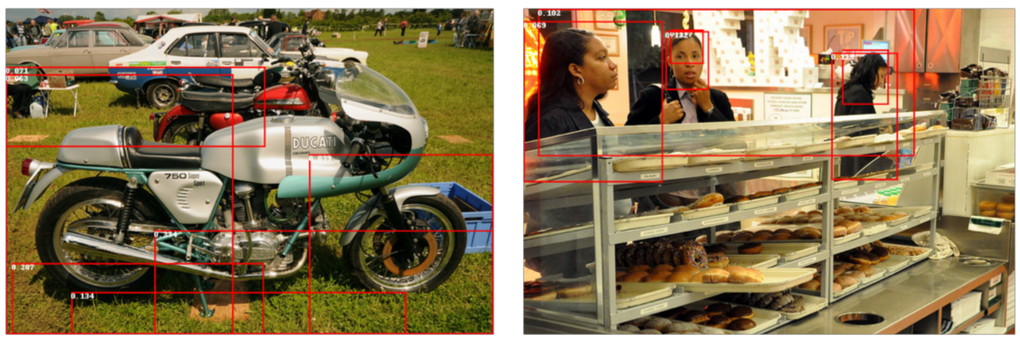
Models	Yes/No	Number	Other	Overall
our full model	**82.81**	42.60	**56.58**	**65.80**
w/o BERT	78.49	40.67	53.46	62.32
w/o Self-Attention	82.31	**43.03**	56.49	65.60

Table 2: Ablation study on the VQA v2.0 test-dev dataset. "w/o BERT" stands for the model using GRU instead of BERT component; "w/o Self-Attention" stands for the model without self-attention mechanism.

☐ Visualization of attention:

What is the dog doing?
Predict: laying down
Ans: laying down

Is the boy playing baseball?
Predict: yes
Ans: yes



What is the bike on?
Predict: grass
Ans: grass

How many women are in the image?
Predict: 2
Ans: 3

[5] A. Fukui, D. Park, D. Yang, et al. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 2016
[15] D. Teney, P. Anderson, X. He, and A. van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In CVPR, June 2018
[24] Chen Zhu, Yanpeng Zhao, Shuaiyi Huang, Kewei Tu, and Yi Ma. Structured attentions for visual question answering. In ICCV 2017