#1561 Co-attending Free-form Regions and Detections with Multi-modal Multiplicative Feature Embedding for Visual Question Answering

Tsinghua University

Pan Lu
Jianyong Wang

The Chinese University of Hong Kong

Hongsheng Li
Xiaogang Wang

East China Normal University

Wei Zhang

https://github.com/lupantech/dual-mfa-vqa

Motivation

Multi-modal feature embedding for VQA

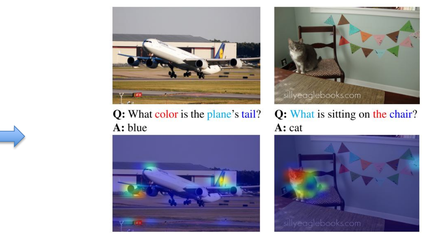
- Addictive, multiplicative, concatenation

Free-form region based attention mechanism

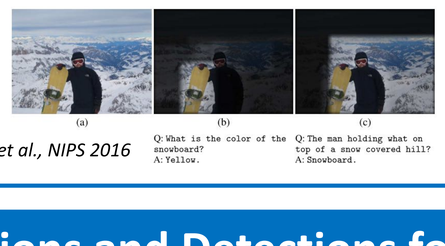
- Attend global visual context and specific foreground objects
- Focus on partial objects or irrelevant contexts sometimes

Detection-based attention mechanism

- Attend pre-specified detection boxes
- There might not exist a detection box such as the background

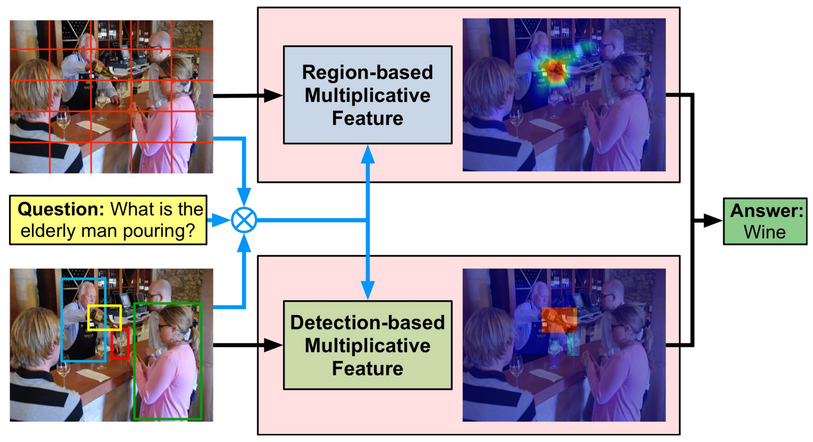


Wang et al., CVPR 2017



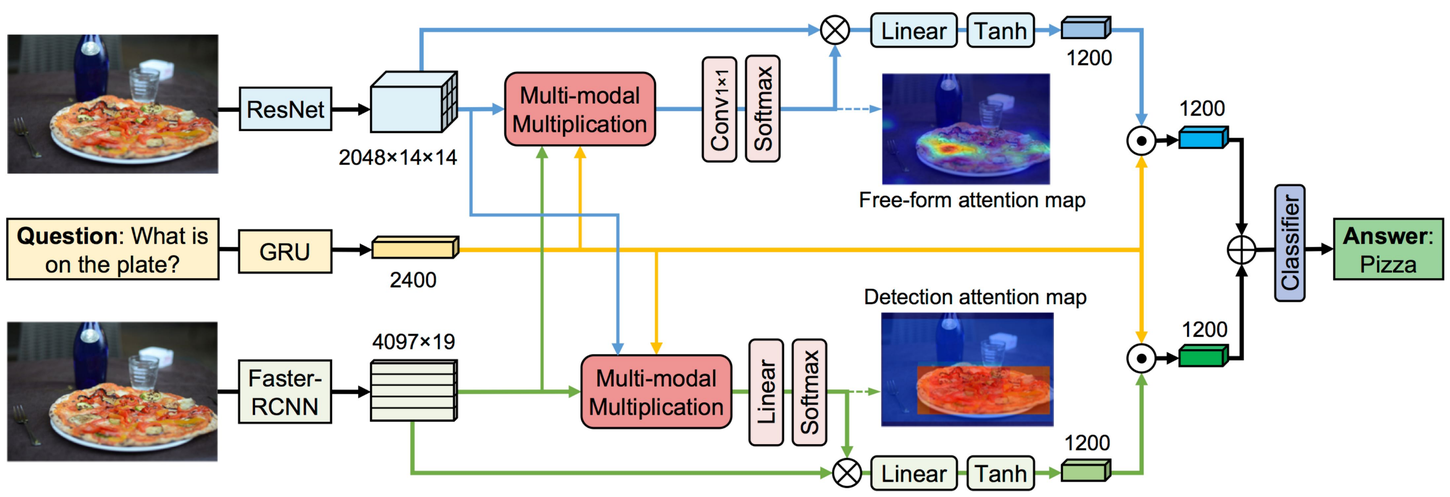
Li et al., NIPS 2016

Our Overall Network Structure



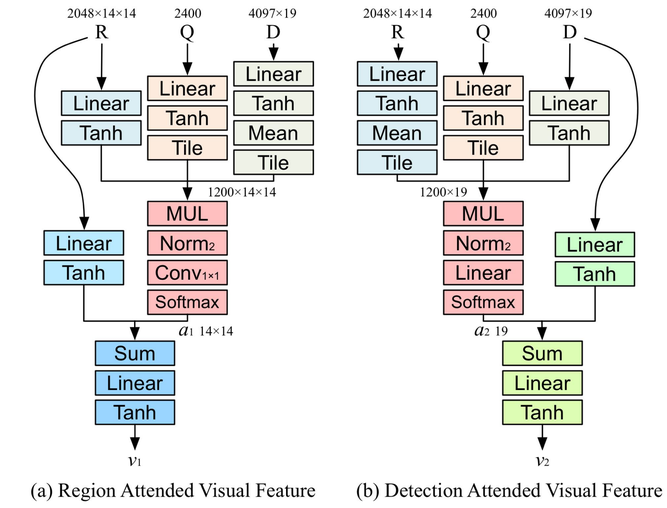
- Co-attention with multi-modal multiplicative embedding

Co-attending Free-form Regions and Detections for VQA



- Two attention branches with the proposed multiplicative feature embedding scheme
- One branch attends free-form image regions
- Another branch attends detection boxes for encoding question-related visual features

Multi-modal Multiplication



(a) Region Attended Visual Feature (b) Detection Attended Visual Feature

- Learning to attend free-form region visual features with multiplicative embedding
- Learning to attend detection-box visual features with multiplicative embedding

Experiments

VQA dataset

- 248,349 training, 121,512 validation and 244,302 testing questions

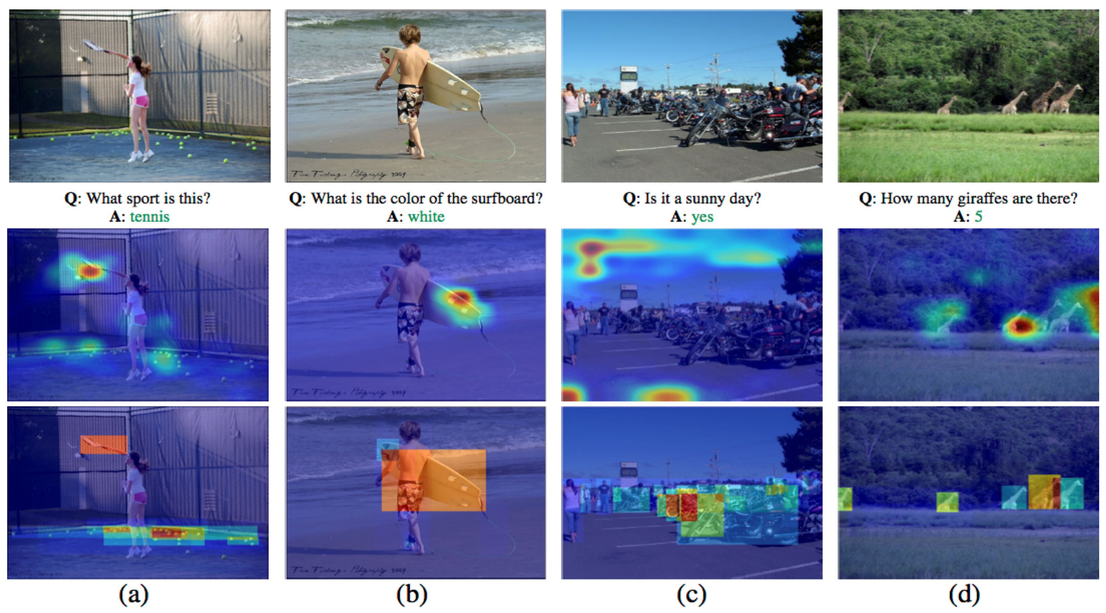
Method	Test-dev Open-Ended All	Y/N	Num.	Other	MC All
LSTM Q+I (Antol et al. 2015)	53.74	78.94	35.24	36.42	57.17
iBOWING (Zhou et al. 2015)	55.72	76.55	35.03	42.62	61.68
DPPnet (Noh, Hongsuck Seo, and Han 2016)	57.22	80.71	37.24	41.69	62.48
FDA (Ilievski, Yan, and Feng 2016)	59.24	81.14	36.16	45.77	64.01
DMN+ (Xiong, Merity, and Socher 2016)	60.30	80.50	36.80	48.30	-
Region Sel. (Shih, Singh, and Hoiem 2016)	-	-	-	-	62.44
QRU (Li and Jia 2016)	60.72	82.29	37.02	47.67	65.43
SAN (Yang et al. 2016)	58.70	79.30	36.60	46.10	-
HieCoAtt (Lu et al. 2016)	61.80	79.70	38.70	51.70	65.80
VQA-Machine (Wang et al. 2017)	63.10	81.50	38.40	53.00	67.70
MCB (Fukui et al. 2016)	64.70	82.50	37.60	55.60	69.10
MLB (Kim et al. 2017)	64.89	**84.13**	37.85	54.57	-
Dual-MLB (our baseline)	65.12	83.32	39.96	55.26	69.55
Dual-MFA (ours)	**66.01**	83.59	**40.18**	**56.84**	**70.04**

COCO-QA dataset

- 78,736 training and 38,948 testing samples

Method	All	Obj.	Num.	Color	Loc.	WUPS0.9	WUPS0.0
2VIS+BLSTM (Ren, Kiros, and Zemel 2015)	55.09	58.17	44.79	49.53	47.34	65.34	88.64
IMG-CNN (Ma, Lu, and Li 2016)	58.40	-	-	-	-	68.50	89.67
DDPnet (Noh, Hongsuck Seo, and Han 2016)	61.16	-	-	-	-	70.84	90.61
SAN (Yang et al. 2016)	61.60	65.40	48.60	57.90	54.00	71.60	90.90
QRU (Li and Jia 2016)	62.50	65.06	46.90	60.50	56.99	72.58	91.62
HieCoAtt (Lu et al. 2016)	65.40	68.00	51.00	62.90	58.80	75.10	92.00
Dual-MFA (ours)	**66.49**	**68.86**	**51.32**	**65.89**	**58.92**	**76.15**	**92.28**

Qualitative Evaluation



- (a), (b): Attend the corresponding image regions with two attention branches, which leads to correct answers with higher confidence
- (c), (d): Despite of only one attention branch attending the correct image region, two attention mechanisms are able to provide complementary information to obtain the correct answer