Contextualized Bilinear Attention Network for Visual Dialog

Gi-Cheon Kang[1], Byoung-Tak Zhang[1, 2]
Seoul National University[1], Surromind Robotics[2]
{gckang, btzhang}@bi.snu.ac.kr

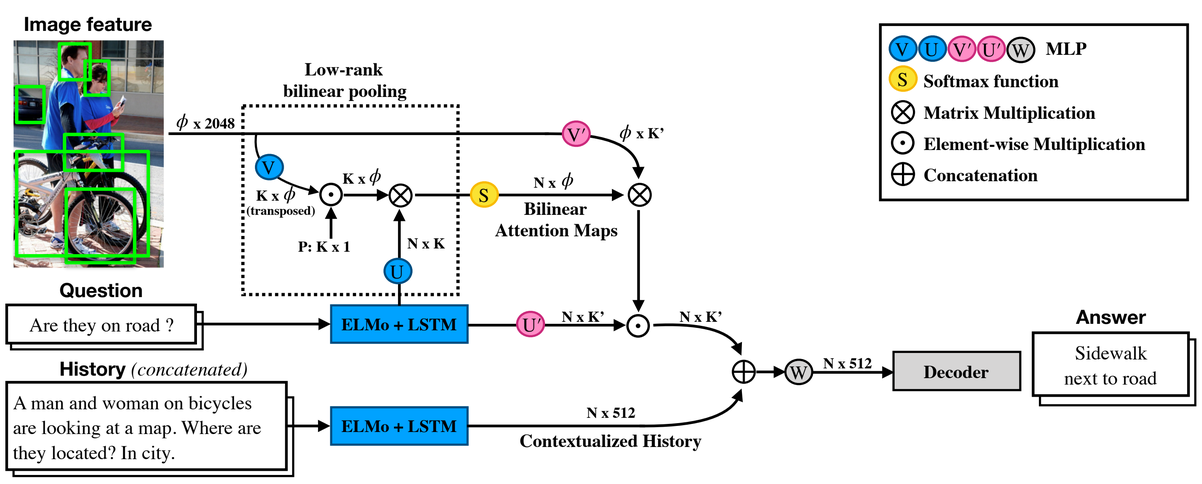
Introduction

An agent that can see everyday scenes and fluently communicate with people is one of the ambitious goals of artificial intelligence.

- **Visual Dialog**
  - Novel AI task introduced as a general version of visual question answering.
  - It requires to answer *a sequence of questions* which has an interdependent property.
- **Two key challenges**
  - Exploiting visually-grounded information.
  - Capturing the temporal topic of dialogs (Das et al., 2017)
- **Methods**
  - We extend the idea of Bilinear Attention Networks, BAN (Kim et al., 2018) to utilize visually-grounded information.
  - We employ newly proposed word embeddings, ELMo (Peters et al., 2018) to utilize a contextualized word representations.

Model

- **Overview of our proposed model**



- **Contextualized Bilinear Attention Network (CBAN)**

Inspired by **Low-rank Bilinear Pooling** (Kim et al., 2017), our model efficiently extracts **bilinear attention maps** of N questions. I and Q denote image feature and N question features, respectively.

$$\mathbf{A} = softmax((P^T \odot VI^T)QU^T)$$

As a sequence of questions has an interdependent property, capturing the context (*e.g.* co-reference and temporal topic) from previous conversion (history) is one of the key challenges. To make the best of ELMo, we define a history as follows. h_n and c denotes history of n*th* round and caption of image, respectively.

$$h_n = (c, (q_1, a_1), \cdots, (q_{n-1}, a_{n-1}))$$
$$H = (h_1, h_2, \cdots h_N)$$

CBAN gets I, Q, H and *attention map* as inputs. E denotes a fused representation of our model.

$$E = CBAN(I, Q, H; \mathbf{A})$$

Image feature	Faster-RCNN feature (pretrained)
Language feature	ELMo embedding + LSTM
Fusion method	Concatenation(BAN(I, Q), H)
Decoder type	Discriminative

Experimental Results

- **Quantitative Analysis**

Test-standard performance on Visual Dialog v1.0 dataset.

Model	ENS	ATT	MRR	R@1	R@5	R@10	Mean
HRE [1]	—	—	54.16	39.93	70.45	81.50	6.41
Memory Network [1]	—	—	55.49	40.98	72.30	83.30	5.92
Late Fusion [1]	—	—	55.42	40.95	72.45	82.83	5.95
Memory Network [1]	—	✔	56.90	42.43	74.00	84.35	5.59
Late Fusion [1]	—	✔	57.07	42.08	74.83	85.05	5.41
CBAN (ours)	—	✔	**57.53**	41.48	**76.95**	**88.52**	**4.49**
CBAN (ours, 2 models)	✔	✔	**58.86**	**42.85**	**78.70**	**90.38**	**4.13**

* ENS and ATT denote an ensemble method and a use of attention mechanism, respectively.
* Above performances are recorded in VisDial challenge leaderboard (https://evalai.cloudcv.org)

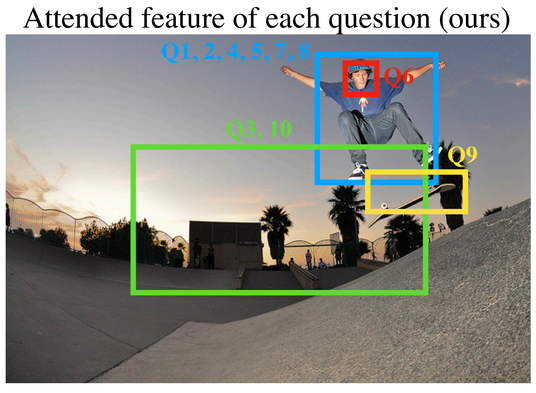
Validation performance on Visual Dialog v1.0 dataset.

Model	ENS	ATT	MRR	R@1	R@5	R@10	Mean
BAN (baseline) [2]	—	✔	54.59	39.74	71.76	82.13	6.20
CBAN (ours)	—	✔	**60.10**	**44.30**	**79.92**	**90.70**	**4.06**
CBAN (ours, 2 models)	✔	✔	**61.38**	**45.61**	**81.54**	**91.58**	**3.79**

- **Qualitative Analysis**

Original Image

Attended feature of each question (ours)



	Caption	A young man jumping his skateboard on a ramp		
	Question	**Answer**	**CBAN**	**BAN**
Q1	Where is he?	In a skate park	O	O
Q2	Is he the only one?	No, I see people in the background	O	O
Q3	Are there others skating?	No	O	O
Q4	**Are they watching him or doing something else?**	**Possibly watching him from a far**	O	X
Q5	Is he young or old?	Young	O	O
Q6	Does he have crazy hair?	No	O	O
Q7	**How about clothes?**	**Pants and Shirts**	O	X
Q8	**Does he look like he knows what he's doing?**	**Yes he does**	O	X
Q9	Can you see his skateboard?	Yes	O	O
Q10	**Does it look like a nice one?**	**Not really, wooden**	O	X

References

1. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2017)
2. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. arXiv preprint arXiv:1805.07932 (2018)
3. Peters, M.E., Neumann, M., lyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: NAACL (2018)
4. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J. W., & Zhang, B. T.: Hadamard product for low-rank bilinear pooling. In: ICLR (2017)