Bilinear attention networks for VizWiz challenge.

Jin-Hwa Kim[1], Yongseok Choi[1], Sungeun Hong[1],
Jaehyun Jun[2], and Byoung-Tak Zhang[2,3]

[1] SK T-Brain [2] Seoul National University [3] Surromind Robotics

Abstract. VizWiz challenge addresses two tasks: visual question answering and visual question answerability. Unlike conventional VQA datasets, VizWiz dataset was collected by the visually impaired using mobile phones and has the following characteristics: (1) the images include a significant amount of blur and unstable lighting while the questions include conversational style. (2) the questions can be unrelated to the images due to the limited visual information. In this paper, we propose bilinear attention networks (BAN), which exploits bilinear interactions between multimodal input channels, followed by two-layer MLPs for each task with a joint loss. Experimental results on VizWiz dataset show that the proposed method significantly outperforms previous methods.

Keywords: visual question answering · bilinear · attention

1 Introduction

The recently introduced VizWiz is designed to assist the visually impaired to overcome their daily visual challenges. The VizWiz dataset is more challenging than the existing VQA datasets because this consists of the images and questions obtained by the visually impaired in real-world scenarios. In this paper, we propose bilinear attention networks (BAN) to use bilinear attention distributions, on top of low-rank bilinear feature pooling technique. BAN exploits bilinear interactions among two groups of input channels, while the low-rank bilinear pooling extracts joint representations for each pair of channels. Finally, we train BAN with a joint loss for the conventional VQA task and the visual question answerability task.

2 Bilinear attention networks

We define BAN as a function of two multi-channel inputs parameterized by a bilinear attention map. Concretely, we generalize a bilinear model based on the unitary attention networks [4] for two multi-channel inputs, $\mathbf{X} \in \mathbb{R}^{N \times \rho}$ and $\mathbf{Y} \in \mathbb{R}^{M \times \phi}$, where $\rho = |\{\mathbf{x}_i\}|$ and $\phi = |\{\mathbf{y}_j\}|$, the numbers of two input channels, respectively. To reduce input channels simultaneously, we introduce a bilinear attention map $\mathcal{A} \in \mathbb{R}^{\rho \times \phi}$ as:

$$\mathbf{f}'_k = (\mathbf{X}^T \mathbf{U}')_k^T \mathcal{A} (\mathbf{Y}^T \mathbf{V}')_k \tag{1}$$

where $(\mathbf{X}^T \mathbf{U}')_k \in \mathbb{R}^{\rho}$, $(\mathbf{Y}^T \mathbf{V}')_k \in \mathbb{R}^{\phi}$, and $\mathbf{f}'_k$ denotes the k-th element of the intermediate joint representation. Using Hadamard product and matrix-matrix multiplication, the bilinear attention map $\mathcal{A}$ is defined as:

$$\mathcal{A} = \text{softmax}\Big(\big((\mathbb{1} \cdot \mathbf{p}^T) \circ \mathbf{X}^T\mathbf{U}\big)\mathbf{V}^T\mathbf{Y}\Big) \tag{2}$$

where $\mathbb{1} \in \mathbb{R}^\rho$, $\mathbf{p} \in \mathbb{R}^d$. The $\mathbf{U}$ and $\mathbf{V}$ project the inputs to d-dimensional space. $\mathrm{softmax}$ function is applied element-wisely. Notice that each logit of the $\mathrm{softmax}$ function is the output of low-rank bilinear pooling [4]. The multiple bilinear attention maps can be obtained by using non-sharable parameters $\mathbf{p}_g$, where g denotes the index of glimpses. In order to integrate the multiple bilinear attention maps, we use residual learning of attention as a variant of multimodal residual network [3]. Finally, to perform the two tasks of VizWiz challenge, we attach a two-layer MLP for each task to BAN and train the entire network using a joint loss for them.

3 Experiments

In Table 1, our BAN significantly outperforms the previous methods in accuracy; however, for the answerability, the simpler model, Q+I [2], might be better to predict. The multitask learning for both the accuracy and answerability is significantly helpful for the answerability compared with the learning of answerability standalone.

Model	Accuracy					Answerability	
	Overall	Yes/no	Number	Other	Unans	AP	F1
Q+I [2]	0.137	0.598	0.045	0.142	0.070	**0.717**	0.648
FT [1,2]	0.475	0.669	**0.220**	0.294	0.776	0.561	0.542
VizWiz [1,2]	0.469	0.596	0.210	0.273	0.805	0.605	0.549
BAN (single)	0.516	0.681	0.179	0.315	0.853	0.588	**0.710**
BAN (ensemble)	**0.520**	**0.691**	0.191	**0.316**	**0.862**	-	-

Table 1: Accuracy and answerability for the test split of VizWiz dataset.

4 Conclusions

We propose BAN with the joint loss for the two tasks in VizWiz. BAN considers bilinear interactions among input channels, while low-rank bilinear pooling extracts joint representations for each pair of channels. Although the proposed method outperforms the previous methods regarding overall performance, this is relatively vulnerable to predicting accuracy of numeric types and answerability. We leave this as a future work.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. CVPR (2018)
2. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. CVPR (2018)
3. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. arXiv (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. ICLR (2016)