Improving the audio visual scene-aware dialog system in DSTC7 by using attentional multimodal fusion and MMI objective

Wang Wenbo[1,a)] Zhuang Bairong[1,b)] Takahiro Shinozaki[1,c)]

Abstract: We show our effort for the 7th Dialog System Technology Challenge (DSTC7) Audio Visual Scene-aware dialog (AVSD) track. In our work, we employ the attentional multimodal fusion and Maximum Mutual Information (MMI). The MMI is utilized for the objective function instead of cross entropy loss in the baseline system. Our results show these extensions are useful for improving the performance of the system.

Keywords: Audio Visual Scene-aware dialog, Attentional multimodal fusion, Maximum Mutual Information

1. Introduction

The AVSD track [1], [2] in the Dialog System Technology Challenges (DSTC) workshop aims to use the video information for scene awareness [3] to generate informative system responses in end-to-end dialog systems [4]. In this paper, we report our exploration on the AVSD track by introducing attentional multimodal fusion [5] and replacing the original objective with MMI [6].

2. Models

2.1 Audio Visual Scene-Aware Dialog System

The baseline AVSD system is similar to [7], and has separated encoders: Question Encoder, Multimodal Encoder, and Context Encoder. The question encoding s^q, multimodal encoding s^{mm}, and dialog context encoding s^c are calculated through their corresponding encoders. The decoder is a stacked LSTM having 2 layers. The detailed architecture of the AVSD system is described in [2].

2.2 The Extended AVSD System

The baseline system combines context features s^{m_k} of different modalities based on Naïve fusion [8]. In this paper, we employ the attentional multimodal fusion introduced in [5] instead of the Naïve fusion. The attentional multimodal fusion is computed as:

$$s^{mm} = \tanh(\sum_{k=1}^{K} \beta_k (W_{sk} s^{m_k} + b_{sk})),$$

and the attention weight β_k is obtained from:

$$\beta_k = softmax(v_k),$$

Table 1 Video Scene-aware Dialog Dataset on Charades

	training	validation	test
# of dialogs	6,172	732	733
# of turns	123,480	14,680	14,660
# of words	1,163,969	138,314	138,790

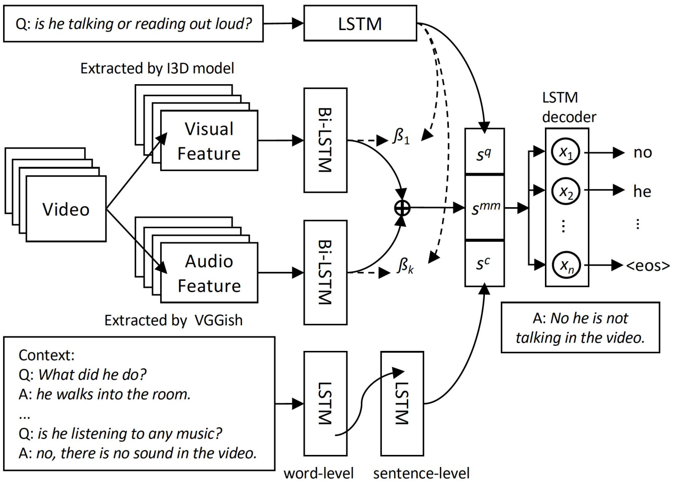

Fig. 1 The architecture of the extended AVSD system with attentional multimodal fusion, where the β denotes the attention weight.

$$v_k = w_b \cdot tanh(W_{aq} s^q + W_{ak} s^{mk} + b_{ak}),$$

where K is the number of modalities. The attention weight β_k enables the decoder network to attend each modality when generating next response word. The architecture of the extended AVSD system is shown in **Fig. 1**.

The cross entropy objective in the baseline system leads to generic and safe responses, since it only takes the source side into the consideration. To promote diversity in the generated responses, we replace it with Maximum Mutual Information (MMI) [6]. In the MMI approach, target T is obtained by Equation (2).

[1] Tokyo Institute of Technology, Tokyo, Japan
[a)] wang.w.ai@m.titech.ac.jp
[b)] zhuang.b.aa@m.titech.ac.jp
[c)] www.ts.ip.titech.ac.jp

Table 2 The results of different methods using visual multimodal features (I3D-RGB and I3D-Flow). In this table, **LSTM** denotes the LSTM unit used for modeling multimodal feature, if no LSTM is specified, there is just a linear transformation before multimodal fusion in the multimodal encoder part. **C** denotes the caption, which is the description of the target video. **AF** denotes the attentional multimodal fusion.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
Baseline(official)	0.273	0.173	0.118	0.084	0.117	0.291	0.766
Baseline(ours)	0.272	0.174	0.118	0.083	0.118	0.292	0.769
+LSTM	0.274	0.174	0.118	0.083	0.117	0.292	0.762
+C	**0.279**	**0.177**	0.12	0.085	0.117	**0.293**	0.77
+LSTM+C	0.277	0.175	0.119	0.084	0.117	**0.293**	0.77
+AF	0.276	**0.177**	**0.122**	**0.087**	0.117	**0.293**	**0.787**
+AF+C	0.278	**0.177**	0.12	0.085	**0.119**	0.291	0.762
+AF+C+LSTM	0.271	0.174	0.119	0.085	0.117	0.291	0.785

Table 3 The results when VGGish audo feature is used in addition to I3D-RGB and I3D-Flow visual features.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
Baseline	0.277	**0.178**	**0.122**	**0.087**	**0.119**	**0.296**	**0.791**
+LSTM	0.277	0.176	0.12	0.085	0.118	0.293	0.77
+C	0.274	0.174	0.119	0.084	0.117	0.292	0.779
+LSTM+C	**0.279**	0.177	0.121	0.085	0.118	0.295	0.77
+AF	0.270	0.172	0.118	0.084	0.116	0.286	0.744
+AF+C	0.271	0.172	0.117	0.083	0.117	0.29	0.759
+AF+C+LSTM	0.276	0.176	0.119	0.084	0.117	0.293	0.766

Table 4 The result of adding MMI objective with different parameter λ using I3D-RGB and I3D-Flow on attentional multimodal fusion system.

λ	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
$\lambda_1 = 0.76$	**0.283**	**0.181**	0.124	**0.089**	0.121	0.296	0.805
$\lambda_2 = 0.46$	0.282	**0.181**	**0.125**	**0.089**	**0.122**	**0.297**	0.819
$\lambda_3 = 0.36$	0.28	0.18	0.124	**0.089**	**0.122**	0.296	**0.821**

$$\hat{T} = \arg\max_{T}(\log p(T|S) - \lambda \log \frac{p(T|S)p(S)}{p(S|T)}), \quad (1)$$

$$= \arg\max_{T}((1-\lambda)\log p(T|S) + \lambda \log p(S|T)), \quad (2)$$

where λ controls the trade-off between the models of $\log P(T|S)$ and $\log P(S|T)$. Since directly optimizing the models following Equation 2 is intractable, we train them independently. More specifically, we first train $p(T|S)$ and $p(S|T)$, where $p(S|T)$ is trained by reversing the QA in the training set. Then, we use $p(T|S)$ to generate an N-best list, and re-rank it by $p(S|T)$.

3. Experimental setup

The data we used is released by the DSTC7 organizer. It is an extension of the existing Charades [9] dataset. **Table 1** summarizes the property of the dataset. The visual features are extracted from the I3D-model including I3D-RGB and I3D-Flow introduced in [10], while the audio features are extracted from the Audio Set VGGish model [11]. When employing MMI objective in the attentional multimodal fusion system, we optimized λ in Equation (2) by the grid search using the validation set.

4. Results

Table 2 shows the results of the attentional multimodal fusion using I3D-RGB and I3D-Flow visual features. The baseline scores we get from our experiment are similar to the officially released scores. By adding the attentional multimodal fusion (AF in the Table) to the system, the performance improves compared with the baseline, which proves this extension is useful. We also find some improvements when adding the caption (C) to the system, which indicates that the scene information provided by the caption helps the system to produce a more reasonable answer. **Table 3** shows the results when the audio feature are added in the system. In this condition, however, the attentional multimodal fusion seems to be useless. The bad performance is probably due to the overdubbed sound, which is not in the original scene [5]. **Table 4** shows the results when the MMI objective is applied with the system using the visual features. The parameter λ is optimized for BLEU1 (λ_1), ROUGE_L (λ_2) and CIDEr (λ_3). Compared with the attentional multimodal fusion system in Table 2 (the line with ”+AF”), the employed MMI makes more improvement on different metrics. Besides, the meaningless answer such as ”I can’t tell ...” reduced from ~350 to ~120, which proves the MMI is also useful for improving the performance of the system.

5. Conclusion

We have investigated extending the Audio Visual Scene-aware Dialogue system in DSTC7 by multimodal attention and MMI objective. By employing attentional multimodal fusion using visual feature, the performance of the system was improved compared with the baseline. The performance was further improved by employing MMI as the objective function, and meaningless answers were reduced from the baseline. Future work includes training the AVSD model in an end-to-end manner. Improving the way of combining multimodal information is also needed.

6. Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 17K20001.

References

[1] Hori, C., Marks, T. K., Parikh, D. and Batra, D.: Video Scene-Aware Dialog Track in DSTC7.

[2] Hori, C., Alamri, H., Wang, J., Wichern, G., Hori, T., Cherian, A., Marks, T., Cartillier, V., Lopes, R., Das, A. et al.: End-to-End Audio Visual Scene-Aware Dialog using Multimodal Attention-Based Video Features (2018).

[3] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C. and Parikh, D.: VQA: Visual question answering, *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433 (2015).

[4] Chen, H., Liu, X., Yin, D. and Tang, J.: A survey on dialogue systems: Recent advances and new frontiers, *ACM SIGKDD Explorations Newsletter*, Vol. 19, No. 2, pp. 25–35 (2017).

[5] Hori, C., Hori, T., Lee, T.-Y., Zhang, Z., Harsham, B., Hershey, J. R., Marks, T. K. and Sumi, K.: Attention-based multimodal fusion for video description, *Computer Vision (ICCV), 2017 IEEE International Conference on*, IEEE, pp. 4203–4212 (2017).

[6] Li, J., Galley, M., Brockett, C., Gao, J. and Dolan, B.: A diversity-promoting objective function for neural conversation models, *arXiv preprint arXiv:1510.03055* (2015).

[7] Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D. and Batra, D.: Visual dialog, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2 (2017).

[8] Yu, H., Wang, J., Huang, Z., Yang, Y. and Xu, W.: Video paragraph captioning using hierarchical recurrent neural networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4584–4593 (2016).

[9] Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I. and Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding, *European Conference on Computer Vision*, Springer, pp. 510–526 (2016).

[10] Carreira, J. and Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset, *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, IEEE, pp. 4724–4733 (2017).

[11] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B. et al.: CNN architectures for large-scale audio classification, *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, IEEE, pp. 131–135 (2017).