Investigation of Attention-Based Multimodal Fusion and Maximum Mutual Information Objective for DSTC7 Track3

Bairong Zhuang, Wenbo Wang, Takahiro Shinozaki
Tokyo Institute of Technology
Tokyo, Japan
http://www.ts.ip.titech.ac.jp

Abstract

In this paper, we show our effort on the Audio Visual Scene-aware dialog (AVSD) task which is proposed in DSTC7. We investigate the effectiveness of different modality fusion methods as well as the different input modalities. We also employ the Maximum Mutual Information (MMI) objective to replace the original objective function in the AVSD system. In our experiments, the proposed simplified attention-based multimodal fusion method outperforms the prior work slightly. Besides, the final system which uses MMI as the new objective obtains a 6.6% relative improvement over the baseline system on BLEU-1 metric.

Index Terms: Audio Visual Scene-Aware Dialog, Attention-Based Multimodal Fusion, Maximum Mutual Information, DSTC7

Introduction

With the development of deep learning technology, dialog systems have attracted much attention (Chen et al. 2017). There are many applications based on dialog systems, such as robot assistants (Nakano et al. 2006) and technical support services (Wen et al. 2017; Eric et al. 2017). However, the information obtained by the dialog systems had been limited to users' text inputs, making it difficult to understand the specific scenes, and thus the systems tend to generate low-relevant responses. To this end, some researchers have been trying to introduce scene-aware technology and provide multimodal information inputs for the dialog systems (Das et al. 2017; Antol et al. 2015). The traditional system requires us to provide intermediate annotations of different modules on such multiple input issues, which is time-consuming and expensive. With the development of end-to-end dialog technology in recent years, only labeled data is needed to feed into a differentiable end-to-end system which makes us able to combine different modules in a single network (Hori et al. 2018).

Focusing on the end-to-end Dialog System with multimodal input features, the Dialog System Technology Challenges (DSTC) workshop proposes the Audio Visual Scene-aware dialog (AVSD) track (Huda et al. 2019) in 2018, aiming to use the video information for scene awareness and generate informative system responses.

In this paper, we explore attention-based multimodal fusion (Hori et al. 2017) and MMI objective (Li et al. 2015). In the experiment, we also investigate the effectiveness of using captions and LSTM in the multimodel encoder module, which is already implemented but not enabled in the official released codes. From our investigation, we find that the systems using LSTM and caption usually outperform the systems that do not use both of them. Our biggest improvement on BLEU-1 metric is 6.6% compared with the baseline system. The improvements mainly come from the MMI objective, and the caption information is also helpful.

The remaining of this paper is organized as follows. We first give a brief description of end-to-end modeling and the baseline AVSD system structure. Then, we explain the attention-based multimodal fusion and MMI objective used in the experiment. We then explain the experimental setup, which includes the detail of the dataset we use, our training objective and training hyper-parameter setting, and evaluation method and metrics. After that, we present analyze the results obtained from our experiment. Finally, we give the conclusion and future work.

Baseline and Extended Models

Basic Approach

The effectiveness of using encoder-decoder model to model sequence data has been proved in various tasks (Sutskever, Vinyals, and Le 2014). In the aspect of using such a model for dialog modeling, the encoder takes the word sequence $X = \{x_1, x_2, \cdots, x_n\}$ as input and summarizes it into a fixed size vector. In the decoding phase, the decoder accepts this context vector as the initial hidden state, then generates the response word by word in an auto-regressive fashion:

$$y_m = \arg\max_{y_m \in V} P(y_m | y_1, y_2, \cdots, y_{m-1}, X), \qquad (1)$$

where V denotes the vocabulary, $y_m \in Y$ denotes the word generated in the m-th step. Mathematically, the encoder-decoder model for dialog modeling can be seen as the conditional probability:

$$\hat{Y} = \arg\max_{Y \in V} P(Y|X). \qquad (2)$$

Practically, the encoder-decoder model is trained by maximizing the likelihood of the target response Y^* using cross entropy loss. In inference time, beam search is usually used.

Audio Visual Scene-aware Dialog System

To use encoder-decoder model for the AVSD system, the vanilla system is extended to model different modalities using respective encoders. Formally, given input data $X = \{(\mathbf{q_1}, \mathbf{av_1}, \mathbf{c_1}), (\mathbf{q_2}, \mathbf{av_2}, \mathbf{c_2}), \cdots, (\mathbf{q_n}, \mathbf{av_n}, \mathbf{c_n})\}$, the AVSD model produces corresponding answer $\mathbf{y_n}$ for each question $\mathbf{q_n}$, where the $\mathbf{q_n}$,$\mathbf{av_n}$, $\mathbf{c_n}$ denotes the word sequence, the audio/visual features, and the dialog context for n-th question, respectively. The architecture of the baseline(Huda et al. 2019) (and our extended) AVSD system is shown in Figure 1.

Question Encoder In the AVSD model, the baseline system uses an RNN as the question encoder. Formally, for a question $q = \{q^1, q^2, \dots, q^n\}$ with n words, the corresponding context vector is computed by

$$s^q = \mathrm{RNN}(\mathrm{embed}(q^1), \mathrm{embed}(q^2), \dots, \mathrm{embed}(q^n)),$$

where the $\mathrm{embed}(\cdot)$ is the embedding layer which translates the word index into a dense representation. The baseline system uses an LSTM as the RNN unit.

Multimodal Encoder For a video v with K different modalities, each modality is represented as $\boldsymbol{\{v^1, v^2, \cdots, v^K\}}$. The visual/audio multimodal features are first transformed into a lower dimensional space $\mathbb{R}^p$, then the context vector for k-th modality s_k^v is computed by the following 2 ways: 1) employ an RNN to summarize the temporal multimodal features 2) take the average over feature vectors of each modality. The first method can be formalized as:

$$s_k^v = \mathrm{RNN}_k(W_{mk}(v^k) + b_{mk}), \tag{3}$$

where the $\boldsymbol{W_{mk}}$ and $\boldsymbol{b_{mk}}$ are learnable parameters. The baseline system uses a bi-directional LSTM as the RNN unit. To combine the context vector of different modalities, prior work (Yu et al. 2016) proposed a simple modality fusion method to compute mixture representation $\boldsymbol{s^{mm}}$ with separate linear transformation (with learnable parameters $\boldsymbol{W_{sk}, b_{sk}}$) applied to each modality:

$$s^{mm} = \tanh(\sum_{k=1}^{K}(W_{sk}s_k^v + b_{sk})), \tag{4}$$

in which the context vectors of different modalities are firstly transformed with the separate linear transformation and activated by a $\tanh$ function. This modality fusionmethod is denoted as Naïve Fusion (Hori et al. 2017). The baseline implements this approach.

Context Encoder To efficiently encode dialog context c, the baseline system uses hierarchical LSTMs (Serban et al. 2016) in which one LSTM encodes each question-answer pair in the word level, and another LSTM summarizes the question-answer encoding in the sentence level. Specifically, given the dialog context $c = \{c^1, c^2, \dots, c^{m_n}\}$ with m_n utterances, each dialog context is summarized by the RNN with corresponding context vector

$$s^c = \mathrm{RNN}(s^{c^1}, s^{c^2}, \dots, s^{c^{m_n}}), \tag{5}$$

where $\boldsymbol{s^{c^m}}$ denotes the context vector by summarizing m-th utterance which is produced by another RNN in the word level. The baseline system uses a stacked 2-layer LSTM as the RNN units for the word level RNN and an LSTM for the sentence-level RNN.

Decoder The decoder used in the baseline system is slightly different from the one that usually used in the seq2seq model. The standard decoder used in the seq2seq model takes the output of the previous step as the input for the current step, which generates output sequence in an auto-regressive manner. The baseline system takes not only the previous step's output but also the context vector generated by the encoder in every decoding step,

$$s^{d_n} = \mathrm{RNN}([\mathrm{embed}(y_{n-1}); s^e]). \tag{6}$$

The baseline system uses a stacked 2-layer LSTM as the RNN unit. The answer is computed by

$$p(y_n|y_{<n-1}, s^e, X) = \mathrm{softmax}(W_{do}s^{d_n} + b_{do}), \tag{7}$$

where the $y_{<n-1}$ denotes the output produced by the previous $n-1$ steps, $\boldsymbol{s^e} = [\boldsymbol{s^q}; \boldsymbol{s^{mm}}; \boldsymbol{s^c}]$ denotes the concatenation of question encoding $\mathbf{s^q}$, multimodal feature encoding $\boldsymbol{s^{mm}}$, and dialog context encoding $\boldsymbol{s^c}$. $\boldsymbol{W_{do}}$ and $\boldsymbol{b_{do}}$ are learnable parameters.

Our Extension to Baseline System

We extend the baseline system in the following two ways. Firstly, we implement attention-based multimodal fusion method (Hori et al. 2017). We then examine this method and propose a slightly modified version. Besides, we use Maximum Mutual Information objective (Li et al. 2015) as the new objective for the baseline system.

Attention-Based Multimodal Fusion In the baseline model, the Naïve Fusion method is used to combine the context vector of each modality. Previous work (Hori et al. 2017) extend the Naïve Fusion by weighing the context vector of each modality instead of treating each modality equally:

$$s^{mm} = \tanh(\sum_{k=1}^{K}\beta_k(W_{sk}s_k^v + b_{sk})). \tag{8}$$

This modality fusion method is denoted as attention-based multimodal fusion in their work. Similar to the temporal attention, the attention weights β_k are obtained by

$$\beta_k = \mathrm{softmax}(v_k), \tag{9}$$

where v_k can be seen as the attention score computed by:

$$v_k = w_b^T \cdot \tanh(W_{aq}s^q + W_{ak}s_k^v + b_{ak}). \tag{10}$$

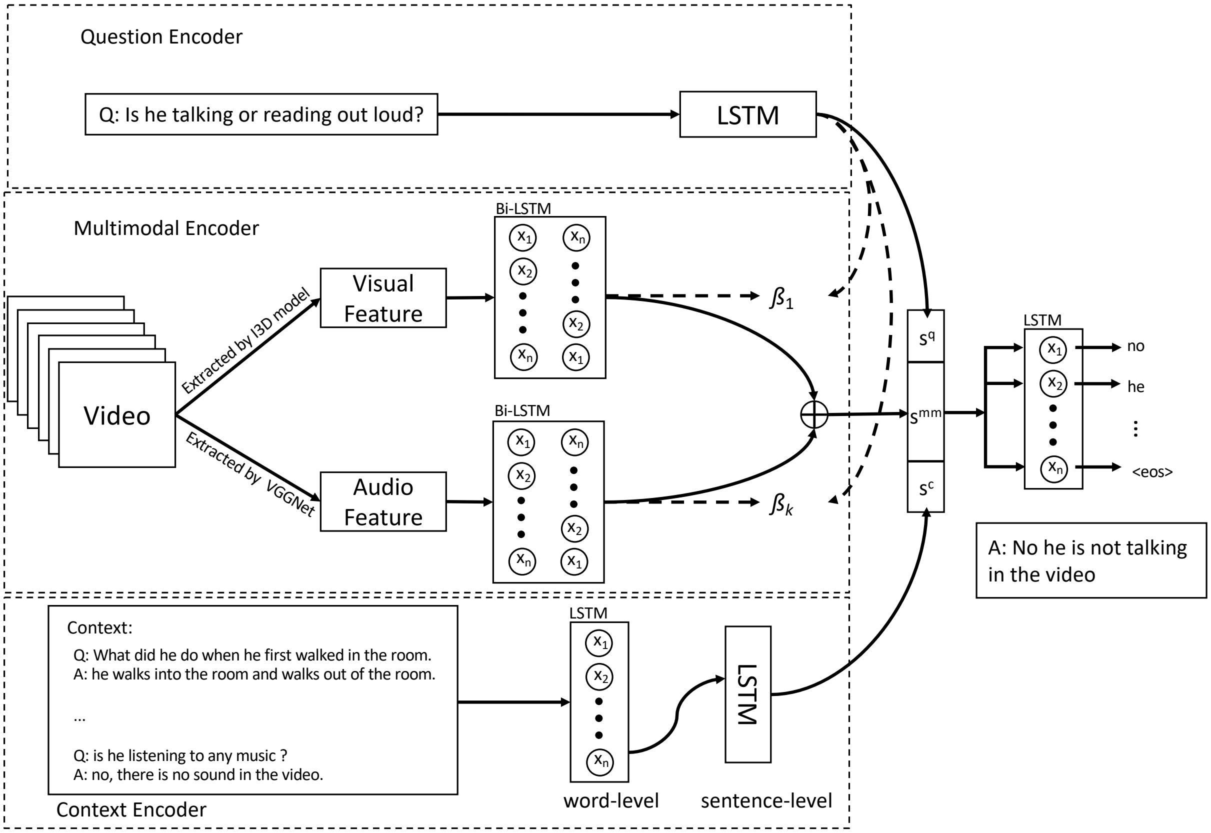
Question Encoder
Q: Is he talking or reading out loud?
LSTM
Multimodal Encoder
Video
Extracted by I3D model
Extracted by VGGNet
Visual Feature
Audio Feature
Bi-LSTM
Bi-LSTM
x_1
x_2
x_n
$ß_1$
$ß_k$
s^q
s^{mm}
s^c
LSTM
no
he
<eos>
A: No he is not talking in the video
Context:
Q: What did he do when he first walked in the room.
A: he walks into the room and walks out of the room.
...
Q: is he listening to any music ?
A: no, there is no sound in the video.
LSTM
LSTM
word-level
sentence-level
Context Encoder

Figure 1: Illustration of model architecture of the baseline and our extended AVSD systems. The Encoder includes three parts: Question Encoder, Multimodal Encoder, and Context Encoder. Question Encoder encodes the question sentence. The Multimodal Encoder takes visual/audio features as input, an LSTM can be chosen to model the temporal information in the multimodal feature. A hierarchical LSTM is used as the Context Encoder which encodes the context in word-level and sentence-level. The vector representations produced by these three encoders are concatenated and fed into the decoder to produce the answer. The multimodal attention weights β are computed by taking current question encoding and modality encoding into consideration. The Naïve Fusion used in the baseline system can be thought as a special case of our extended version with modality weight β fixed to 1 (Hori et al. 2018).

The attention weights are computed by taking current question encoding $\boldsymbol{s^q}$ and each input modality $\boldsymbol{s_k^v}$ into consideration, which makes the modality fusion process more flexible. We implement this attention-based multimodal fusion as the extension of the baseline. In the attention-based multimodal fusion method, it transforms each context vector into attention space using separate weight matrices. To examine the importance of computing modality attention score using separate weight metrics, we simplify this method by computing attention score v_k for k-th modality using the shared weight matrix:

$$v_k = w_b^T \cdot \tanh(W_{aq}s^q + W_{am}s_k^v + b_{am}), \tag{11}$$

where $\boldsymbol{W_{am}}$ and $\boldsymbol{b_{am}}$ are learnable parameters, $\boldsymbol{s^q}$ and $\boldsymbol{s^{mk}}$ are context vectors for question and multimodal feature respectively. We denote this modality fusion method as simplified attention-based multimodal fusion.

Maximum Mutual Information objective The Maximum Mutual Information (MMI) objective has shown great potential in the dialog modeling tasks (Galley et al. 2017). The original objective function in the baseline model is the log-likelihood of the target T given the source S, which in the decoding process, we can formalize it as a statistical decision problem

$$\hat{T} = \arg\max_T(\log p(T|S)). \tag{12}$$

However, in the baseline system, this objective often leads to generic and safe responses, since it only takes the source side into consideration. We implement the MMI objective to promote diversity in the generated response [1]. The MMI objective is employed to replace the original objective function. In MMI, parameters are chosen to maximize mutual

[1] Although correctness may be more important than the diversity for the AVSD task, we tried it anyway.

information between source S and target T:

$$\log\frac{p(S,T)}{p(S)p(T)} = \log p(T|S) - \log p(T). \tag{13}$$

This let the model avoid producing meaningless answer with high log-likelihood. Then the MMI objective can be written as follow:

$$\hat{T} = \arg\max_{T}(\log p(T|S) - \log p(T)). \tag{14}$$

Following the literature (Li et al. 2015) which generalize the MMI objective with hyper-parameter λ that controls the tradeoff between $P(T|S)$ model and $p(T)$ model, the formula can be written as

$$\hat{T} = \arg\max_{T}(\log p(T|S) - \lambda \log p(T)). \tag{15}$$

By Bayesian formula that:

$$p(T) = \frac{p(T|S)p(S)}{p(S|T)}, \tag{16}$$

the above objective function can be rewritten as

$$\begin{aligned} \hat{T} &= \arg\max_{T}(\log p(T|S) - \lambda \log\frac{p(T|S)p(S)}{p(S|T)}) && (17) \\ &= \arg\max_{T}((1-\lambda)\log p(T|S) + \lambda \log p(S|T)). && (18) \end{aligned}$$

Note that formula 18 is intractable which cannot be directly optimized (Li et al. 2015). Alternatively, the $p(S|T)$ model and the $p(T|S)$ model are trained separately, only applied MMI in the decoding phase.

Experimental Setup

Dataset The data we used is released by DSTC7 organizer which is an extension based on an existing dataset, Charades (Sigurdsson et al. 2016). This dataset containing 11,848 videos, which is split into 7985 for training, 1863 for validation, and 2000 for testing, respectively. The 7043 videos in the training set and all 1863 videos from the validation set are selected as the video data for the dataset used in the DSTC7.

The visual feature is extracted from the "Mix_5c" layer of the I3D-model (Carreira and Zisserman 2017), then preprocessed into zero mean and unit norm, according to the description from the organizer. Besides, I3D-RGB (I3D feature computed on a stack of 16 video frames) and I3D-Flow (I3D feature computed on a stack of frames of optical flow fields) are treated as two separate modalities, which is input to the multimodal encoder (Hori et al. 2018). The audio feature is extracted from the Audio Set (Hershey et al. 2017) VGGish model (VGGNet without the last group of convolutional/pooling layers), this model operates on 0.96s log Mel spectrogram patches extracted from 16kHz audio, then outputs a 128-dimensional embedding vector. The input frames of VGGish network are with 50% overlap (Hori et al. 2018). Each video in the training set includes 10 rounds of QA, a caption, and a summarization for the video from the questioner. The statistics of the dialog part are listed as table 1.

Table 1: Statistics of dialog

	Train	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

Table 2: Hyper-parameter setting

Hyper-parameter	Value
Embedding size	100
# of Question Encoder Layer	1
# of Multimodal Encoder Layer	1
# of Context Encoder Layer (word, sent.)	(2,1)
Encoder projection size (modality-wise)	(512,512,64)
Encoder hidden size (modality-wise)	(256,256,128)
Attention size	128
# of Decoder Layer	2
Decoder hidden size	128
Batch size	64
Optimizer	Adam
Dropout	0.5

Training To train the model on objective $(1-\lambda)\log p(T|S) + \lambda \log p(S|T)$, direct optimization is intractable. Instead, we train the $p(T|S)$ model first and reverse the QA in the training set for training the $p(S|T)$ model. Then, we used the $p(T|S)$ to generate an N-best list, which is further reranked by the $p(S|T)$ model. We search the λ through the grid search and select the λ according to the performance on the validation set. The model architecture for $p(T|S)$ and $p(S|T)$ model are the same (the only difference is we train the two systems with reversed QA-pair). We use the same hyper-parameter to train the system in all the experiments. The hyper-parameter setting for most of our experiments is summarized in table 2

Evaluation We evaluated our results using the evaluation code in the MS COCO caption (Chen et al. 2015) GitHub repository, which provides objective measures such as BLEU, METEOR, ROUGE_L, and CIDEr. The results in the next section except the MMI experiments were decoded in the same parameter setting, which the beam size was set to 5 and the length penalty factor was set to 1.0. In the experiments for examining Maximum Mutual Information objective, we set the beam size from 5 to 20 in the decoding time so that we can search large enough space to generate more diversity answers. The baseline model was implemented in PyTorch and the experiments were run on a server equipped with NVIDIA GeForce GTX1080.

Results

Investigation on Official Released System Table 3 shows our investigation of the baseline model. According to the official score sheet, only V and V+A are denoted as the baseline. Since the other systems shown in the Table 3

Table 3: Results of different setting in the system released by the DSTC7 organizer. In the table, **V** denotes the visual feature, **A** denotes the audio feature, **C** denotes the caption, which is the description of target video. **LSTM** denotes the LSTM used for modeling multimodal feature, if no LSTM is specified, there is just a linear transformation before multimodal fusion in the multimodal encoder module.

Methods	MM feature types	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
V (baseline)	i3d	0.272	0.174	0.118	0.083	0.118	0.292	0.769
V+C	i3d	**0.279**	0.177	0.120	0.085	0.117	0.293	0.770
V+LSTM	i3d	0.274	0.174	0.118	0.083	0.117	0.292	0.762
V+C+LSTM	i3d	0.277	0.175	0.119	0.084	0.117	0.293	0.770
V+A (baseline)	i3d&vggish	0.277	**0.178**	**0.122**	**0.087**	**0.119**	**0.296**	**0.791**
V+A+C	i3d&vggish	0.274	0.174	0.119	0.084	0.117	0.292	0.779
V+A+LSTM	i3d&vggish	0.277	0.176	0.120	0.085	0.118	0.293	0.770
V+A+C+LSTM	i3d&vggish	**0.279**	0.177	0.121	0.085	0.118	0.295	0.770

(e.g. use LSTM in the multimodal encoder module) are already implemented in the baseline released by the organizer, we also discuss them in this section. Firstly, we validate the baseline system with the visual feature. The BLEU score we obtained is similar to the officially released score. The performance improves on BLEU, ROUGE_L, and CIDEr when we added captions i.e. the (V+C), which indicates that the scene information provided by the captions helps the system to produce a more meaningful answer. The performance is slightly drops on METEOR metric. To examine the effectiveness of using LSTM as the multimodal feature summarizer, we employed an LSTM in the multimodal encoder module with/without caption information. The BLEU score of the V+LSTM and the V+C+LSTM improves when compared with the baseline. Note that the performance on other metric e.g. METEOR are still lower than the baseline. Furthermore, the performance of the system that leverages caption information outperforms the system that does not use caption information.

Besides, we investigate the role of the audio feature in the AVSD system by adding the audio feature system. The performance of V+A system improves when compared with the system that only uses the visual feature. Further, we examine the role of the captions in this visual-audio system. The performance of system V+A+C does not improve compared to the V+A system, which is different from the system that using only visual feature and captions (V+C). Also, we attempt to use LSTM in the multimodal encoder module to help the system to learn the temporal information in the multimodal feature. However, the performance gets slightly worse. Lastly, we use LSTM in the multimodal encoder module and feed captions into the system at the same time (i.e. V+A+C+LSTM). The BLEU-1 score outperforms above-mentioned systems.

According to the experiments, the AVSD system could be improved by adding captions to the system. The system using both LSTM and captions outperforms the system that does not use them. The best system in the view of BLEU-1 score is the one that uses both visual and audio feature with captions and the LSTM in the multimodal encoder module (V+A+C+LSTM).

Investigation on Attention Based Multimodal Fusion Methods Table 4 shows the experimental results of our investigation of fusion methods for multimodal features. By using attention-based multimodal fusion we implemented by ourselves, the ROUGE_L and CIDEr score (i.e. the V+AF) improves and another improvement can be made by adding captions. However, note that this system still can not outperform the system that only uses visual feature and captions (V+C). Further, the BLEU score drops dramatically when we use an LSTM in the multimodal encoder module (V+C+LSTM+AF). One possible explanation for this is that the unnormalized audio feature may mislead the computation for modality fusion method.

When audio feature is used in the model, either adding the captions (i.e. the V+A+C+AF or employing an LSTM (i.e. V+A+C+LSTM+AF) in the multimodal encoder module can generally improve the performance of AVSD system. Although the overall performance is still at a relatively low level compared with the one that just uses the visual feature.

Also, the performance of the simplified attention-based multimodal fusion on visual feature (V+SAF) gets slightly better compared with the one using attention-based multimodal fusion. The performance drops when we add captions which is different from the previous experiments. We also observed that when adding the audio features in the AVSD system with attention-based modality fusion methods, both of the attention-based multimodal fusion method and the simplified attention-based multimodal fusion method get worse performances, while the baseline system with Naïve Fusion obtains a better performance.

Investigation on Maximum Mutual Information Objective Table 5 shows the experiment results of applying MMI objective to the best systems in the previous experiments, i.e. V+C, V+A+C, V+C+AF, and V+SAF. Another significant peroformance improvements could be obtained by adding MMI objective as the new objective function in the previous best system. The baseline system which uses caption information in the training set and MMI objective as the new objective is the best system, which shows the effectiveness of the MMI objective.

Table 4: Results of the system performance between attention based modality fusion methods. In the table, **AF** denotes the attention based multimodal fusion, **SAF** denotes the simplified attention based modality fusion

Methods	MM feature types	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
V+AF	i3d	0.276	**0.177**	**0.122**	**0.087**	0.117	**0.293**	**0.787**
V+C+AF	i3d	**0.278**	**0.177**	0.120	0.085	**0.119**	0.291	0.762
V+C+LSTM+AF	i3d	0.271	0.174	0.119	0.085	0.117	0.291	0.785
V+A +AF	i3d&vggish	0.270	0.172	0.118	0.084	0.116	0.286	0.744
V+A+C+AF	i3d&vggish	0.271	0.172	0.117	0.083	0.117	0.290	0.759
V+A+C+LSTM+AF	i3d&vggish	0.276	0.176	0.119	0.084	0.117	**0.293**	0.766
V+SAF	i3d	0.277	0.176	0.120	0.085	0.118	0.290	0.765
V+C+SAF	i3d	0.274	0.174	0.119	0.085	0.118	0.291	0.775
V+C+LSTM+SAF	i3d	0.276	0.174	0.119	0.084	0.118	0.290	0.755
V+A+SAF	i3d&vggish	0.272	0.173	0.118	0.083	0.116	0.288	0.754
V+A+C+SAF	i3d&vggish	0.276	0.175	0.119	0.084	0.116	0.292	0.765
V+A+C+LSTM+SAF	i3d&vggish	0.271	0.171	0.117	0.083	0.115	0.287	0.747

Table 5: Results of the system performance by adding MMI in the best setting of Table 3 and Table 4

Methods	MM feature types	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
V+C+MMI	i3d	**0.290**	**0.184**	**0.125**	**0.089**	0.121	**0.298**	0.800
V+A+C+LSTM+MMI	i3d&vggish	0.286	0.182	0.124	0.088	0.120	0.297	0.789
V+C+AF+MMI	i3d	0.287	0.183	**0.125**	**0.089**	**0.122**	0.293	0.795
V+SAF+MMI	i3d	0.283	0.181	0.124	**0.089**	0.121	0.296	**0.805**

Conclusion

In this paper, we described our effort for the Audio Visual Scene-aware dialog task of DSTC7. We investigated the baseline system released by the DSTC7 organizer, the modality fusion method proposed by (Hori et al. 2017), and the simplified version of the modality fusion method we proposed in this work. Also, we examined the Maximum Mutual Information objective which was proposed by (Li et al. 2015), and got an improvement over the baseline system.

Future work includes investigating a way to training the AVSD model in an end-to-end manner. Since in the current system, we use the visual and audio feature extracted from the pre-trained model. Exploring a new way to leverage and incorporate multimodal information in AVSD system is also needed.

Acknowledge

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

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

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

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Chen, H.; Liu, X.; Yin, D.; and Tang, J. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter* 19(2):25–35.

Das, A.; Kottur, S.; Moura, J. M.; Lee, S.; and Batra, D. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. *arXiv preprint arXiv:1703.06585*.

Eric, M.; Krishnan, L.; Charette, F.; and Manning, C. D. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 37–49.

Galley, M.; Brockett, C.; Dolan, B.; and Gao, J. 2017. The msr-nlp system at dialog system technology challenges 6. In *Proceedings of the 6th Dialog System Technology Challenges (DSTC6) Workshop*.

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. 2017. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, 131–135. IEEE.

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

Hori, C.; Alamri, H.; Wang, J.; Winchern, G.; Hori, T.;

Cherian, A.; Marks, T. K.; Cartillier, V.; Lopes, R. G.; Das, A.; et al. 2018. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. *arXiv preprint arXiv:1806.08409*.

Huda, A.; Chiori, H.; Tim K., M.; Dhruv, B.; and Devi, P. 2019. Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7. In *Dialog System Technology Challenge 7 at AAAI2019*.

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

Nakano, M.; Hoshino, A.; Takeuchi, J.; Hasegawa, Y.; Torii, T.; Nakadai, K.; Kato, K.; and Tsujino, H. 2006. A robot that can engage in both task-oriented and non-task-oriented dialogues. In *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*, 404–411. IEEE.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, 3776–3784.

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

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Wen, T.; Vandyke, D.; Mrkšíc, N.; Gašíc, M.; Rojas-Barahona, L.; Su, P.; Ultes, S.; and Young, S. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017-Proceedings of Conference*, volume 1, 438–449.

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