End-to-End Multimodal Dialog Systems with Hierarchical Multimodal Attention on Video Features

Hung Le[1], Steven C.H. Hoi[1], Doyen Sahoo[1], Nancy F. Chen[2]
[1]Singapore Management University, Singapore
[2]Institute for Infocomm Research (I2R), Singapore
hungle.2018@phdis.smu.edu.sg, {chhoi,doyens}@smu.edu.sg, nfychen@i2r.a-star.edu.sg

Abstract

We present our work on the Dialog System Technology Challenges 7 (DSTC7). We participated in Track 3, which evaluated how dialog systems understand video scenes and response to users about the video visual and audio content. Our system is built upon the baseline system (Hori et al. 2018) with changes adopted similarly to (Anderson et al. 2018). The model utilizes different types of attentions on video caption and the video audio and visual input that contribute to the improved evaluation results. We also applied a nonlinear feature fusioning of the visual and audio features to improve the results further. Our proposed model showed improved performance in terms of both objective evaluation and human rating, surpassing the performance of the baseline.

Introduction

The Dialog System Technology Challenge 7 (DSTC7) proposed a track that focused on multi-modal dialog systems. Arising from the related tasks in visual Question-Answering (VQA) (Antol et al. 2015)(Goyal et al. 2017), image captioning (Vinyals et al. 2015)(Xu et al. 2015), video captioning (Hori et al. 2017)(Li et al. 2018), and visual dialogs (Das et al. 2017a)(Das et al. 2017b), the track offered an interesting dialog setting that integrates not only visual features but also audio features from video input. Compared to visual dialog (Das et al. 2017b), the proposed task in this track consists of more modalities with much larger feature space. Our entries to Track 3 of DSTC7 follow this setting and was trained exclusively from the provided official data and did not utilize any external data.

Our approach for this track is summarized in this paper. Our two entries to this track were built upon the baseline model (Hori et al. 2018), and exploited different attention mechanisms on question features, caption features, and visual and audio features of the input video. The attention strategies are adopted similarly to (Anderson et al. 2018), including question-guided attention techniques on caption and video features. In the VQA setting, the usage of these attentions was shown to improve the accuracy in selecting the correct answers. In the context of DSTC7, we aims to explore how these attention mechanisms could be utilized in a dialog context to generate system responses rather than in a QA setting. The provision of both visual and audio features also allowed us to explore fusion techniques to combine these features better than the baseline. Our experiments showed that using linear layer with ReLU activation and Hadamard-product helped to fuse the features and increased the results significantly.

In this report we detailed our proposed model and different parameter settings of the model. We also provided some qualitative study to analyze how our system improved the quality of the responses from the baseline model.

End-to-End Multimodal Dialog System

This section details several changes we made from the baseline approach (Hori et al. 2018). The overview of the model can be seen in Figure 1.

Gated Recurrent Unit

Instead of using Long short-term memory (LSTM) as the unit module for the recurrent network, we replaced LSTM with Gated Recurrent Unit (GRU) in the encoders (for question and dialog history). GRUs have shown to achieve superior performance at affordable computational cost (Cho et al. 2014). We describe here in mathematical details of the GRU for complete notation of the proposed model. Given a sequence of input words S, in each encoding step n, the GRU will recurrently process the respective input s_n and the previous hidden state h_{n-1}. For simplicity, we denote s_n as both the real word as well as the representation vector of the word using an embedding matrix or one-hot representation. We denote the embedding dimension as V. The hidden state h_n for each encoding step n is given by:

$$r_n = \sigma(I_r s_n + H_r h_{n-1}), \tag{1}$$

$$u_n = \sigma(I_u s_n + H_u h_{n-1}), \tag{2}$$

$$\overline{h}_n = \tanh(I s_n + H(r_n \cdot h_{n-1})), \tag{3}$$

$$h_n = (1 - u_n) \cdot h_{n-1} + u_t \cdot \overline{h}_n \tag{4}$$

where σ is the logistic sigmoid, $\cdot$ represents the elementwise scalar product between vectors, $I, I_u, I_r \in \mathbb{R}^{d_h \times V}$ and $H, H_r, H_u \in \mathbb{R}^{d_h \times d_h}$. The I matrices encode the word s_n while the H matrices are used to retain or forget the information in h_{n-1}. Hence, r_n denotes the *reset gate*, u_n the *update gate*, $\overline{h}_n$ the *candidate update*, and h_n the *final update*.

The reset gate and update gate are computed in parallel. Provided the current word s_n, if it is learned to forget information of the previous sequence h_{n-1}, the elements of r_n will be closer to 0. The update gate u_n judges whether the current word contains relevant information that should be stored in h_n. In the final update, if the elements of u_n are close to 0, the network keeps the last recurrent state h_{n-1}. The gating behavior in GRU showed to provide robustness to noise in the source sequence.

At each dialog turn t, for each question Q_t, the question encoder reads the words of the questions sequentially and updates its hidden state according to:

$$h_{t,n}^{qes} = GRU_{qEnc}(h_{t,n-1}^{qes}, s_{t,n}), n = 1, ..., N_t^{qes} \quad (5)$$

To encode the dialog history, each question and answer for each dialog turn $1, ..., t-1$ is encoded by a separate encoder.

$$h_{t,n}^{qa} = GRU_{qaEnc}(h_{t,n-1}^{qa}, s_{t,n}), n = 1, ..., N_t^{qa} \quad (6)$$

A separate GRU takes as input the sequence of past question and answer representations $Q_1, A_1, ..., Q_{t-1}, A_{t-1}$ and computes the sequence of dialog-turn recurrent states to summarize the dialog up to turn t into h_t^{his}. For all encoders, we initialized the hidden states to zero.

$$h_{t,0}^{qes} = 0 \quad (7)$$

$$h_{t,0}^{qa} = 0 \quad (8)$$

$$h_0^{his} = 0 \quad (9)$$

Caption Encoder

Instead of concatenating the video caption as the first turn in the dialog history like in the baseline (Hori et al. 2018), we decided to use a separate encoder to encode the video caption. For each dialog, a GRU encoder reads the words of the caption of the respective video input sequentially and updates its hidden states:

$$h_{t,n}^{cap} = GRU_{capEnc}(h_{t,n-1}^{cap}, s_{t,n}), n = 1, ..., N_t^{cap} \quad (10)$$

We also initialized the hidden state $h_{t,0}^{cap} = 0$.

Question Self-Attention

We added a self-attention mechanism in question encoder. Specifically, in each dialogue turn, the model attends over all positions in the question sequence, each represented by the question encoder hidden state $h_n^{qes}(n = 1, ..., N^{qes})$. The set of all question hidden states h^{qes} are passed through two convolutional layers with kernel size 1 and ReLU and softmax activation. The result scalar attention α_n^{qes} is associated with the position n^{th} in the question.

$$\alpha^{qes} = softmax(Conv(ReLU(Conv(h^{qes})))) \quad (11)$$

$$\hat{h}^{qes} = \sum_{n=1}^{N^{qes}} \alpha_n^{qes} h_n^{qes} \quad (12)$$

The quetion hidden states are weighted by the softmax result and sum to obtain a single vector $\hat{h}^{qes}$ representing the attended question features q.

Question-to-Multimodal Attention

We extended the baseline multimodal attention (Hori et al. 2018) by implementing a question-guided attention mechanism commonly used in many VQA models (Teney et al. 2017)(Anderson et al. 2018).The attention mechanism is used to direct the model to specific input feature sequences in each modality k (input sequence $x_k = x_{k1}, ..., x_{kL}$ for $k = 1, ..., K$). The number of modalities is denoted by K and the number of feature sequences is L. First, both question features q and modality feature x_{kl} are passed through separate linear layers with ReLU activation to project them to the same dimensional space D_k. For each modality $k = 1, ..., K$ and $l = 1, .., L$:

$$\tilde{q}_k = ReLU(W_{kq}q + b_{kq}) \quad (13)$$

$$\tilde{x}_{kl} = ReLU(W_{kx}x_{kl} + b_{kx}) \quad (14)$$

where $W_{kq} \in \mathbb{R}^{D_k \times d_q}$, and $W_{kx} \in \mathbb{R}^{D_k \times d_k}$. The question features is then expanded to have the same sequential dimension L as the modality feature $\tilde{q}_k^{exp} \in \mathbb{R}^{L \times D_k}$ and we then use Hadamard product to create a feature vector f_k to jointly combine question and modality features . The vector is then passed through two convolutional layers with kernel size 1 and ReLU and softmax activation to obtain a scalar attention weight α_{kl} associated with input sequence x_{kl}.

$$f_k = \tilde{x}_k \cdot \tilde{q}_k^{exp} \quad (15)$$

$$\alpha_k = softmax(Conv(ReLU(Conv(f_k)))) \quad (16)$$

$$\hat{x}_k = \sum_{l=1}^{L} \alpha_{kl} x_{kl} \quad (17)$$

The attention weights are normalized over all input sequence with the softmax function. The input features are then weighted by the normalized values and sum to obtain a single vector $\hat{x}_k$ representing the attended features of the input video for a modality k.

After obtaining the attended modality features for all modalities, we combined these features by first passing each of them to a linear layer with weight normalization (Salimans and Kingma 2016) followed by ReLU. All modalities are projected to the same dimensional space D. Then we use Hadamard product to combine the features from different modalities. The result is a single vector $\hat{z}$ representing the combined modality features of the input video.

$$\tilde{z}_k = ReLU(weightNorm(W_{kz}\hat{x}_k + b_{kz})) \quad (18)$$

$$\tilde{z} = \prod^{K} \tilde{z}_k \quad (19)$$

Question-to-Caption Attention

We also used a question-guided attention on the caption sequence. Here the attention attends to information from different positions in the caption, representing by hidden states obtained from the caption encoder $(h_1^{cap}, ..., h_{N^{cap}}^{cap})$. First, both question features q and caption hidden state h_n^{cap} are passed through separate linear layers with ReLU activation to project them to the same dimensional space D^{cap}. The question features is then expanded to have the same sequential dimension N^{cap} as the caption features $\tilde{q}_{cap}^{exp} \in$

$\mathbb{R}^{N^{cap} \times D^{cap}}$ and we then used Hadamard product to create a vector for question-caption features f_{cap}. The rest of the attention is similar to our Question-to-Multimodal Attention described above.

$$f_{cap} = \tilde{h}_n^{cap} \cdot \tilde{q}_{cap}^{exp} \tag{20}$$

$$\alpha^{cap} = softmax(Conv(ReLU(Conv(f_{cap})))) \tag{21}$$

$$\hat{h}^{cap} = \sum_{n=1}^{N^{cap}} \alpha_n^{cap} h_n^{cap} \tag{22}$$

Response Decoder

To generate each system response, each dialog history H, question Q, and video V are paired with a sequence of output words to predict a target sequence T. A GRU decoder is used to define a distribution over output words. For each decoding step m:

$$h_m^{res} = GRU_{resDec}(h_{m-1}^{res}, [y_{m-1}, g]) \tag{23}$$

$$g = \hat{h}^{qes} \oplus \tilde{z} \oplus h_T^{his} \oplus \hat{h}^{cap} \tag{24}$$

where g is the concatenation of question encoding, audio-visual fusioned encoding, dialog history encoding up to the last dialogue turn T, and caption encoding. The decoder sequentially predicts each token using softmax function:

$$p(T|H,Q,V) = \prod_{m=1}^{M} \frac{exp(f(h_{m-1}^{res}, e_{y_m}))}{\sum_{y'} exp(f(h_{m-1}^{res}, e_{y'}))} \tag{25}$$

where e_{y_m} is the output word embedding, h_{m-1}^{res} is the output hidden vector of the decoder at decoding step $m-1$, and f is the activation function between h_{m-1}^{res} and e_{y_m}.

Question, dialog history, and video caption encoders and the response decoder use different GRUs with separate parameters to capture different patterns of word composition. Similarly to (Hori et al. 2018), we use a beam search technique with beam size 5.

Experiments

We used the standard objective function log-likelihood of the target sequence T given the dialog history H, question Q, and video V, which at decoding time provides the statistical decision problem:

$$\hat{T} = \operatorname*{argmax}_T \{\log p(T|H,Q,V)\} \tag{26}$$

For each encoder and decoder, we used an independent single forward GRU layer. The number of hidden units is set to 512 for all the encoders and decoder. We also separate the parameters of the word embedding for question, dialog history, caption encoders and response decoder. We chose to initialize all word embeddings with 200-dimensional Glove embedding (Pennington, Socher, and Manning 2014) pre-trained on Wikipedia and Gigaword [1]. The large size of the training dataset helps to bootstrap the embeddings to contain more meaningful semantic information in each word. We trained each model up to 15 epochs with a decaying learning rate schedule. The learning rate is initialized to 0.001. We used the ADAM optimizer (Kingma and Ba 2014) to train the model. The batch size is set to 64 during training. For each training, we selected the best model with the lowest perplexity on the official validation dataset.

[1] `https://nlp.stanford.edu/projects/glove/`

Data

The main objective for Track 3 of DSTC7 is training an end-to-end multimodal dialog system based on Charades videos (Sigurdsson et al. 2016). We downloaded the data from the official links provided by the organizers. Table 1 summarizes the data provided for this track. Each dialog consists of 10 questions about a given video and corresponding 10 responses. Each dialog was yielded by two Amazon Mechanical Turk (AMT) workers. One of the workers played the role of an answerer who already watched the entire video while the other did not. Each answerer had to answer the other worker's questions based on the previous dialog history and the input video (including audio and visual features and/or video caption). For each dialog of the official test set, we generated a response corresponding to the position of the *UNDISCLOSED* token i.e. 1710 responses in total for each of our submissions. We used the official training dataset to train our system and the official validation dataset to validate and select the best models. We did not merge validation data to the official training data so that we can compare the results to the baselines (Hori et al. 2018). We also utilized the audio and visual feature extractors provided by the organizers. Particularly, we used the I3D_rgb and I3D_flow features from the "Mixed_5c" layer of the I3D network (Carreira and Zisserman 2017) for visual features and Audio Set VGGish (Hershey et al. 2017) for audio features.

Table 1: DSTC7 Video Scene-aware Dialog Dataset

	Official Training	**Official Validation**	**Official Test**	**Prototype Test**
# of Dialogs	7,659	1,787	1,710	733
# of Turns	153,180	35,740	13,490	14,660
# of Words	1,450,754	339,006	110,252	138,790

Official Results

We evaluated our submissions and the baselines using corpus-level BLEU1 to BLUE4 (Papineni et al. 2002), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), ROUGE-L (Lin 2004), and METEOR (Banerjee and Lavie 2005). Results for these metrics were provided by the DSTC organizers themselves. We submitted two systems to Track 3, representing the two settings: *Video+Text* and *Text Only*. For *Video+Text* setting, in addition to the dialog data, we use the I3D_rgb features and VGGish features for visual and audio features. In this setting, we did not submit the system that also uses video caption data as we did not find significant improvement during testing with the prototype test data. For *Text Only* setting, we used the dialog data as well as the video caption to train our model. We did not use the video summary data. We compared these systems to the baseline (Hori et al. 2018).

Table 2: Objective evaluation results on official test data. The highest value in each metric is highlighted in bold.

Model	Video	Text	BLEU B-1	B-2	B-3	B-4	METEOR	ROUGE_L	CIDEr
Baseline	I3D_rgb_flow	Dialog	0.621	0.480	0.379	0.305	0.217	0.481	0.733
Baseline	I3D_rgb_flow+VGGish	Dialog	0.626	0.485	0.383	0.309	0.215	0.487	0.746
Ours	I3D_rgb+VGGish	Dialog	0.631	**0.491**	**0.390**	**0.315**	0.239	0.509	0.848
Ours	-	Dialog+Caption	**0.633**	0.490	0.386	0.310	**0.242**	**0.515**	**0.856**

Table 3: Human evaluation results on official test data.

Model	Video	Text	Human Rating
Baseline	I3D_rgb_flow+VGGish	Dialog	2.848
Ours	-	Dialog+Caption	3.080
Official Test	-	-	3.938

The objective and subjective evaluation results are shown in Table 2 and 3 respectively. The ground truth responses from the official test data was also evaluated by human judges and the results were provided by the organizers. All of our submissions show improvement over the baselines in terms of BLEU, CIDEr, METEOR, and ROUGE-L. Among our systems' results, the *Video+Text* system performs better than the *Text Only* system in terms of BLEU scores, with an exception for BLEU-1 where *Text Only* system is slightly better than *Video+Text* system. The *Text Only* system outperforms the *Video+Text* system in terms of METEOR, ROUGE-L, and CIDEr. As ROUGE-L is a recall oriented metric designed for summarization and METEOR is a translation metric, they may not be completely suitable to evaluate generated dialog responses. This might explain the inconsistency between these metrics and BLEU scores when we compare *Video+Text* and *Text Only* system results. The difference between *Video+Text* and *Text Only* results is also not significant. As we expect the information conveyed from video visual and audio features is more than video caption alone, the performance of *Video+Text* system could have been further improved. For human evaluation, the results are consistent with objective scores in which our proposed *Text Only* model outperforms the baseline. However, there is still a significant gap of human rating (0.858 difference) between our generated responses and the official test responses.

Prototype Results

Table 4 shows the results of our proposed models trained on the official training data and evaluated on the prototype test data. The evaluation metrics are the same as the official results, including BLEU1-4, METEOR, ROUGE-L, and CIDEr. The evaluation codes were provided by the organizer and based on MS COCO caption generation [2]. Here we analyzed how changes in different modules affect the model performance. *Model #1* is essentially our prototype results running with the baseline model (Hori et al. 2017). As we changed from LSTM to GRU (*Model #2*) in all encoders and decoder, we did not observed much changes in terms of evaluation metrics. However, as GRU is more computationally efficient, we applied GRU in the remaining experiments. As we applied Question-to-Multimodal Attention (*Model #3*), the performance increased slightly across all the metrics except for BLEU1. When we combined Question-to-Multimodal Attention with Non-linear Multimodal Feature Fusioning as described above (*Model #4*), the results increased significantly in terms of BLEU scores. However, as we added I3D_flow features of the input video (*Model #5*), the performance became worse. We speculate that our Multimodal Feature Fusioning method is not suitable to combined more than two modalities, and hence, adding a third feature such as I3D_flow affected the results. When we added caption features with question-guided attention mechanism, the model performance clearly improved (*Model #6 and #7*). We also experimented with Caption-to-Multimodal Attention by replace q in Equation 13 to $\hat{h}^{cap}$ (*Model #7*). However, the results were worse than using the proposed Question-to-Multimodal Attention.

When using pretrained Glove embedding, we observed improved results with 200-dimensional embedding (*Model #9*). With 100-dimensional Glove embedding (*Model #8*), the model is not as good as one without pretrained embedding (*Model #4*). This could be caused by the 100-dimensional embedding space not being able to capture enough useful semantic meaning in the training corpus. Similarly to (*Model #5*), we did not see improvement when adding I3D_flow into the input video features with pretrained word embedding (*Model #11*). Surprisingly, as we added caption features with attention (*Model #12*), the performance became worse except for BLEU1. This is inconsistent with our previous finding in cases without pretrained word embedding. Among the *Video+Text* setting models, *Model #10* showed the best performance and was used as our submission to the DSTC7. We also experimented with only input text without the input video (*Model #12* and *#13*). As we used pretrained 200-dimensional Glove embedding (*Model #13*), we achieved better performance and used this model as our submission for the *Text Only* setting.

Discussion

Using the prototype test data, we tested our best *Video+Text* setting model and compared some sample responses with the baseline model responses as well as the reference responses in Table 5. In terms of correctness, our responses are able

[2] `https://github.com/tylin/coco-caption`

Table 4: Objective evaluation results for models trained on official data and evaluated on prototype test data

Model No.	Video	Text	RNN	cap-att	mm-att	mm-fusion	word-emb	BLEU B-1	BLEU B-2	BLEU B-3	BLEU B-4	METEOR	ROUGE-L	CIDEr
1	I3D_rgb+VGGish	Dialog	LSTM	-	Baseline	Baseline	No	0.272	0.176	0.120	0.086	0.122	0.298	0.842
2	I3D_rgb+VGGish	Dialog	GRU	-	Baseline	Baseline	No	0.266	0.174	0.120	0.086	0.121	0.299	0.843
3	I3D_rgb+VGGish	Dialog	GRU	-	QuesProj+Conv	Baseline	No	0.269	0.175	0.121	0.087	0.123	0.301	0.851
4	I3D_rgb+VGGish	Dialog	GRU	-	QuesProj+Conv	FC+HdmProd	No	0.291	0.186	0.126	0.090	0.127	0.301	0.824
5	I3D_rgb_flow+VGGish	Dialog	GRU	-	QuesProj+Conv	FC+HdmProd	No	0.284	0.183	0.125	0.090	0.123	0.296	0.802
6	I3D_rgb+VGGish	Dialog+Caption	GRU	QuesProj+Conv	QuesProj+Conv	FC+HdmProd	No	0.304	0.198	0.137	0.100	0.131	0.312	0.891
7	I3D_rgb+VGGish	Dialog+Caption	GRU	QuesProj+Conv	CapProj+Conv	FC+HdmProd	No	0.298	0.194	0.135	0.097	0.129	0.309	0.867
8	I3D_rgb+VGGish	Dialog	GRU	-	QuesProj+Conv	FC+HdmProd	Glove100	0.276	0.181	0.125	0.091	0.124	0.304	0.870
9	I3D_rgb+VGGish	Dialog	GRU	-	QuesProj+Conv	FC+HdmProd	Glove200	0.307	0.204	0.144	0.106	0.136	0.320	0.995
10	I3D_rgb_flow+VGGish	Dialog	GRU	-	QuesProj+Conv	FC+HdmProd	Glove200	0.303	0.201	0.141	0.103	0.133	0.317	0.962
11	I3D_rgb+VGGish	Dialog+Caption	GRU	QuesProj+Conv	QuesProj+Conv	FC+HdmProd	Glove200	0.314	0.204	0.142	0.102	0.136	0.317	0.940
12	-	Dialog+Caption	GRU	QuesProj+Conv	-	-	No	0.293	0.194	0.136	0.100	0.130	0.313	0.933
13	-	Dialog+Caption	GRU	QuesProj+Conv	-	-	Glove200	0.312	0.203	0.141	0.102	0.135	0.316	0.931

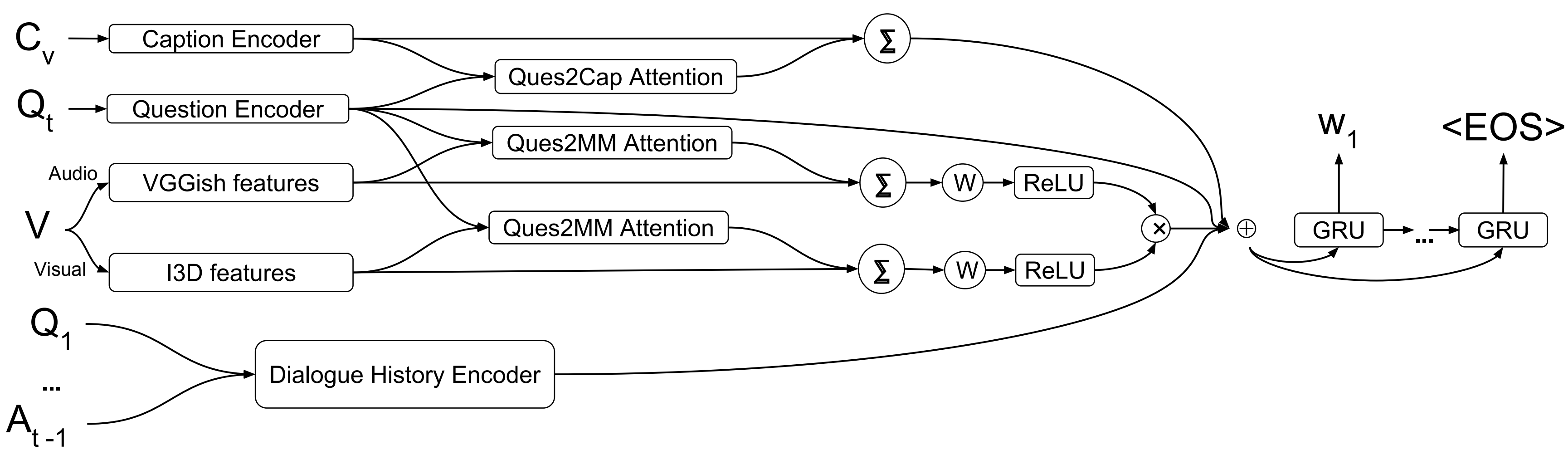

Figure 1: Overview of the proposed end-to-end multimodal dialog systems with attention mechanisms on multimodal features and video caption embedding

Table 5: Sample dialogues for an input video with responses generated by the best baseline model $A^{baseline}$, our best *Video+Text* model A^{ours}, and the reference response A^{ref} from the prototype test dataset

Video Caption	**"A person is working on their laptop, another person come in carrying a bag of groceries. they set the groceries down and start cooking."**
Q_1	how many people in the video?
A_1^{ref}	there are two person in video
$A_1^{baseline}$	there are one person in the video
A_1^{ours}	there are two people in the video
Q_2	what gender are they?
A_2^{ref}	there are both males in video
$A_2^{baseline}$	they are both men in the video
A_2^{ours}	one is a male and the other is a male
Q_3	are they talking in the video?
A_3^{ref}	no, both are not talks to each other
$A_3^{baseline}$	yes they are talking in the video
A_3^{ours}	no one is talking in the video
Q_4	what is the second person doing?
A_4^{ref}	second person bring a vegetable bag come to kitchen room
$A_4^{baseline}$	he is eating a sandwich
A_4^{ours}	the person is standing in front of the refrigerator
Q_5	are they drinking something in the video?
A_5^{ref}	no, they are both not a drink
$A_5^{baseline}$	no, he is eating a sandwich
A_5^{ours}	no, they are not drinking anything in the video

to express the correct answers better than the baseline responses in A_1, A_3, and A_5. This indicates that our model is able to capture the video features, either visual feature (e.g. action "drinking", number of people) or audio feature (action "talking") better than the baseline approach. For Q_2, both the baseline model and our model are able to express the correct answers. However, the baseline answer is preferred as its response is more concise than our response. For Q_4, both baseline response and our response do not match the reference response. However, we could argue that our response is a better possible response in this context as "standing in front of the refrigerator" is more appropriate than "eating a sandwich" action for this particular input video.

In addition, we also observe that our generated responses have a large proportion of negative answers i.e. answers that response "no" to yes/no questions. This might be due to the high frequence of negative responses in the training corpus. We also noticed our models tend to generate a universal answer "yes that is all happening in the video" to questions such as "is that all happened in the video?" This type of questions might require further cross-references to reason over the dialog history before generating a correct response. Using a hierarchical encoder for the dialog history might not be sufficient for this type of questions.

Conclusion

DSTC7 Track 3 offered a valuable opportunity to investigate multimodal dialog systems in a video-oriented setting rather than just visual setting (Das et al. 2017b) (Das et al. 2017a). It set a good framework to explore how state-of-the-art feature extraction models such as VGGish and I3D can be pretrained to extract the visual and audio features efficiently. We also found that techniques used in visual QA models such as (Anderson et al. 2018) (Teney et al. 2017) could be adapted into this setting to improve the model performance. We hope to explore in this multimodal dialog setting further in the future with larger scale datasets and probably in other variations of dialog settings e.g. open-domain dialogs, task-oriented dialogs. Besides bootstrapping with pretrained word embeddings, we could also pretrain parts of the model on a larger dialog corpus that covers similar topics and types of questions and responses. An example corpus is the Movie QA dataset (Tapaswi et al. 2016) containing Q-A pairs constructed to query about movie contents. This corpus can be used to pretrain the model before further training with the DSTC7 training dataset. Alternatively, unsupervised pretraining with language models such as BERT (Devlin et al. 2018) has shown significant improvement in multiple NLP tasks and could be applied into multimodal dialog settings.

Acknowledgements

The first author is supported by A*STAR Graduate Scholarship. We also thank the insightful comments from the reviewers and E. Hovy.

References

[Anderson et al. 2018] Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, volume 3, 6.

[Antol et al. 2015] 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.

[Banerjee and Lavie 2005] Banerjee, S., and Lavie, A. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

[Carreira and Zisserman 2017] 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.

[Cho et al. 2014] Cho, K.; van Merriënboer, B.; Gülçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. Doha, Qatar: Association for Computational Linguistics.

[Das et al. 2017a] Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017a. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.

[Das et al. 2017b] Das, A.; Kottur, S.; Moura, J. M. F.; Lee, S.; and Batra, D. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. *2017 IEEE International Conference on Computer Vision (ICCV)* 2970–2979.

[Devlin et al. 2018] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Goyal et al. 2017] Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, volume 1, 3.

[Hershey et al. 2017] 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 et al. 2017] 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 et al. 2018] 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*.

[Kingma and Ba 2014] Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Li et al. 2018] Li, Y.; Yao, T.; Pan, Y.; Chao, H.; and Mei, T. 2018. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7492–7500.

[Lin 2004] Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

[Papineni et al. 2002] Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

[Pennington, Socher, and Manning 2014] Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

[Salimans and Kingma 2016] Salimans, T., and Kingma, D. P. 2016. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, 901–909.

[Sigurdsson et al. 2016] 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.

[Tapaswi et al. 2016] Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Teney et al. 2017] Teney, D.; Anderson, P.; He, X.; and van den Hengel, A. 2017. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *arXiv preprint arXiv:1708.02711*.

[Vedantam, Lawrence Zitnick, and Parikh 2015] Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

[Vinyals et al. 2015] Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

[Xu et al. 2015] Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.