Multimodal Dialogue State Tracking By QA Approach with Data Augmentation

Xiangyang Mou[1], **Brandyn Sigouin**[1], **Ian Steenstra**[1], **Hui Su**[1,2]
[1]Rensselaer Polytechnic Institute, 110 Eighth Street, Troy, NY 12180 USA
[2]IBM Thomas J Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598 USA
{moux4, sigoub, steeni}@rpi.edu
huisuibmres@us.ibm.com

Abstract

Recently, a more challenging state tracking task, Audio-Video Scene-Aware Dialogue (AVSD), is catching an increasing amount of attention among researchers. Different from purely text-based dialogue state tracking, the dialogue in AVSD contains a sequence of question-answer pairs about a video and the final answer to the given question requires additional understanding of the video. This paper interprets the AVSD task from an open-domain Question Answering (QA) point of view and proposes a multimodal open-domain QA system to deal with the problem. The proposed QA system uses common encoder-decoder framework with multimodal fusion and attention. Teacher forcing is applied to train a natural language generator. We also propose a new data augmentation approach specifically under QA assumption. Our experiments show that our model and techniques bring significant improvements over the baseline model on the DSTC7-AVSD dataset and demonstrate the potentials of our data augmentation techniques.

Introduction

Given a conversation flow, question-answering dialog systems are an ideal mechanism for investigating the nuances of dialog state-tracking. This is based on the hypothesis that the natural language response to any question depends on the point in time in the conversation that the question is asked. A simple example of this is if one asks, "is there a cat in the video?". One may ask a natural follow-up question such as "what color is it?", where the subject of the question is a pronoun, but the true meaning of which is stored in the dialog around the question instead of directly within it. The logical response to this may depend on such information. However, without considering the previous question(s), it is difficult for a generative model to produce information about the subject as there is little to no relevant context from which to deduce this. Even if there is information about "color" in the video modality, the word "it" is still ambiguous without understanding the current state of the dialog. To capture such a relationship as it pertains to natural language generation, we investigate dialog history encoding techniques in order to fuse text with the other modalities. We believe that by successfully answering questions in the Audio Visual Scene-aware Dialog track of DSTC8, it will provide evidence that dialog context in QA setting does in fact store information pertinent to the current stage of a conversation.

Furthermore, we introduce a new data augmentation technique for dialog state-tracking problems. We believe that for QA problems, the presence of information in a dialog matters more than its temporal location. For each question in the dialog, we encode the QA pairs in the dialog history up until the point of the question we aim to answer. Leveraging this claim, we shuffle the dialog histories effectively increasing the size of our dataset. Additionally, we find that teacher forcing during the training procedure is important for modeling natural language sentence generation. Unlike our argument with dialog state tracking, we believe that there is a time dependency present within a sentence. During generation, the current word or token being predicted depends on the word(s) generated previously and the respective order. It should be noted that in the case of the first word, we use the special start-of-sentence token as the default first dependency. To capture this relationship during training, each token prediction must be treated as its own event; thus, the model should assume that the previously generated words are correct.

Background

Dialog State Tracking aims to model natural language by leveraging the argument that conversational patterns maintain information. Thus, there should be contextual information hidden in the dialog around the current conversation point. One of the key domains of interest for this proposal is that of Dialog Question Answering systems. In common conversations, questions asked later than the first utterance likely pertain to earlier utterances in some way, such as follow-up questions. Realistically, these relationships are not always clear and it is challenging to evaluate if there is indeed a dependency from utterance to utterance. Alternatively, one must consider the possibility that, in these types of conversations, utterances may be singleton disjoint events. This argument is important for extending encoder-decoder architectures to dialog systems. Originally used for

natural language translation, these models rely on the formation of a context vector from sentence components, where a time dependency most likely exists (Cho et al. 2014). Intuitively, if one is to extend this theory to the macro-level of full conversation, it is natural to assume that a similar relationship must exist. Question Answering systems provide a measurable way to evaluate these dialog state tracking hypotheses.

The Audio Visual Scene-Aware Dialog track of the Dialog Systems Technology Challenge 8 (AVSD, DSTC8) exists to encourage further research into the complex domain of natural language generation from multimodal data. DSTC8 and the AVSD track are an extension of DSTC7. This investigation is based on data from DSTC7. The provided modalities in the challenge dataset include visual data in the form of processed video frames, the audio data extracted from those videos, and three text modalities consisting of a summary, caption, and dialog history. The dialog history is comprised of ten question-answer pairs per example. The challenge participant is free to use any or all of the modalities, but is encouraged to attempt synthesis with the text and video-derived inputs. Ideally, the challenge aims to fuse computational linguistics, computer vision, and signal processing to generate meaningful natural language. The development of this technology is important for the emerging fields of human-agent interaction beyond just the scope of language to language interactions. One may envision an agent which aids a user through interactions with visual data. The user should be able to freely inquire about that information while expecting a reasonably confident response. This work as a whole largely extends the encoder-decoder model used for natural language translation tasks (Cho et al. 2014). However, instead of just deriving a context vector from language alone, AVSD encourages using context from a much larger scope, which presents the unique challenge of multimodal fusion.

Related Work

Dialog State Tracking, as a research domain, is broadly defined as the deduction of evidence from linguistic information over the course of a conversation to complete a task (Williams et al. 2013). Within the context of AVSD, the state of the dialog is part of the internal representation of the data from which sentence generation is based upon. Essentially the goal of state tracking is the maintenance a belief state (Mrkšić et al. 2016). These belief states act as a probability space from which a natural language generator, in our case a decoder, derives its context. Related topics include, but are not limited to, image captioning (You et al. 2016) and, to an extent, the visual-question answer challenge (Antol et al. 2015). Both of these topics cover the fusion of Computer Vision with Natural Language Understanding to generate new information. Key distinctions between VQA and AVSD is that VQA does not incorporate a dialog flow necessary for state-tracking and does not require the generation of language. A unique aspect of AVSD that is important to emphasize is that it is an open-domain QA setting and is not restricted to goal-specific tasks.

Other works that were extremely influential in the development of this model are the developments of the Gated-Recurrent Unit (GRU) and attention mechanisms (Bahdanau, Cho, and Bengio 2014). Bahdanau, Cho, and Bengio demonstrate the effectiveness of Recurrent Neural Networks with fewer parameters and how to address the issue of capturing meaningful temporal information within a sequence of natural language tokens. Sanabria, Palaskar and Metze (Sanabria, Palaskar, and Metze 2019), extended this technology to the previous iteration of the AVSD challenge to encode text modalities. The use of GRUs is shown to effectively address the issues pertaining to the larger size of traditional LSTM encoders, which can become very costly when used within multimodal models. Furthermore, they use the idea of attention to calculate the importance of a data frame on the overall goal of the encoding. By extending this idea to multimodal data, one can tie together diverse modalities via a common attention source. In theory, this should reduce the complexity of learning from just raw modality fusion.

Proposed Approach

In general, our model follows an encoder-decoder framework (Fig.1) which can be commonly seen in language generation tasks (Wen et al. 2015; Wu et al. 2016). In encoding, bidirectional Gated Recurrent Units (BiGRU) are used for visual, audible and textual sequence embedding of which are further masked by question-guided attention. Early multimodal fusion, among different modalities, is performed to form the context representation for the decoder. The decoder takes in the context and question information in order to generate a response to the given question using a GRU. In addition, a scheduled sampling strategy (Bengio et al. 2015) is applied within the training phase in an effort to increase the efficiency of the training and robustness of the inference.

Feature Encoding With Soft Attention

Originally, in AVSD, a total of 7 different features are provided for each sample, including optical flow of video, RGB frames of video, audio, captioning, annotator generated summary, dialogue history and the question (Alamri et al. 2018). Empirically, the best result is usually achieved by an optimal combination of features. In our work, the caption is not used because much of the information in this modality is duplicative of information found within other text modalities, such as the summary.

For textual inputs including question, summary and individual sentences in dialogue history, we choose a pre-trained fastText model (Mikolov et al. 2018) for word embedding. We find the fact that there are a fair number of typos including missing and reversed letters within individual words caused by annotators during data collection. The typos would generate out-of-vocabulary (OOV) words and mislead the essential meaning of the sentence. The fastText embedding features a character-level encoding and is therefore considered a more suitable language model for AVSD in terms of minimizing the negative effect of OOV words in language modelling. In our work, we take the advantage of

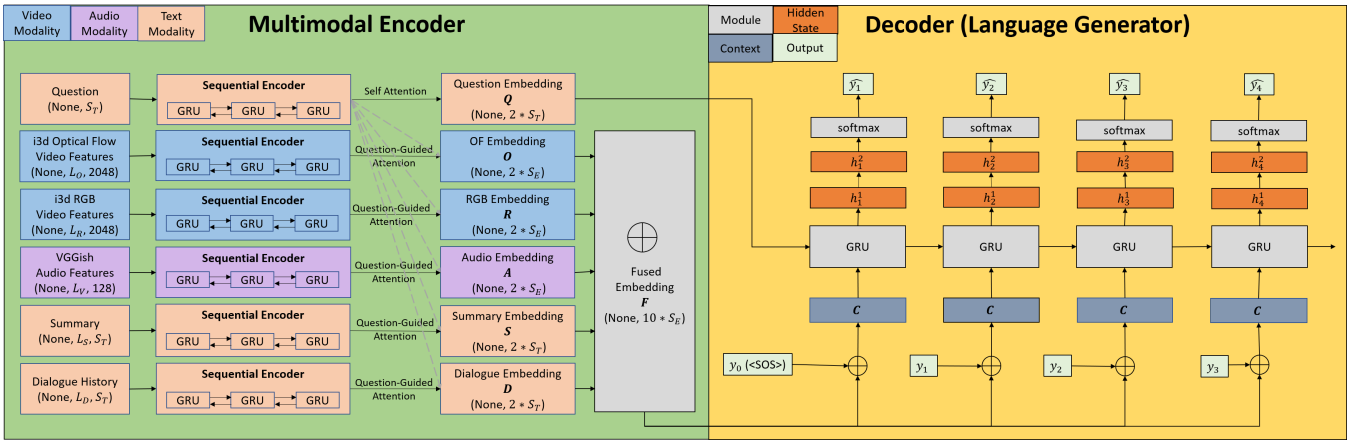

Figure 1: Model Architecture. The number and notation in the brackets, e.g. (None, L_s, S_T), describe the feature dimension. $\bigoplus$ means simple concatenation among different modalities. y_i represents the i-th word in the ground truth.

an existing library [1] based on (Patel et al. 2018) due to its OOV handling. It would find the closest known word vector to replace an OOV word and to the best extent, restore the sentence-level representation.

For question embedding, we input the sequence of word vectors into a BiGRU, apply a 2-layer convolutional self-attention mask to the outputs of BiGRU and then take the average of the attended BiGRU outputs to obtain the final question representation. Let $\mathbf{q} = [q_1, q_2, \ldots, q_n] \in R^{n \times d_w}$ denote a sequence of question word vectors with a length of n; where $q_i \in R^{1 \times d_w}$ is the i-th word vector with a dimension of d_w:

$$\tilde{\mathbf{q}} = BiGRU(\mathbf{q}) \tag{1}$$

$$m_q = ReLU(Conv2D(ReLU(Conv2D(\tilde{\mathbf{q}})))) \tag{2}$$

$$\mathbf{Q} = ReLU(mean(\tilde{\mathbf{q}} \odot m_q)) \tag{3}$$

where $\tilde{\mathbf{q}} \in R^{n \times D}$ are the outputs of all BiGRU cells; $m_q \in R^{n \times D}$ is the attention weight; $\mathbf{Q} \in R^{1 \times D}$ is the final question sentence representation; D is the dimension of the output of each BiGRU cell; $Conv2D$ represents a 2-dimensional convolution layer with a size of 1×1 and $\odot$ indicates element-wise product.

For summary sentence embedding, we would like it to focus more on question-related words; therefore, we choose to use question-guided general attention (Luong, Pham, and Manning 2015) rather than self-attention. Then, instead of using the output of the last BiGRU cell, we do max pooling over the outputs of all BiGRU cells to form their final representations out of our belief that the embedding from max pooling includes the dominant signals across all dimensions. Let $\mathbf{s} = [s_1, s_2, \ldots, s_n] \in R^{n \times d_w}$ denote a sequence of summary word vectors with a length of n; where $s_i \in R^{1 \times d_w}$ is the i-th word vector with the dimension of d_w:

$$\tilde{\mathbf{s}} = BiGRU(\mathbf{s}) \tag{4}$$

$$m_s = softmax(\tilde{\mathbf{s}} W_s \tilde{\mathbf{q}}^T) \tag{5}$$

[1] https://github.com/plasticityai/magnitude

$$\mathbf{S} = MaxPool(ReLU(\left[m_s^T \tilde{\mathbf{s}}; \tilde{\mathbf{q}}\right] W_{os})) \tag{6}$$

where $\tilde{\mathbf{s}} \in R^{n_s \times D}$ are the outputs of all BiGRU cells; $W_s \in R^{D \times D}$ is the trainable weight; $attn \in R^{n_s \times n_q}$ is the question-guided attention with n_s as length of summary and n_q as length of question; $W_o \in R^{2D \times D}$ is another trainable weight; $\mathbf{S} \in R^{1 \times D}$ is the final summary sentence representation and D is the dimension of the output of each BiGRU cell.

Similarly, for dialogue history, the same question-guided attention is applied following Eq.4, Eq.5 and Eq.16. The only difference is that, instead of using the sequence of word vectors, we use the sequence of sentence vectors to encode dialogue history. Each question and answer is treated as a single sentence and has a single sentence representation vector regardless of the actual number of sentences it contains. If no dialogue history is available for a specific question, a zero vector would be used as the representation whose elements are all zeros:

$$\begin{cases} \mathbf{d} = [q_1; a_1; \ldots q_{n-1}; a_{n-1}], & n > 1 \\ \mathbf{d} = \mathbf{0}, & n = 1 \end{cases} \tag{7}$$

$$\tilde{\mathbf{d}} = BiGRU(\mathbf{d}) \tag{8}$$

$$m_d = softmax(\tilde{\mathbf{d}} W_d \tilde{\mathbf{q}}^T) \tag{9}$$

$$\begin{cases} \mathbf{D} = MaxPool(ReLU(\left[m_d^T \tilde{\mathbf{d}}; \tilde{\mathbf{q}}\right] W_o)), & n > 1 \\ \mathbf{D} = \mathbf{0}, & n = 1 \end{cases} \tag{10}$$

where $\mathbf{d}$ is equivalent to $\mathbf{s}$ in Eq.4; q_i, a_i are the question and answer sentence vector in dialogue history; n means the n-th question-answer pair in the dialogue history starting from 1; $\mathbf{0}$ is a vector whose elements are zeros; $\tilde{\mathbf{d}} = BiGRU(\mathbf{d})$; $attn = softmax(\tilde{\mathbf{d}} W_d \tilde{\mathbf{q}}^T)$.

For video and audio modalities, we do not train our own video feature extractor but directly use the features provided by the AVSD dataset, namely i3d-flow, i3d-rgb and VGGish. i3d-flow and i3d-rgb are generated by the state-of-the-art video feature extractor (Carreira and Zisserman 2017) and VGGish by the state-of-the-art audio feature extractor

(Hershey et al. 2017). Since they are all frame-wise and of variable length, we use another BiGRU to capture the temporal dependency on top of individual modality. Following the same processing procedure as shown in Eq.4, Eq.5 and Eq.16, we define:

$$\mathbf{o} = [o_1; o_2; \dots o_l] \tag{11}$$

$$\mathbf{r} = [r_1; r_2; \dots r_m] \tag{12}$$

$$\mathbf{a} = [a_1; a_2; \dots a_n] \tag{13}$$

where $\mathbf{o}$, $\mathbf{r}$, $\mathbf{a}$ are equivalent to $\mathbf{s}$ in Eq.4 and their corresponding outputs are denoted as $\tilde{\mathbf{o}}$, $\tilde{\mathbf{r}}$, $\tilde{\mathbf{a}}$; o_i, r_i, a_i denote individual frame representation for i3d-flow, i3d-rgb and audio respectively. The final i3d-flow, i3d-rgb and audio representation, are denoted as $\mathbf{O}, \mathbf{R}, \mathbf{A}$, s.t:

$$\mathbf{O} = MaxPool(ReLU(\left[m_o^T \tilde{\mathbf{o}}; \tilde{\mathbf{q}}\right] W_{oo})) \tag{14}$$

$$\mathbf{R} = MaxPool(ReLU(\left[m_r^T \tilde{\mathbf{r}}; \tilde{\mathbf{q}}\right] W_{or})) \tag{15}$$

$$\mathbf{A} = MaxPool(ReLU(\left[m_a^T \tilde{\mathbf{a}}; \tilde{\mathbf{q}}\right] W_{oa})) \tag{16}$$

where W_{oo}, W_{or}, W_{oa} are the trainable weights; m_o^T, m_r^T, m_a^T are the question-guided attention masks for i3d-flow, i3d-rgb and audio modalities, following the same strategy as in Eq.5

multimodal Fusion

The context vector contains information from different modalities and will be used for natural answer generation. We form it by simple concatenation in order to achieve early fusion across multimodalities, i.e.

$$\mathbf{C} = [\mathbf{O}; \mathbf{R}; \mathbf{A}; \mathbf{S}; \mathbf{D}] \tag{17}$$

Decoder

Because our system is open-domain and supposed to generate answers of free-form, any extraction-based language generation approach (Wang and Jiang 2016) would be out of our consideration. In our work, we adopt a two-layer BiGRU as the natural language generator. One good point of a RNN is that it can take in variable-length input and also generate variable-length output. More importantly, RNN is known for its ability of modelling long-term spatial or temporal dependency within a sequence so that the language generated could be more fluent and readable. A two-RNN-layer at the output is a commonly-used setting in related areas such as a language generation system and image captioning. In addition, it is said to be beneficial in decreasing the opportunity of repeated words within generated language; which, is one of the difficulties in the natural language generation task.

Given the input question, context and the preceding words, the language generator models the probability of each next word. We would like the GRU network to be able to reason over the context and predict the next work based on the question and the preceding part of the answer.

$$P(\omega_1, \dots, \omega_n) = \Pi_{i=1}^{n} P(\omega_i | \omega_{0 \sim i-1}; \mathbf{C}; \mathbf{Q};) \tag{18}$$

In our work, we initialize the GRU hidden state with the question vector Q and take the concatenation of the context $\mathbf{C}$ and the 1-gram preceding word as the input to GRU cell. We believe that the context contains more information than the question and do not want the context to forget any information along the progressive prediction procedure.

$$z_t = \sigma(W_z[\mathbf{C}; \mathbf{w}_{t-1}] + U_z h_{t-1}) \tag{19}$$

$$r_t = \sigma(W_r[\mathbf{C}; \mathbf{w}_{t-1}] + U_r h_{t-1}) \tag{20}$$

$$h_t = \sigma(W[\mathbf{C}; \mathbf{w}_{t-1}] + r_t \odot U h_{t-1}) \tag{21}$$

Eq.19, Eq.20 and Eq.21 show how our GRU cell updates its hidden state at time t. $\mathbf{w}_{t-1}$ is the word vector at time $t-1$ in the prediction; $[\mathbf{C}; \mathbf{w}_{t-1}]$ represents the input to the GRU cell; h_t and h_{t-1} are the hidden states at different time; W_z, W_r, W, U_z, U_r, U are the trainable weights in GRU cell.

Experiments

Training

In our training phase, we use an Adam optimizer to minimize the cross entropy error between the predicted word and the ground truth. The F1 score on the validation set is used to terminate the training procedure. To increase the training efficiency and accuracy, we use teacher forcing (Williams and Zipser 1989); it uses the ground truth word to predict the next word during training. More specifically in Eq.19, Eq.20 and Eq.21, $\mathbf{w}_{t-1} = y_{t-1}$ rather than $\hat{y}_{t-1}$ just as shown in Fig.1.

As a comparison, we also try a scheduled sampling technique (Bengio et al. 2015) which introduces probability into teacher forcing. Different from a traditional teacher forcing technique that always uses the ground truth word, there is a certain probability in scheduled sampling to use the predicted word as the input to the GRU cell. (Bengio et al. 2015) claims that scheduled sampling could improve the generalization and robustness. However, we do not see significant improvements in our tests. Therefore, we remain to use teacher forcing for our experiments.

Data Augmentation

Data augmentation is a widely used technique in deep learning. Most of the time, it is of great help and can outperform the baseline significantly for data driven approaches. After examining the AVSD dataset, we find a way to enlarge the size of the training set by several orders of magnitude. The training set of AVSD, provides 10 question-answer pairs for each video.

- The most **basic** way of using the training data is to treat the first 9 pairs as dialogue history and take the last question as what needs to be answered.
- A **quick improvement** would be treating the first n pairs as dialogue history and take the $n+1$-th question as what needs to be answered. This could augment the data by 10 times.
- In our work, since AVSD is regarded as a question answering problem, we do not necessarily care about the sequence order of the dialogue history. Each question-answer pair is being seen as a knowledge point. With this slight difference, we can **shuffle** the first n pairs for the $(n+1)^{th}$ question; ideally it

should not be a problem for a human to answer the $(n+1)^{th}$ question. In other words, the dialogue history $[q_1; a_1; q_2; a_2; q_3; a_3; q_4; a_4]$ can be seen as no obvious difference from $[q_1; a_1; q_4; a_4; q_3; a_3; q_2; a_2]$, or any other order to the $(n+1)^{th}$ question. Theoretically, for a training sample whose dialogue history length is 9, we can generate $P_9^9 - 1 = 362879$ similar samples out of it, following the **shuffle** idea. Thus, the approach could enlarge the training set by tens of thousands times on average.

In our work, we take the **quick improvement** version as the baseline; assuming that most of the people would adopt such technique, and take **shuffle** version as our new data augmentation approach. In our experiments, we only double the training set considering the training time.

Results

Evaluation Metrics We use 7 metrics for evaluating our model which are widely used when evaluating image and video captioning, as well as language generation: BLEU(1-4) (Papineni et al. 2002), METEOR (Denkowski and Lavie 2014), ROUGE-L (Lin 2004), and CIDEr (Vedantam, Zitnick, and Parikh 2014).

Model Performance In order to further evaluate our model's performance, we compare our results with the baseline model; as well as other participants in the DSTC7 challenge on the track of AVSD under two different settings – with and without video related modalities (i3d-flow, i3d-rgb and audio). Our best model fully utilizes the combination of teacher forcing, max pooling, BiGRU and data augmentation techniques. Table.1 shows that our best model outperforms the baseline model significantly and our scores are comparatively better than the majority of other models, demonstrating that our model successfully captures salient signals among multimodalities. But comparing to (Nguyen et al. 2018) under the Text-Only setting, whose architecture is specifically designed to encode the conversation flow, we see a big decrease in all scores. We hypothesize that as we regard the dialogue history as a group of discrete QA pairs, we miss certain inner temporal dependencies, impacting tasks like coreference resolution. While our data demonstrates that the ordering of the dialog history as a whole may not contribute as much information as previously assumed, it is important to consider that phenomena like coreference resolution most likely does depend on temporal location. Fig. 2 shows an example of the importance of deducing meanings of ambiguous terms within a dialog history. Another example is in the case of an ambiguous pronoun. Both situations usually rely on the presence of a previous noun or a certain portion in an image or video. After such information has been deduced, our data suggests that the order of that processed information would likely have little effect on the performance of the system. However, without doing this before shuffling, information that can be obtained through techniques such as coreferencing may be untapped.

When comparing our model under different settings, we notice a significant empirical improvement in every metric when shifting from Video-Text to just Text-Only. This is counter-intuitive because the scores should decrease as the context should be less-complete without the video and audible inputs. On the contrary, the two other models shown evaluating on text-only do see a slight decrease in their scores when compared to the same models that make use of the video-derived modalities. The difference could be the result of our model not capturing the real attention of the video-related signals within relevance to the current questions at the hidden layer of the BiGRU (Nguyen et al. 2018). Without the proper video encoding, the video embeddings could be simply adding noise to our model. Fig. 3 provides an example of such an error. This provides an interesting insight: the video encoder seems to struggle at capturing temporal actions, limiting the amount of useful information present in the encoding. Since this encoding cannot add meaningful information to the context in respect to the current question, it functionally becomes noisy data. Given such findings, the development of more robust video-audio encoding techniques would be a logical next step in this research. The success of the text-only model in this context can likely be explained by the model being able to capture relationships between keywords in the input modalities and the words that appear in the ground truths. In 4 example (a), the model is still able to produce a relevant answer, albeit an empirically incorrect one, regarding audio in the video without actually being able to access that modality. It is safe to conclude here that our model successfully captures textual context, however will struggle in producing the most robust responses when critical information is exclusive to other modalities.

Q1: Is there a guy or a girl in the video?
A1: A man with beard.
Q2: Is that man alone?
A2: _____

Figure 2: An example of dialog history with a coreference resolution issue from the AVSD dev set. If Q_2 was the given question, in order to offer a correct answer, a human would need to figure out whom "that man" exactly referred to.

Fig. 4 provides examples for the importance of utilizing a combination of an image encoder, audio encoder and text understanding. Leaving any of these out could result in inadequate answers compared to the ground truths. For instance, example (a) shows the importance of an audio encoder. The question can only be answered correctly with the information from the audio, but the vocal attention is overshadowed by other modalities. Example (b) illustrates the importance of using image encoding with text understanding because they work together to find that the cloth is black (or dark).

Ablation Study We conduct our ablation study under the Video-Text setting. As shown in Table.2, the techniques of teacher forcing, maximum pooling, average pooling, scheduled sampling, data augmentation and RNN variations has been testified. We find that the best overall scores are from the model that uses teacher forcing, maximum pooling, data

	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Video + Text	(Nguyen et al. 2018)	**0.695**	**0.553**	**0.444**	**0.36**	**0.249**	**0.544**	0.997
	(Le et al. 2019)	0.631	0.491	0.390	0.315	0.239	0.509	0.848
	Our Model	0.586	0.436	0.333	0.262	0.206	0.46	0.704
	(Pasunuru and Bansal 2019)	N/A	N/A	N/A	0.118	0.150	0.378	**1.158**
	(Lin et al. 2019)	0.333	0.196	0.131	0.093	0.129	0.334	0.88
	(Zhuang, Wang, and Shinozaki 2019)	0.29	0.184	0.125	0.089	0.121	0.298	0.8
	(Yeh et al. 2019)	0.237	0.161	0.116	0.088	0.121	0.31	1.015
	Basline (Hori et al. 2019)	0.256	0.161	0.109	0.078	0.113	0.277	0.727
Text Only	(Nguyen et al. 2018)	**0.686**	**0.52**	**0.416**	**0.340**	0.228	**0.518**	0.851
	(Le et al. 2019)	0.633	0.49	0.386	0.31	**0.242**	0.515	**0.856**
	Our Model	0.631	0.478	0.37	0.291	0.224	0.496	0.789
	Basline (Hori et al. 2019)	0.245	0.152	0.103	0.073	0.109	0.271	0.705

Table 1: Model Performance. Models are ranked by an overall performance rather than any single metrics.

	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Video + Text	TF + MaxPool + BiGRU (Baseline)	0.587	**0.438**	0.334	0.261	0.204	0.451	0.684
	+ Data Augmentation ($\times 2$)	0.586	0.436	0.333	**0.262**	**0.206**	**0.46**	**0.704**
	BiGRU $\rightarrow$ BiLSTM	0.585	0.437	**0.335**	0.262	0.202	0.45	0.684
	TF $\rightarrow$ SS	**0.588**	0.437	0.333	0.259	0.197	0.449	0.655
	MaxPool $\rightarrow$ AveragePool	0.378	0.25	0.165	0.111	0.152	0.345	0.3414
	- TF	0.448	0.231	0.124	0.063	0.12	0.353	0.224

Table 2: Ablation Study. "TF" means "Teacher Forcing". "SS" means "Scheduled Sampling". "No-TF" means No Teacher Forcing or Scheduled Sampling. "$\times 2$" means "Data Augmentation by a factor of 2".

Question: What happens in the end of the video?

Summary: A woman is drinking from a glass and she walks to the sink area in the kitchen. she closes a cabinet door and walks away from the kitchen with the glass on her hand.

GT Answer: she look at the camera and walked away.

Text + Video: she is still standing there holding the stove.

Text Only: she walks out of the room

Figure 3: A question whose answer requires the understanding of dynamics in the video. But the **Text + Video** model provides an answer describing all static actions, which shows that the captured feature more focuses on the image individually and doesn't well represent the dependency among the frames. It could partially explain the performance decrease compared to the **Text Only** model.

Q: Does he ever say something?
A: He says something but in a different language.

Text + Video:
No, he is not talking.

Text Only:
No, he does not speak at all.

(a)

Q: What color was his shirt?
A: It is a black jacket and I cannot see his shirt.

Text + Video:
It is a dark color.

Text Only:
It is a blue shirt on the other side of the room.

(b)

Figure 4: **A** is the ground-truth answer to the question **Q**. (a) is an instance where information from audible inputs is directly correlated with the proper response. (b) shows the necessity for proper visual-textual reasoning.

augmentation and BiGRUs.

We are interested in the performance difference between BiLSTM and BiGRU since both are widely used RNN variants in others' work. (Weiss, Goldberg, and Yahav 2018) claims that LSTM with ReLU activation function is strictly stronger for NLP tasks than GRU because of its unbound computational ability; however, our results share more or less identical in terms of the end-goal performance. Given that our initial model has very low CIDEr score of 0.224, we experiment by including teacher forcing. This leads to a notable increase in all our metrics by a range of around $30\% \sim 300\%$. Time wise, we find in our experiments that the model with teacher forcing needs fewer epochs to reach the same performance as without it. Because we find teacher forcing to be an improvement, we also try scheduled sampling. However, we find that it does not improve our scores beyond the improvements from strict teacher forcing. Since scheduled sampling has a certain probability to use the predicted token instead of a ground-truth token as the last token, scheduled sampling could work if our baseline model (TF + MaxPool + GRU) has a fairly high generative accuracy to begin with as there would be less uncertainty during next-token prediction. We find that using the prediction is too noisy and only makes the training procedure less efficient; as well as not being beneficial to cover a limited number of outliers in inference. We could try lowering our probability of picking the predicted word (0.2), but lowering by too much could defeat the purpose of using scheduled sampling.

Once we switch from average pooling to maximum pooling, every evaluation metric increases dramatically, most notably the BLEU and CIDEr scores. This verifies our belief that max pooling includes the dominant signals across all dimensions from the outputs of the BiGRU cells. Lastly, with the inclusion of data augmentation with a factor of 2, our scores increase even further. Therefore, an enlarged dataset through shuffling of the n pairs for the $(n+1)^{th}$ question does result in a quick score enhancement over the baseline. Additionally, this supports our theory that the information within a dialog history matters more than the order it appears in. Given this performance, we will experiment with other factors, such as 4 and 5, in order to find the extent of how much improvement can be made.

Conclusion

In this paper, we evaluate various techniques such as max pooling, use of BiGRU/BiLSTM encoders, teacher forcing/scheduled sampling, and our proposed data augmentation technique on question answering from multimodal data. Our goal was to analyze dialog state tracking through the perspective of QA within the AVSD track of the DSTC8. Through this research, we empirically conclude that our approach performs satisfactorily and improves upon the work of DSTC7. We find that increasing the dataset by a factor of 2 by shuffling dialog history, combined with teacher forcing, max pooling, and GRU encoders, produces the best results within the scope of our tests. Teacher forcing and our proposed technique of shuffling dialog histories result in a substantial improvement over the baseline model and our own tests which do not use these methods. This appears to reinforce our hypothesis that for QA problems, the order of dialog histories is less important than the raw information present within. Additionally, the success derived from the incorporation of teacher forcing suggests that individual tokens within a sentence do have a very important temporal dependence which is critical for generating accurate natural language.

In the future, we would like to conduct further investigation into the fusion of visual and textual data. Specifically, we would like to experiment with approaches pertaining to video-derived modalities, in hope to produce more informative responses with specific details extracted from the videos.

Acknowledgment

This work is supported by Cognitive and Immersive Systems Lab (CISL), a collaboration between IBM and RPI, and also a center in IBM's AI Horizons Network.

References

Alamri, H.; Hori, C.; Marks, T. K.; Batra, D.; and Parikh, D. 2018. Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7. In *DSTC7 at AAAI2019 Workshop*, volume 2.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate.

Bengio, S.; Vinyals, O.; Jaitly, N.; and Shazeer, N. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, 1171–1179.

Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation.

Denkowski, M., and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 376–380. Baltimore, Maryland, USA: Association for Computational Linguistics.

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 *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, 131–135. IEEE.

Hori, C.; Alamri, H.; Wang, J.; Wichern, G.; Hori, T.; Cherian, A.; Marks, T. K.; Cartillier, V.; Lopes, R. G.;

Das, A.; and et al. 2019. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Le, H.; Hoi, S.; Sahoo, D.; and Chen, N. 2019. End-to-end multimodal dialog systems with hierarchical multimodal attention on video features. In *DSTC7 at AAAI2019 workshop*.

Lin, K.-Y.; Hsu, C.-C.; Chen, Y.-N.; and Ku, L.-W. 2019. Entropy-enhanced multimodal attention model for scene-aware dialogue generation. In *DSTC7 at AAAI2019 workshop*.

Lin, C.-Y. 2004. Rouge: a package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Mikolov, T.; Grave, E.; Bojanowski, P.; Puhrsch, C.; and Joulin, A. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Mrkšić, N.; Séaghdha, D. ; Wen, T.-H.; Thomson, B.; and Young, S. 2016. Neural belief tracker: Data-driven dialogue state tracking.

Nguyen, D. T.; Sharma, S.; Schulz, H.; and Asri, L. E. 2018. From film to video: Multi-turn question answering with multi-modal context.

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*, ACL '02, 311–318. Stroudsburg, PA, USA: Association for Computational Linguistics.

Pasunuru, R., and Bansal, M. 2019. Dstc7-avsd: Scene-aware video-dialogue systems with dual attention. In *DSTC7 at AAAI2019 workshop*.

Patel, A.; Sands, A.; Callison-Burch, C.; and Apidianaki, M. 2018. Magnitude: A fast, efficient universal vector embedding utility package. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 120–126.

Sanabria, R.; Palaskar, S.; and Metze, F. 2019. Cmu sinbad's submission for the dstc7 avsd challenge. In *DSTC7 at AAAI2019 workshop*, volume 6.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2014. Cider: Consensus-based image description evaluation. *CoRR* abs/1411.5726.

Wang, S., and Jiang, J. 2016. Machine comprehension using match-lstm and answer pointer.

Weiss, G.; Goldberg, Y.; and Yahav, E. 2018. On the practical computational power of finite precision rnns for language recognition. 740–745.

Wen, T.-H.; Gasic, M.; Mrksic, N.; Su, P.-H.; Vandyke, D.; and Young, S. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

Williams, R. J., and Zipser, D. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation* 1(2):270–280.

Williams, J.; Raux, A.; Ramachandran, D.; and Black, A. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, 404–413. Metz, France: Association for Computational Linguistics.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yeh, Y.-T.; Lin, T.-C.; Cheng, H.-H.; Deng, Y.-H.; Su, S.-Y.; and Chen, Y.-N. 2019. Reactive multi-stage feature fusion for multimodal dialogue modeling. *arXiv preprint arXiv:1908.05067*.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image captioning with semantic attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhuang, B.; Wang, W.; and Shinozaki, T. 2019. Investigation of attention-based multimodal fusion and maximum mutual information objective for dstc7 track3. In *DSTC7 at AAAI2019 workshop*.