DSTC7-AVSD: Scene-Aware Video Dialogue Systems with Dual Attention

Ramakanth Pasunuru
www.rama-kanth.com

Mohit Bansal
www.cs.unc.edu/~mbansal/
Dialogue Context

Multimodal Context

Text

Gaze

Speech

Gesture
Dialogue Context

Text
[Lowe et al., 2015]  
[Vinyals & Le, 2015]  
[Serban et al., 2016]

Gaze
[Deubel & Schneider, 1996]  
[Randall et al., 2014]  
[Wood et al., 2016]

Speech
[Janin et al., 2003]  
[Young, 2005]  
[Celikyilmaz et al., 2015]

Gesture
[Johnston et al., 2002]  
[Cassell, 1999]
Visual Context

Image-based Context

[Das et al., 2017]

[A sample multi-turn dialog. A list of second turn responses.]

[Mostafazadeh et al., 2017]

[De Vries et al., 2017]

[Celikyilmaz et al., 2014]
Abstract

We present a new AI task – Embodied Question Answering (EmbodiedQA) – where an agent is spawned at a random location in a 3D environment and asked a question (e.g., ‘What color is the car?’). In order to answer, the agent must first intelligently navigate to explore the environment, gather necessary visual information through first-person (egocentric) vision, and then answer the question (‘orange’).

EmbodiedQA requires a range of AI skills – language understanding, visual recognition, active perception, goal-driven navigation, commonsense reasoning, long-term memory, and grounding language into actions. In this work, we develop a dataset of questions and answers in House3D environments [1], evaluation metrics, and a hierarchical model trained with imitation and reinforcement learning.

1. Introduction

The embodiment hypothesis is the idea that intelligence emerges in the interaction of an agent with an environment and as a result of sensorimotor activity. Smith and Gasser [2]

Our long-term goal is to build intelligent agents that can perceive their environment (through vision, audition, or other sensors), communicate (i.e., hold a natural language dialog grounded in the environment), and act (e.g., aid humans by executing API calls or commands in a virtual or embodied environment). In addition to being a fundamental scientific goal in artificial intelligence (AI), even a small advance towards such intelligent systems can fundamentally change our lives – from assistive dialog agents for the visually impaired, to natural-language interaction with self-driving cars, in-home robots, and personal assistants.

As a step towards goal-driven agents that can perceive, communicate, and execute actions, we present a new AI task – Embodied Question Answering (EmbodiedQA) – along with a dataset of questions in virtual environments, evaluation metrics, and a deep reinforcement learning (RL) model.

Concretely, the EmbodiedQA task is illustrated in Fig. 1 – an agent is spawned at a random location in an environment (a house or building) and asked a question (e.g., ‘What color is the car?’). The agent perceives its environment through first-person egocentric vision and can perform a few atomic actions (move-forward, turn, strafe, etc.). The goal of the agent is to intelligently navigate the environment and gather visual information necessary for answering the question.
Visual+Audio Context

Person A (Questioner)
1. How many people are in the video?
2. Is he speaking with anyone?
3. What room is he in?
4. What is the man doing?
5. Does he start the video in the hallway?
6. Where does he put the tie and shirt?
7. Does he leave the hallway?
8. Does he open the closet door?
9. Can you tell what he grabs from the closet?
10. Is there anything else I should know?

Person B (Answerer)
1. There is only one man in the video
2. No there is no sound
3. He is in a hallway
4. He is taking off his tie and shirt
5. Yes he starts in the hallway
6. He puts it in a closet
7. After he puts his stuff in the closet he grabs something out of the closet
8. No it is already open
9. He grabs a box and then starts walking toward the camera
10. No that is from start to finish

[Alamri et al., 2018]
Visual+Audio Context

**Task:**

**Input**
- Question
- Video
- Chat History Summary

**Output**
- Answer

---

[Alamri et al., 2018]
Model

What room is he in?

Q: How many people are in the video?
A: There is only one man in the video

Q: Is he speaking with anyone?
A: No there is no sound

There is a person walking in the hallway. He takes off his tie and shirt and puts it in the closet. He then grabs a box from the closet and starts walking to the camera.

Question Encoder

Video Encoder

Chat-History Encoder

Summary Encoder
Model

What room is he in?

Question Encoder

Video Encoder

Chat-History Encoder

Summary Encoder

Cross Attention

What room is he in?

Q: How many people are in the video?
A: There is only one man in the video

Q: Is he speaking with anyone?
A: No there is no sound

There is a person walking in the hallway. He take off his tie and shirt and put it in the closet. He then grabs a box from the closet and starts walking to the camera.

He is in a hallway

[Seo et al., 2017]
Note that we do not use audio features in our models.

[Bahdanau et al., 2015; Seo et al., 2017]
Results

<table>
<thead>
<tr>
<th>Model</th>
<th>METEOR</th>
<th>CIDEr</th>
<th>BLEU-4</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video Only</td>
<td>12.43</td>
<td>95.54</td>
<td>8.83</td>
<td>34.23</td>
</tr>
<tr>
<td>Video + Chat History</td>
<td>14.13</td>
<td>105.39</td>
<td>10.58</td>
<td>36.54</td>
</tr>
<tr>
<td>Video + Chat History + Summary</td>
<td>14.94</td>
<td>112.80</td>
<td>11.22</td>
<td>37.53</td>
</tr>
<tr>
<td>Video + Chat History + Summary + Cross-attention</td>
<td>14.95</td>
<td>115.82</td>
<td>11.38</td>
<td>37.87</td>
</tr>
</tbody>
</table>

Our models’ performance on AVSD dataset’s public test set. All of these models use the question information.
Results

<table>
<thead>
<tr>
<th>Model</th>
<th>METEOR</th>
<th>CIDEr</th>
<th>BLEU-4</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video Only</td>
<td>12.43</td>
<td>95.54</td>
<td>8.83</td>
<td>34.23</td>
</tr>
<tr>
<td>Video + Chat History</td>
<td>14.13</td>
<td>105.39</td>
<td>10.58</td>
<td>36.54</td>
</tr>
<tr>
<td>Video + Chat History + Summary</td>
<td>14.94</td>
<td>112.80</td>
<td>11.22</td>
<td>37.53</td>
</tr>
<tr>
<td>Video + Chat History + Summary + Cross-attention</td>
<td>14.95</td>
<td>115.82</td>
<td>11.38</td>
<td>37.87</td>
</tr>
</tbody>
</table>

Our models’ performance on AVSD dataset’s public test set. All of these models use the question information (no audio information).
Other Methods

• Policy gradient based reinforcement learning

• Contextualized ELMo word embeddings

• Using external data

• Pointer-generator copy model
Policy Gradients

\[
\nabla_\theta L(\theta) = -\mathbb{E}_{w^s \sim p_\theta} [r(w^s) \cdot \nabla_\theta \log p_\theta(w^s)]
\]

[Williams, 1992]
Contextualized ELMo Word Embeddings

Q: How many people are in the video?  
A: There is only one man in the video

Q: Is he speaking with anyone?  
A: No there is no sound

There is a person walking in the hallway. He take off his tie and shirt and put it in the closet. He then grabs a box from the closet and starts walking to the camera.

He is in a hallway

[ELMo] [ELMo] [ELMo]

Similarity matrix

Cross Attention

[ELMo]
[ELMo]

[Q Peters et al., 2018]
Using External Data (MSR-VTT)

Q: How many people are in the video?
A: There is only one man in the video.

Q: Is he speaking with anyone?
A: No, there is no sound.

There is a person walking in the hallway. He takes off his tie and shirt and puts them in the closet. He then grabs a box from the closet and starts walking towards the camera.

He is in a hallway.

[Xu et al., 2016]
Using External Data (MSR-VTT)

Q: How many people are in the video?
A: There is only one man in the video

Q: Is he speaking with anyone?
A: No there is no sound

There is a person walking in the hallway. He takes off his tie and shirt and put it in the closet. He then grabs a box from the closet and starts walking to the camera.

He is in a hallway

[Xu et al., 2016]
Pointer-generator Copy Model

• Lot of words in the question can also be present in the answer
• The final word distribution is a weighted combination of the vocab distribution and attention distribution
• Question-based pointer
• Joint question- and summary-based pointer

[See et al., 2017]
Future Work

• Further analyze and improve these promising approaches with specific RL rewards, contextualized large language models, and joint copy models

• We will add Audio features to our final model

• Effective ways of extending cross-attention to multiple modalities (question+summary; question+chat-history)
Thanks!

Acknowledgment: DARPA YFA17-D17AP00022, Faculty awards from Google, Facebook, and Salesforce