CMU Sinbad’s Submission for the DSTC7 AVSD Challenge

Ramon Sanabria*, Shruti Palaskar* and Florian Metze

Language Technologies Institute
Carnegie Mellon University

*equal contribution
Task Description

- Generate system responses in a dialog about an input video.
- Dialog systems need to understand scenes.
- Task:
  - Visual question answering (VQA).
  - Video description.
Investigating Different Visual Features
Visual Features

Place/Scene Recognition
- [http://places2.csail.mit.edu/download.html](http://places2.csail.mit.edu/download.html)
- **Source:** Image.
- **Target:** Scene Context.

Object Recognition
- **Source:** Image.
- **Target:** Object.

Action Features
- Hara et al. 2018
- **Source:** Video (16 frames).
- **Target:** Action.

riding arena

Tree

Arranging flowers
DSTC7 Baseline

Model Details:
- Sequence-to-Sequence model.
- Simple concatenation.
- 2 layers 128 units in encoder.
- 2 layers 128 units.
- Alamri et al. 2018.
## Feature Comparison

<table>
<thead>
<tr>
<th></th>
<th>Bleu-4</th>
<th>Rouge-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.084</td>
<td>0.291</td>
</tr>
<tr>
<td>Place</td>
<td>0.082</td>
<td>0.286</td>
</tr>
<tr>
<td>Object</td>
<td>0.083</td>
<td>0.287</td>
</tr>
<tr>
<td>Action</td>
<td>0.079</td>
<td>0.284</td>
</tr>
<tr>
<td>All</td>
<td><strong>0.085</strong></td>
<td>0.287</td>
</tr>
</tbody>
</table>

- Results of proto set.
Our models
Model

- Sequence-to-Sequence model with attention (Bahdanau et al. 2015)
- Bidirectional GRU encoder decoder
- 2 layers 256 units in encoder
- 2 layers 256 units of conditional GRU decoder
Text-only Model

Answer

Attentive Decoder

2-layer BiGRU

Text Encoder

Summary + Question

ROUGE-L: 0.330

BLEU-4: 0.105
Action-only Model

Decoder

ResNeXt
action prediction

2-layer BiGRU

Answer

BLEU-4
0.085

ROUGE-L
0.294
Hierarchical Attention for Text + Video

- Text Encoder
- ResNeXt action prediction
- Multimodal Decoder
- Text Encoder
- Vocabulary

Libovicky et al. 2017

- BLEU-4: 0.112
- ROUGE-L: 0.338
# All models, Comparison

![Table showing comparison of models]

<table>
<thead>
<tr>
<th></th>
<th>DSTC</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bleu-4</td>
<td>Rouge-L</td>
</tr>
<tr>
<td>Text-only</td>
<td></td>
<td>0.105</td>
<td>0.330</td>
</tr>
<tr>
<td>Video-only</td>
<td></td>
<td>0.085</td>
<td>0.294</td>
</tr>
<tr>
<td>Text + Video</td>
<td></td>
<td>0.112</td>
<td>0.338</td>
</tr>
</tbody>
</table>

- Video-only competitive with other models
- Text + Video improves over Text-only (but not drastically)
Investigating use of external data
External Data: How2 Dataset

I’m very close to the green but I didn’t get it on the green so now I’m in this grass bunker.

Eu estou muito perto do green, mas eu não pus a bola no green, então agora estou neste bunker de grama.

In golf, get the body low in order to get underneath the golf ball when chipping out of thick grass from a side hill lie.

- Video
- Speech
- English Transcript
- Portuguese Transcript
- Summary

Sanabria et al. 2018
Model Fine-tuned with How2 Data

Summary

Answer

Transcript
Video
Transcript + Video
Audio
Translation

Summary + Question
Video
Audio
Caption
Summarization with How2 Data

- **Summarization**
  - Present subset of information in a more compact form (maybe across modalities)

- **“Description” field**
  - 2-3 sentences of meta-data: template based, uploader provides
  - “Informative” and abstractive summary of a how-to video
  - Should generate interest of a potential viewer

Libovicky et al. 2018
Attention over the Video Features

Talking and preparing the brush

Close-up of brushstrokes with hand

Close-up of brushstrokes no hand

Black frames at the end
Text-only Model Fine-tuned with How2

Transcript \rightarrow Summary

Summary + Question \rightarrow Answer

Answer

Attentive Decoder

2-layer BiGRU

Text Encoder

BLEU-4

0.114

ROUGE-L

0.3337
Video-only Model Fine-tuned with How2

- Video -> Answer
- Video -> Summary

- ResNeXt action prediction
- Bidirectional RNN
- Decoder
- Vocabulary

- BLEU-4: 0.086
- ROUGE-L: 0.300
Text + Video Model Fine-tuned with How2

Multimodal Decoder

Vocabulary

ResNeXt action prediction

Text Encoder

BLEU-4
0.113

ROUGE-L
0.339

Libovicky et al. 2017
## All models, Comparison + Human Ratings

<table>
<thead>
<tr>
<th></th>
<th>DSTC only</th>
<th>DSTC + How2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bleu-4</td>
<td>Rouge-L</td>
</tr>
<tr>
<td>Text-only</td>
<td>0.105</td>
<td>0.330</td>
</tr>
<tr>
<td>Video-only</td>
<td>0.085</td>
<td>0.294</td>
</tr>
<tr>
<td>Text + Video</td>
<td><strong>0.112</strong></td>
<td><strong>0.338</strong></td>
</tr>
</tbody>
</table>

- How2 data improves Text-only model
- Video-only competitive with other models
- Text + Video improves over Text-only but not drastically
- Our model performs well on human evaluation as well
<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>is he talking or reading out loud?</td>
<td>no, he is not talking at all.</td>
</tr>
<tr>
<td>what 's in the mug?</td>
<td>i don't know, i can't see the inside.</td>
</tr>
<tr>
<td>hello. did someone come to the door?</td>
<td>no and it is a window that he is standing in front of.</td>
</tr>
<tr>
<td>are they talking in the video?</td>
<td>not really no i don't hear anything</td>
</tr>
</tbody>
</table>
References

- “Attention Strategies for Multi-Source Sequence-to-Sequence Learning”, Jindřich Libovický, Jindřich Helcl
- “How2: A Large-scale Dataset for Multimodal Language Understanding”, Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliot, Loic Barrault, Lucia Specia, Florian Metze
- “Multimodal Abstractive Summarization for Open-Domain Videos”, Jindrich Libovicky, Shruti Palaskar, Spandana Gella, Florian Metze
Thank you to the organizers!

Data  https://github.com/srvk/how2-dataset

Code  https://github.com/lium-lst/nmtpytorch