Multimodal Abstractive Summarization for Open-Domain Videos

Jindřich Libovický, Shruti Palaskar, Spandana Gella, Florian Metze
spalaska@cs.cmu.edu
Charles University, Carnegie Mellon University, University of Edinburgh

Introduction

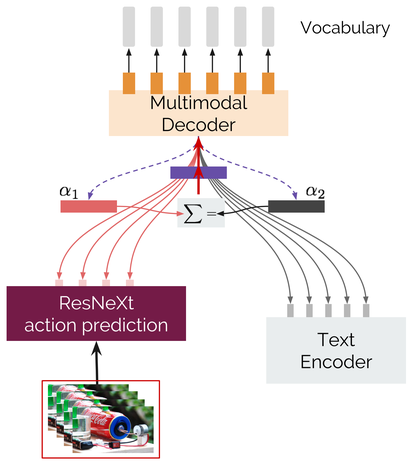
- Introducing Abstractive Summarization for Open-Domain Videos
- Provide **fluent textual summary** from multimodal information
- The **How2 corpus** [1] of instructional videos, transcripts and summaries is useful for this task
- **Propose Content-F1**: a more informative measure of multimodal abstractive summaries

The How2 dataset for Summarization

- 2000 hours of short instructional videos
- Many different topics like cooking, sports, music...
- Human annotated summary although somewhat template-like
- Summary is the description in video meta-data

Models

- Sequence-to-Sequence Model for summarization
- Hierarchical attention [2] for multi-modality



- Using action features as video representations
- Text-only, Video-only, Text-and-Video models

Using How2 dataset for Summarization

Transcript:
Today we are going to show you how to make Spanish omelet. I'm going to dice a little bit of peppers here. I'm not going to use a lot ... You can use red peppers if you like to get a little bit color in your omelet. Some people do and some people don't ... You are going to take the onion also and dice it really small. You don't want big chunks of onion in there cause it is just pops out of the omelet ... So we have small pieces of onions and peppers ready to go.

Summary:
how to cut peppers to make a spanish omelette; get expert tips and advice on making cuban breakfast recipes in this free cooking video .

- Transcript is conversational and has many more details about the procedure
- Summary is a high-level overview of entire video
- Text and Vision modalities contain complementary information

Results

Model No.	Description	ROUGE-L	Content F1
1	Random Baseline using Language Model	27.5	8.3
2a	Rule-based Extractive summary	16.4	18.8
2b	Next-neighbor summary	31.8	17.9
3	Using extracted sentence from 2a only (Text-only)	46.4	36
4	First 200 tokens (Text-only)	40.3	27.5
5	Complete Transcript (Text-only, 650 tokens)	**53.9**	**47.4**
6	Action Features only (Video)	38.5	24.8
7	Action Features + RNN (Video)	**46.3**	**34.9**
8	Text + Action with Hierarchical Attn	**54.9**	**48.9**
9	Text + Action RNN with Hierarchical Attn	53.4	46.8

Model 1	Random baseline is a language model trained only on the summaries
Model 2a	Rule-based extracted summary for extractive baseline, generally second sentence from transcript is chosen as target
Model 2b	Target summary replaced by the nearest neighbor summary to test generalizability
Model 7	Video-only model performs competitive with text-only model
Model 8	Hierarchical attention for text-and-video improves over text-only

Content-F1 Score

- Summary is a high-level overview of the video
- ROUGE is often high due to repetitive catch-phrases: *learn from expert, in this free video, get tips from professional etc.*
- Content-F1 computes F1 score of only content words (zero weight to function words)
- Content-F1 ignores fluency; ROUGE prefers style

Output Examples

Model	Output
Reference	watch and learn how to tie thread to a hook to help with fly tying as explained by out expert in this free how - to video on **fly tying tips and techniques** .
Random Baseline	learn tips on how to **play the bass drum** beat variation on the guitar in this free video clip on **music theory** and guitar lesson .
Text-only	learn from our expert **how to tie a thread for fly fishing** in this free how - to video on fly tying tips and techniques .
Action Features + RNN	learn about the **equipment needed for fly tying** , as well as other fly fishing tips from our expert in this free how - to video on fly tying tips and techniques .
Hierarchical Attention	learn from our expert how to **attach thread to fly fishing for fly fishing** in this free how - to video on fly tying tips and techniques .

DSTC-7 AVSD

- Applied these models to DSTC7 AVSD challenge
- Ranked first in objective and human evaluation
- Additionally, we pretrained models on the How2 data and observed moderate gains in DSTC data

References

[1] Sanabria et al., How2: a large-scale dataset for multimodal language understanding, NeurIPS ViGIL 2018
[2] Libovický & Helcl, Attention strategies for multi-source sequence-to-sequence learning, ACL 2017