Multimodal Abstractive Summarization for How2 Videos

Shruti Palaskar[1], Jindřich Libovický[2], Spandana Gella[3], Florian Metze[1]
spalaska@cs.cmu.edu, libovicky@ufal.mff.cuni.cz
[1]Carnegie Mellon University, [2]Charles University, [3]Amazon AI

Language Technologies Institute

FACULTY OF MATHEMATICS AND PHYSICS Charles University

Introduction

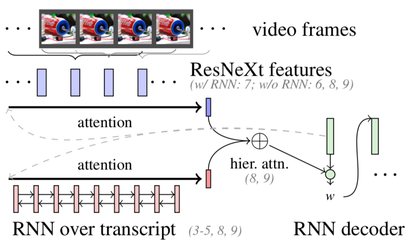
- Introducing Abstractive Summarization for Open-Domain Videos
- Provide **fluent textual summary** from multimodal information
- The **How2 corpus** [1] of instructional videos, transcripts and summaries is useful for this task
- **Propose Content-F1**: a more informative measure of multimodal abstractive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

The How2 dataset

- 2000 hours of short instructional videos
- Different topics like cooking, sports, music...
- Manual summaries although somewhat template-like
- Summary is the description in video meta-data
- Contains speech, text and video frames

Multimodal Hierarchical Attention

- Hierarchical attention [2] for multi-modality



- Action features as video representations
- Text-only, Video-only, Text-and-Video models

Transfer Learning

- Pre-training on the How2 data summarization task and fine-tuning on the Charades dataset video question answering task led to moderate gains
- Charades dataset is a multimodal dataset with audio, video, questions, answers and summary

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 and Model Analysis

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	36.0
4	First 200 tokens (Text-only)	40.3	27.5
5a	S2S Complete Transcript (Text-only, 650 tokens)	**53.9**	**47.4**
5b	PG Complete Transcript (Text-only)	50.2	42.0
5c	ASR output Complete Transcript (Text-only)	46.1	34.7
6	Action Features only (Video)	38.5	24.8
7	Action Features + RNN (Video)	**46.3**	**34.9**
8	Ground-truth transcript + Action with Hierarchical Attn	**54.9**	**48.9**
9	ASR output + Action with Hierarchical Attn	46.3	34.7

Model 1-2b Different baseline models
Model 3-5c Different text-only models, with and without speech input
Model 6-7 Video-only models without text input
Model 8-9 Multimodal models, with and without speech input

Content-F1 Score

- ROUGE is often high due to repetitive catch-phrases: *learn from expert, free video, tips from professional*
- Content-F1 computes F1 score of only content words (no function words), ignores fluency

Output Analysis

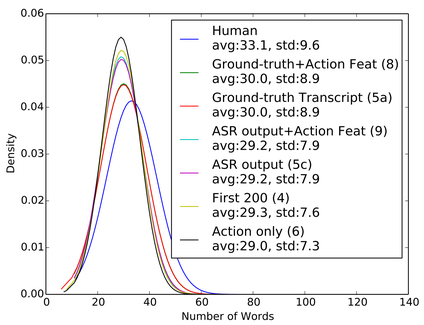


Figure: Word distribution of different models

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 ...
Hierarchical Attention	learn from our expert how to **attach thread to fly fishing for fly fishing** in this free ...

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