Multimodal Abstractive Summarization for How2 Videos

ACL19

Shruti Palaskar、Jindrˇich Libovický、Spandana Gella、Florian Metze

School of Computer Science, Carnegie Mellon University
Faculty of Mathematics and Physics, Charles University
Amazon AI

Xiachong Feng

Outline

- Author
- Background
- Task
- Dataset
- Metric
- Experiment

Author

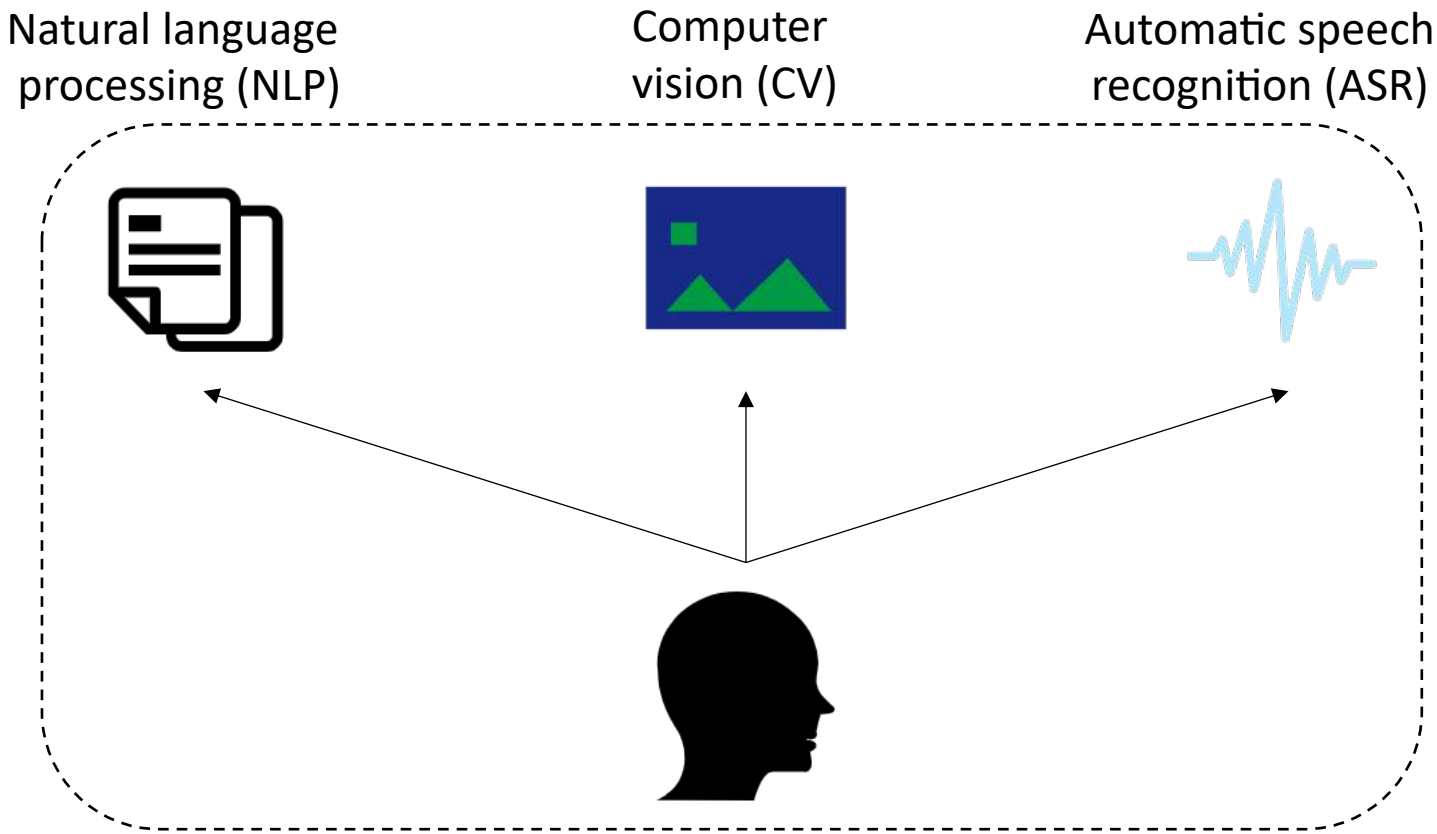
Shruti Palaskar

- **PhD student** at the Language Technologies Institute of the School of Computer Science at **Carnegie Mellon University**.
- **multimodal machine learning**, speech recognition and natural language processing

Updates

[Mar 2019]	I will be giving a talk about our work on Multimodal Acoustic Word Embeddings at the 6th Amazon Graduate Student Symposium in Seattle. Slides here!
[Feb 2019]	We will be holding the How2 Challenge and Workshop at ICML 2019. If you work on anything multimodal, hope to see you there!
[Jan 2019]	Come check out the special session on *Multimodal Representation Learning for Language Generation and Understanding* at ICASSP 2019.
[Dec 2018]	Received the Facebook Fellowship for academic years 2019-2021. Thank you Facebook!
[Nov 2018]	The How2 dataset of open-domain instructional videos has been released! Check it out!
[Nov 2018]	Our paper on Multimodal Abstractive Summarization has been accepted at the NeurIPS 2018 ViGIL workshop for **Spotlight** presentation!
[Oct 2018]	Ramon and I won the **first place** in the audio-visual track of DSTC7. We will present this at AAAI 2019 in Hawaii.
[Sep 2018]	**PhD student panelist** at the Young Female Researchers in Speech Workshop at Interspeech 2018
[Sep 2018]	Our paper on Acoustic-to-Word Speech Recognition is accepted at SLT 2018
[Jul 2018]	Received the 2018-2019 Center for Machine Learning and Health **PhD Fellowship**. Thank you CMLH!
[Sep 2016]	Received the CMU LTI **Graduate Research Fellowship** for acamedic years 2016-2018

Background



Human information processing is inherently multimodal, and language is best understood in a situated context.

Task

- Multimodal summarization
  - Video summarization
  - Text 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 , i 'm going to use very very little . a little bit more then this maybe . you can use red peppers if you like to get a little bit color in your omelet . some people do and some people do n't …. t is the way they make there spanish omelets that is what she says . i loved it , it actually tasted really good . you are going to take the onion also and dice it really small . you do n't want big chunks of onion in there cause it is just pops out of the omelet . so we are going to dice the up also very very small . so we have small pieces of onions and peppers ready to go .

Video

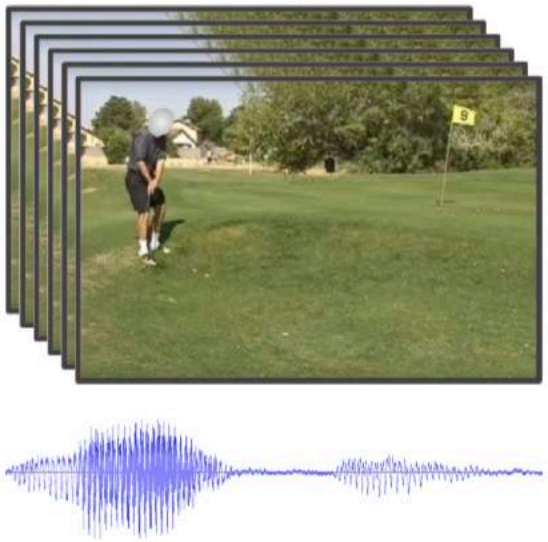
Summary

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

Figure 1: How2 dataset example with different modalities. "Cuban breakfast" and "free cooking video" is not mentioned in the transcript, and has to be derived from other sources.

Search and Retrieve Relevant Videos

Dataset-How2



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.

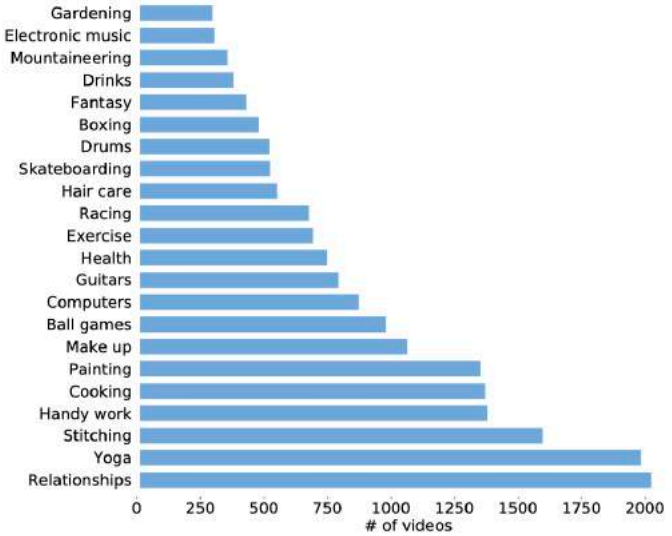
In golf, get the body low in order to get underneath the golf ball when chipping out of thick grass from a side hill lie.

Figure 1: How2 contains a large variety of instructional videos with utterance-level English subtitles (in bold), aligned Portuguese translations (in italics), and video-level English summaries (in the box). Multimodality helps resolve ambiguities and improves understanding.

Dataset

- 2,000 hours of short instructional videos, spanning different domains such as cooking, sports, indoor/outdoor activities, music, etc.
- Each video is accompanied by a human-generated transcript and a 2 to 3 sentence summary

Training	73993
Validation	2965
Testing	2156
Input avg	291 words
Summary avg	33 words



(a) Topic distribution.

Model

- Video-based Summarization
- Speech-based Summarization

Video-based Summarization

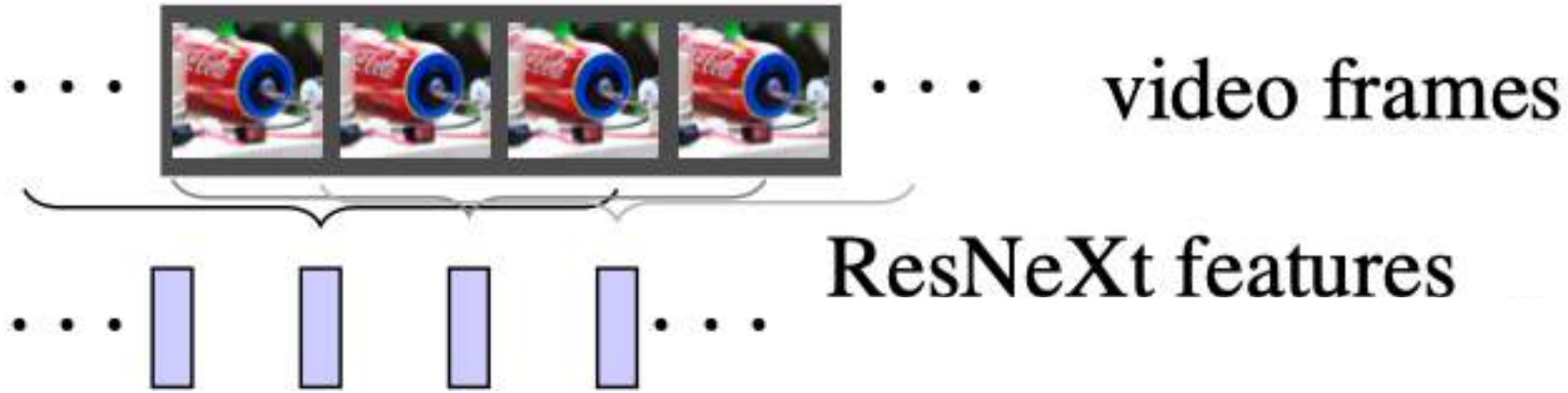
- **Pre-trained action recognition model**: a ResNeXt-101 3D Convolutional Neural Network
- Recognize 400 different human actions

Actions

(g) riding a bike

Video-based Summarization

- 2048 dimensional, extracted for every 16 non-overlapping frames



Speech-based Summarization

- Pretrained speech recognizer
- use the state-of-the-art models for distant-microphone conversational speech recognition, ASpIRE and EESEN.

Summarization Models

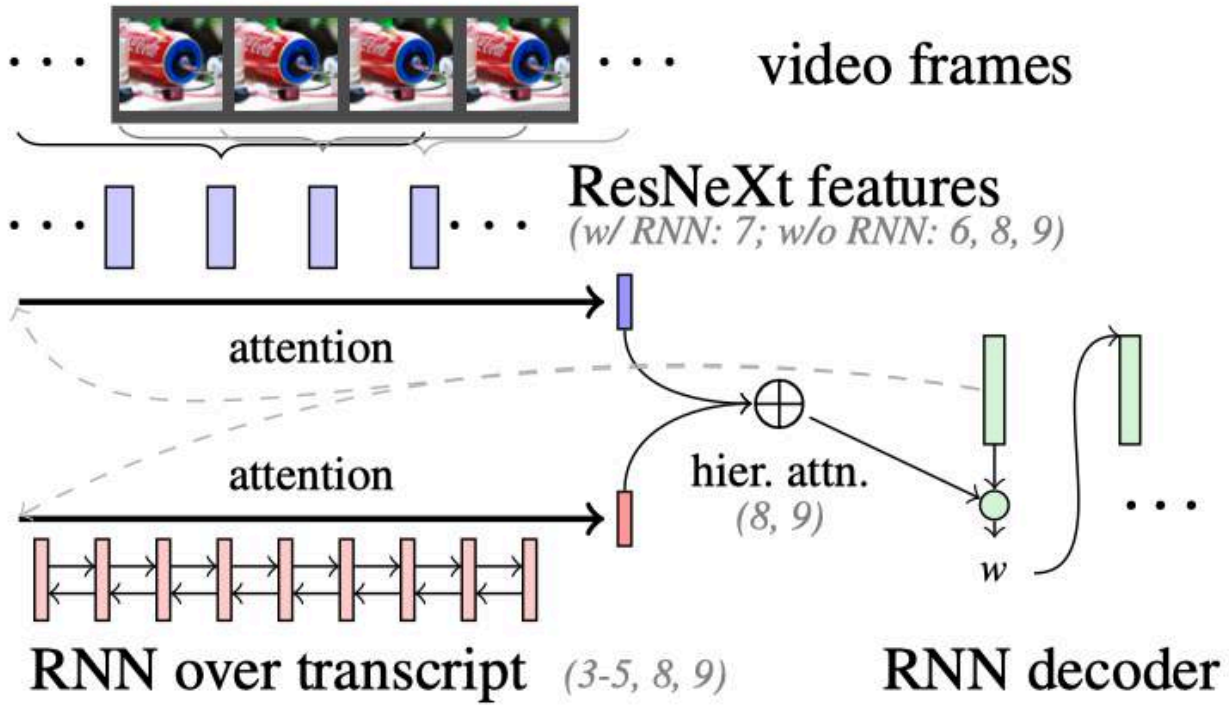


Figure 2: Building blocks of the sequence-to-sequence models, gray numbers in brackets indicate which components are utilized in which experiments.

Content F1

1. Use the METEOR toolkit to obtain the alignment between *ref* and *gen.*
2. Remove function words and task-specific stop words.
3. F1 score over the alignment.

Experiment

- RNN language model on all the summaries and randomly sample tokens from it.
- The output obtained is fluent in English leading to a high ROUGE score, but the content is unrelated which leads to a low Content F1 score

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

Table 1: ROUGE-L and Content F1 for different summarization models: random baseline (1), rule-based extracted summary (2a), nearest neighbor summary (2b), different text-only (3,4,5a), pointer-generator (5b), ASR output transcript (5c), video-only (6-7) and text-and-video models (8-9).

Experiment

- Sentence containing words “how to” with predicates *learn, tell, show, discuss or explain*, usually the second sentence in the transcript.

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

Table 1: ROUGE-L and Content F1 for different summarization models: random baseline (1), rule-based extracted summary (2a), nearest neighbor summary (2b), different text-only (3,4,5a), pointer-generator (5b), ASR output transcript (5c), video-only (6-7) and text-and-video models (8-9).

Experiment

- trained with the summary of the nearest neighbor of each video in the Latent Dirichlet Allocation (LDA) based topic space as a target.

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

Table 1: ROUGE-L and Content F1 for different summarization models: random baseline (1), rule-based extracted summary (2a), nearest neighbor summary (2b), different text-only (3,4,5a), pointer-generator (5b), ASR output transcript (5c), video-only (6-7) and text-and-video models (8-9).

Experiment

- The text-only model performs best when using the complete transcript in the input (650 tokens).
- This is in contrast to prior work with news-domain summarization.

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

Table 1: ROUGE-L and Content F1 for different summarization models: random baseline (1), rule-based extracted summary (2a), nearest neighbor summary (2b), different text-only (3,4,5a), pointer-generator (5b), ASR output transcript (5c), video-only (6-7) and text-and-video models (8-9).

Experiment

- PG networks do not perform better than S2S models on this data which could be attributed to the abstractive nature of our summaries and also the lack of common n-gram overlap between input and output which is the important feature of PG networks
- ASR: degrades noticeably

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

Table 1: ROUGE-L and Content F1 for different summarization models: random baseline (1), rule-based extracted summary (2a), nearest neighbor summary (2b), different text-only (3,4,5a), pointer-generator (5b), ASR output transcript (5c), video-only (6-7) and text-and-video models (8-9).

Experiment

- almost competitive ROUGE and Content F1 scores compared to the text-only model showing the importance of both modalities in this task.

single mean-pooled feature vector

sequence of feature vectors

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

Table 1: ROUGE-L and Content F1 for different summarization models: random baseline (1), rule-based extracted summary (2a), nearest neighbor summary (2b), different text-only (3,4,5a), pointer-generator (5b), ASR output transcript (5c), video-only (6-7) and text-and-video models (8-9).

Experiment

- Hierarchical attention model that combines both modalities obtains the highest score.

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

Table 1: ROUGE-L and Content F1 for different summarization models: random baseline (1), rule-based extracted summary (2a), nearest neighbor summary (2b), different text-only (3,4,5a), pointer-generator (5b), ASR output transcript (5c), video-only (6-7) and text-and-video models (8-9).

Human Evaluation

- Informativeness, relevance, coherence, and fluency

Model (No.)	INF	REL	COH	FLU
Text-only (5a)	3.86	**3.78**	3.78	3.92
Video-only (7)	3.58	3.30	3.71	3.80
Text-and-Video (8)	**3.89**	3.74	**3.85**	**3.94**

Table 2: Human evaluation scores on 4 different measures of Informativeness (INF), Relevance (REL), Coherence (COH), Fluency (FLU).

Word distributions

- very similar in length showing that the improvements in Rouge-L and Content-F1 scores stem from the difference in content rather than length.

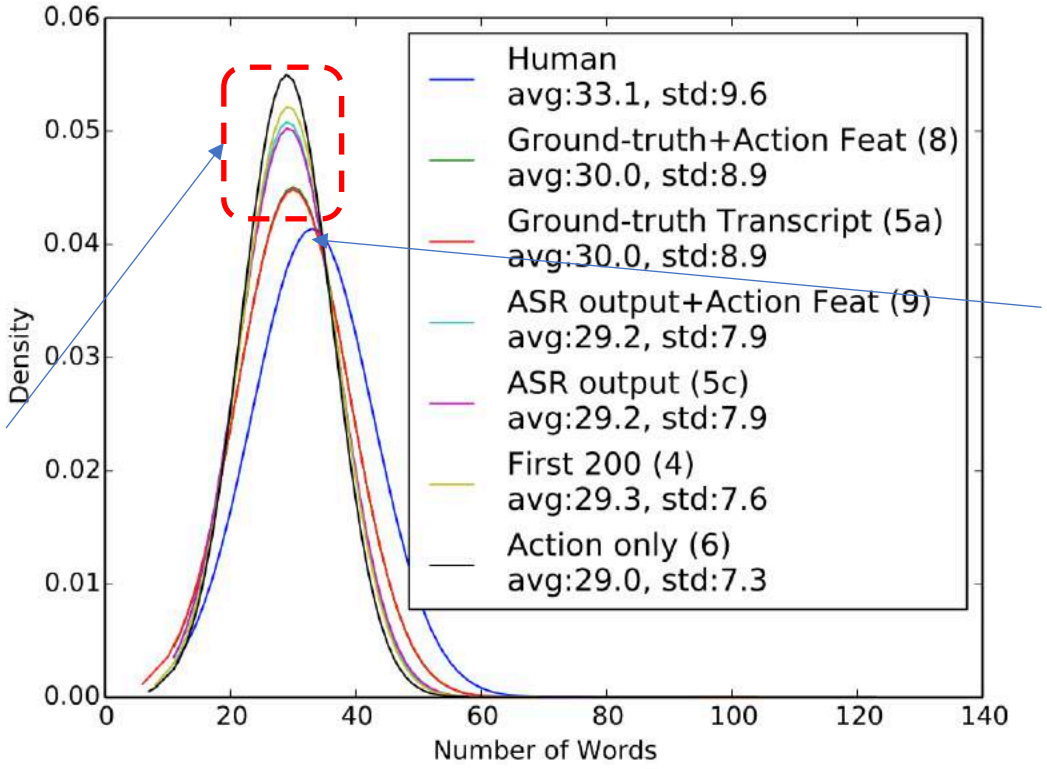


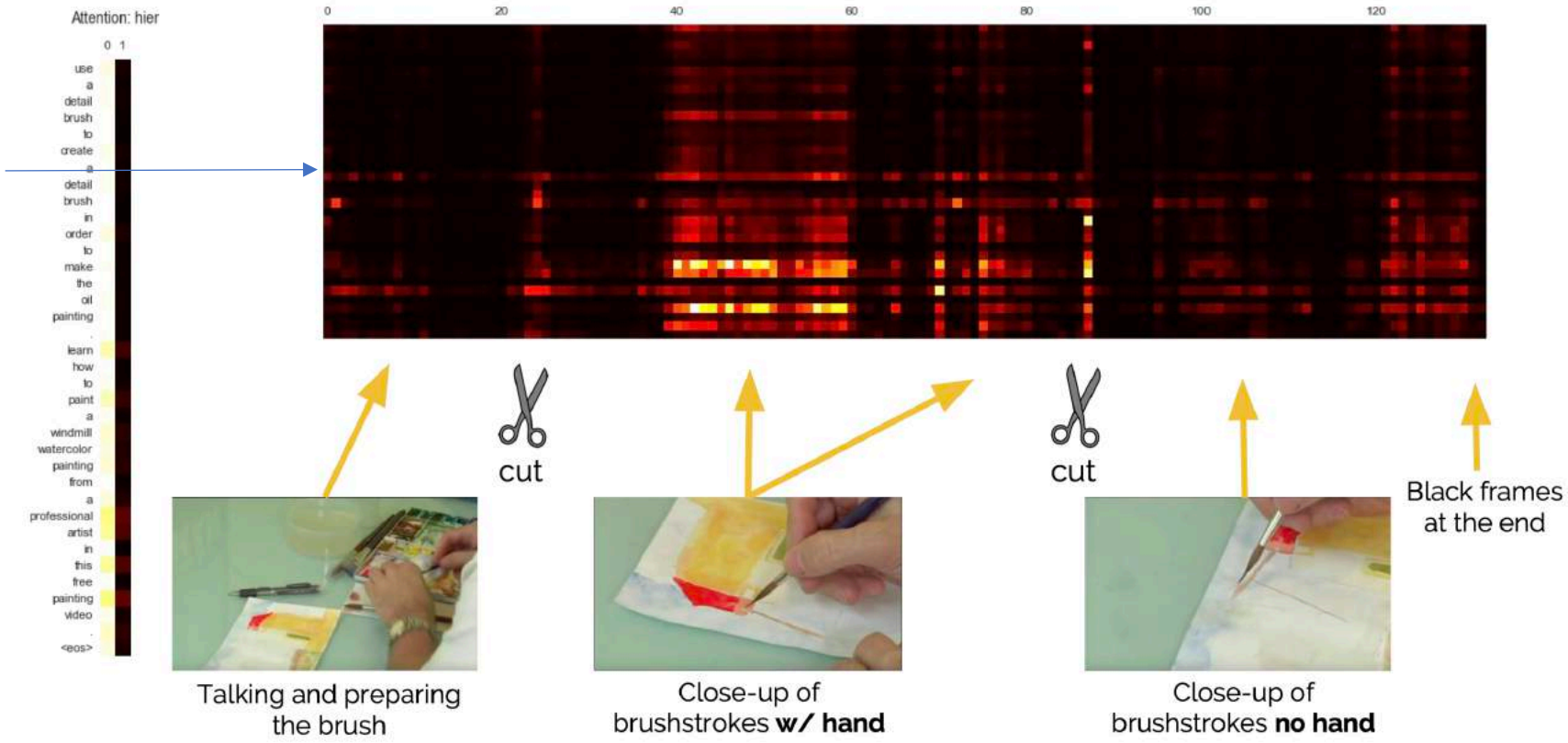
Figure 3: Word distribution in comparison with the human summaries for different unimodal and multimodal models. Density curves show the length distributions of human annotated and system produced summaries.

- most model outputs are shorter than human annotations

Attention Analysis-painting.

input time-steps (from the transcript).

output summary of the model



- less attention in the first part of the video where the speaker is introducing the task and preparing the brush.
- the camera focuses on the close-up of brush strokes with hand, model pays higher attention over consecutive frames.
- the close up does not contain the hand but only the paper and brush, less attention which could be due to unrecognized actions in the close-up.

Case Study

No.	Model	R-L	C-F1	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 .
8	Ground-truth text + Action Feat.	54.9	48.9	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 .
5a	Text-only (Ground-truth)	53.9	47.4	learn from our expert how to tie a thread for fly fishing in this free how - to video on fly tying tips and techniques .
9	ASR output + Action Feat.	46.3	34.7	learn how to tie a fly knot for fly fishing in this free how-to video on fly tying tips and techniques .
5c	ASR output	46.1	34.7	learn tips and techniques for fly fishing in this free fishing video on techniques for and making fly fishing nymphs .
7	Action Features + RNN	46.3	34.9	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 .
6	Action Features only	38.5	24.8	learn from our expert how to do a double half hitch knot in this free video clip about how to use fly fishing .
2b	Next Neighbor	31.8	17.9	use a sheep shank knot to shorten a long piece of rope . learn how to tie sheep shank knots for shortening rope in this free knot tying video from an eagle scout .
1	Random Baseline	27.5	8.3	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 .

Table A2: Example outputs of ground-truth text-and-video with hierarchical attention (8), text-only with ground-truth (5a), text-only with ASR output (5c), ASR output text-andv-video with hierarchical attention (9), action features with RNN (7) and action features only (6) models compared with the reference, the topic-based next neighbor (2b) and random baseline (1). Arranged in the order of best to worst summary in this table.

Thanks!