Multimodal Abstractive Summarization for How2 Videos

Shruti Palaskar[1] **Jindřich Libovický**[2] **Spandana Gella**[3]* **Florian Metze**[1]
[1]School of Computer Science, Carnegie Mellon University
[2]Faculty of Mathematics and Physics, Charles University
[3] Amazon AI
spalaska@cs.cmu.edu, libovicky@ufal.mff.cuni.cz
sgella@amazon.com, fmetze@cs.cmu.edu

Abstract

In this paper, we study abstractive summarization for open-domain videos. Unlike the traditional text news summarization, the goal is less to "compress" text information but rather to provide a fluent textual summary of information that has been collected and fused from different source modalities, in our case video and audio transcripts (or text). We show how a multi-source sequence-to-sequence model with hierarchical attention can integrate information from different modalities into a coherent output, compare various models trained with different modalities and present pilot experiments on the *How2 corpus* of instructional videos. We also propose a new evaluation metric (Content F1) for abstractive summarization task that measures semantic adequacy rather than fluency of the summaries, which is covered by metrics like ROUGE and BLEU.

1 Introduction

In recent years, with the growing popularity of video sharing platforms, there has been a steep rise in the number of user-generated instructional videos shared online. With the abundance of videos online, there has been an increase in demand for efficient ways to search and retrieve relevant videos (Song et al., 2011; Wang et al., 2012; Otani et al., 2016; Torabi et al., 2016). Many cross-modal search applications rely on text associated with the video such as description or title to find relevant content. However, often videos do not have text meta-data associated with them or the existing ones do not provide clear information of the video content and fail to capture subtle differences between related videos (Wang et al., 2012). We address this by aiming to generate a short text summary of the video that describes the most salient content of the video. Our work benefits users through better contextual information and user experience, and video sharing platforms with increased user engagement by retrieving or suggesting relevant videos to users and capturing their attention.

Summarization is a task of producing a shorter version of the content in the document while preserving its information and has been studied for both textual documents (automatic text summarization) and visual documents such as images and videos (video summarization). Automatic text summarization is a widely studied topic in natural language processing (Luhn, 1958; Kupiec et al., 1995; Mani, 1999); given a text document the task is to generate a textual summary for applications that can assist users to understand large documents. Most of the work on text summarization has focused on single-document summarization for domains such as news (Rush et al., 2015; Nallapati et al., 2016; See et al., 2017; Narayan et al., 2018) and some on multi-document summarization (Goldstein et al., 2000; Lin and Hovy, 2002; Woodsend and Lapata, 2012; Cao et al., 2015; Yasunaga et al., 2017).

Video summarization is the task of producing a compact version of the video (visual summary) by encapsulating the most informative parts (Money and Agius, 2008; Lu and Grauman, 2013; Gygli et al., 2014; Song et al., 2015; Sah et al., 2017). Multimodal summarization is the combination of textual and visual modalities by summarizing a video document with a text summary that summarizes the content of the video. Multimodal summarization is a more recent challenge with no benchmarking datasets yet. Li et al. (2017) collected a multimodal corpus of 500 English news videos and articles paired with manually annotated summaries. The dataset is small-scale and has news articles with audio, video, and text summaries, but there are no human annotated audio-transcripts.

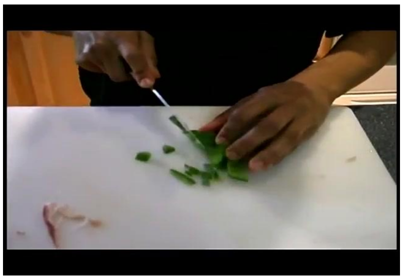
*Work done while SG was at University of Edinburgh

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6587–6596
Florence, Italy, July 28 - August 2, 2019.

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.

Related tasks include image or video captioning and description generation, video story generation, procedure learning from instructional videos and title generation which focus on events or activities in the video and generating descriptions at various levels of granularity from single sentence to multiple sentences (Das et al., 2013; Regneri et al., 2013; Rohrbach et al., 2014; Zeng et al., 2016; Zhou et al., 2018; Zhang et al., 2018; Gella et al., 2018). A closely related task to ours is video title generation where the task is to describe the most salient event in the video in a compact title that is aimed at capturing users attention (Zeng et al., 2016). Zhou et al. (2018) present the YouCookII dataset containing instructional videos, specifically cooking recipes, with temporally localized annotations for the procedure which could be viewed as a summarization task as well although localized with time alignments between video segments and procedures.

In this work, we study multimodal summarization with various methods to summarize the intent of *open-domain instructional videos* stating the exclusive and unique features of the video, irrespective of modality. We study this task in detail using the new How2 dataset (Sanabria et al., 2018) which contains human annotated video summaries for a varied range of topics. Our models generate natural language descriptions for video content using the transcriptions (both user-generated and output of automatic speech recognition systems) as well as visual features extracted from the video. We also introduce a new evaluation metric (Content F1) that suits this task and present detailed results to understand the task better.

2 Multimodal Abstractive Summarization

The How2 dataset (Sanabria et al., 2018) contains about 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 is available for every video written to generate interest in a potential viewer.

The example in Figure 1 shows the transcript describes instructions in detail, while the summary is a high-level overview of the entire video, mentioning that the peppers are being “cut”, and that this is a “Cuban breakfast recipe”, which is not mentioned in the transcript. We observe that text and vision modalities both contain complementary information, thereby when fused, helps in generating richer and more fluent summaries. Additionally, we can also leverage the speech modality by using the output of a speech recognizer as input to a summarization model instead of a human-annotated transcript. The How2 corpus contains 73,993 videos for training, 2,965 for validation and 2,156 for testing. The average length of transcripts is 291 words and of summaries is 33 words. A more general comparison of the How2 dataset for summarization as compared with certain common datasets is given in (Sanabria et al., 2018).

Video-based Summarization. We represent videos by features extracted from a pre-trained action recognition model: a ResNeXt-101 3D Convolutional Neural Network (Hara et al., 2018) trained to recognize 400 different human actions in the Kinetics dataset (Kay et al., 2017). These features are 2048 dimensional, extracted for every 16 non-overlapping frames in the video. This results in a sequence of feature vectors per video rather than a single/global one. We use these sequential features in our models described in Section 3.

Speech-based Summarization. We leverage the speech modality by using the outputs from a pretrained speech recognizer that is trained with other data, as inputs to a text summarization model. We use the state-of-the-art models for distant-microphone conversational speech recognition, ASpIRE (Peddinti et al., 2015) and EESEN (Miao et al., 2015; Le Franc et al., 2018). The word error rate of these models on the How2 test data is 35.4%. This high error mostly stems from normalization issues in the data. For example, recognizing and labeling “20” as “twenty” etc. Handling these effectively will reduce the word error rates significantly. We accept these as is for this task.

Transfer Learning. Our parallel work Sanabria et al. (2019) demonstrates the use of summarization models trained in this paper for a transfer learning based summarization task on the Charades dataset (Sigurdsson et al., 2016) that has audio, video, and text (summary, caption and question-answer pairs) modalities similar to the How2 dataset. Sanabria et al. (2019) observe that pre-training and transfer learning with the How2 dataset led to significant improvements in unimodal and multimodal adaptation tasks on the Charades dataset.

3 Summarization Models

We study various summarization models. First, we use a Recurrent Neural Network (RNN) Sequence-to-Sequence (S2S) model (Sutskever et al., 2014) consisting of an encoder RNN to encode (text or video features) with the attention mechanism (Bahdanau et al., 2014) and a decoder RNN to generate summaries. Our second model is a Pointer-Generator (PG) model (Vinyals et al., 2015; Gülçehre et al., 2016) that has shown strong performance for abstractive summarization (Nallapati et al., 2016; See et al., 2017). As our third model, we use hierarchical attention approach of Libovický and Helcl 2017 originally proposed for multimodal machine translation to combine textual and visual modalities to generate text. The model first computes the context vector independently for each of the input modalities (text and video). In the next step, the context vectors are treated as states of another encoder, and a new vector is computed. When using a sequence of action features instead of a single averaged vector for a video, the RNN layer helps capture context. In Figure 2 we present the building block of our models.

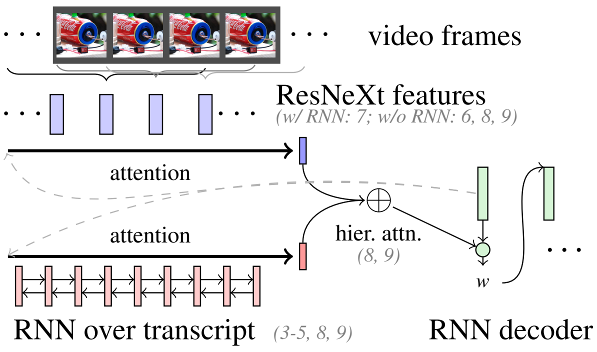


Figure 2: Building blocks of the sequence-to-sequence models, gray numbers in brackets indicate which components are utilized in which experiments.

4 Evaluation

We evaluate the summaries using the standard metric for abstractive summarization ROUGE-L (Lin and Och, 2004) that measures the longest common sequence between the reference and the generated summary. Additionally, we introduce the Content F1 metric that fits the template-like structure of the summaries. We analyze the most frequently occurring words in the transcription and summary. The words in transcript reflect the conversational and spontaneous speech while the words in the summaries reflect their descriptive nature. For examples, see Table A1 in Appendix A.2.

Content F1. This metric is the F1 score of the content words in the summaries based over a monolingual alignment, similar to metrics used to evaluate quality of monolingual alignment (Sultan et al., 2014). We use the METEOR toolkit (Banerjee and Lavie, 2005; Denkowski and Lavie, 2014) to obtain the alignment. Then, we remove function words and task-specific stop words that appear in most of the summaries (see Appendix A.2) from the reference and the hypothesis. The stop words are easy to predict and thus increase the ROUGE score. We treat remaining content words from the reference and the hypothesis as two bags of words and compute the F1 score over the alignment. Note that the

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).

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).

score ignores the fluency of output.

Human Evaluation. In addition to automatic evaluation, we perform a human evaluation to understand the outputs of this task better. Following the abstractive summarization human annotation work of Grusky et al. (2018), we ask our annotators to label the generated output on a scale of $1-5$ on informativeness, relevance, coherence, and fluency. We perform this on randomly sampled 500 videos from the test set. We evaluate three models: two unimodal (text-only (5a), video-only (7)) and one multimodal (text-and-video (8)). Three workers annotated each video on Amazon Mechanical Turk. More details about human evaluation are in the Appendix A.5.

5 Experiments and Results

As a baseline, we train an RNN language model (Sutskever et al., 2011) 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 in Table 1. As another baseline, we replace the target summary with a rule-based extracted summary from the transcription itself. We used the sentence containing words "how to" with predicates *learn*, *tell*, *show*, *discuss* or *explain*, usually the second sentence in the transcript. Our final baseline was a model trained with the summary of the nearest neighbor of each video in the Latent Dirichlet Allocation (LDA; Blei et al., 2003) based topic space as a target. This model achieves a similar Content F1 score as the rule-based model which shows the similarity of content and further demonstrates the utility of the Content F1 score.

We use the transcript (either ground-truth transcript or speech recognition output) and the video action features to train various models with different combinations of modalities. 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 (Nallapati et al., 2016). We also observe that 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. We also use the automatic transcriptions obtained from a pretrained automatic speech recognizer as input to the summarization model. This model achieves competitive performance with the video-only models (described below) but degrades noticeably than

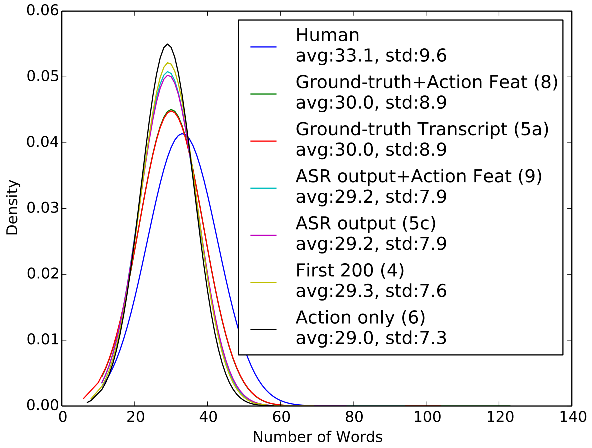


Figure 3: Word distribution in comparison with the human summaries for different unimodal and multimodal models. Density curves show the length distributions of human annotated and system produced summaries.

ground-truth transcription summarization model. This is as expected due to the large margin of ASR errors in distant-microphone open-domain speech recognition.

We trained two video-only models: the first one uses a single mean-pooled feature vector representation for the entire video, while the second one applies a single layer RNN over the vectors in time. Note that using only the action features in input reaches almost competitive ROUGE and Content F1 scores compared to the text-only model showing the importance of both modalities in this task. Finally, the hierarchical attention model that combines both modalities obtains the highest score.

In Table 2, we report human evaluation scores on our best text-only, video-only and multimodal models. In three evaluation measures, the multimodal models with the hierarchical attention reach the best scores. Model hyperparameter settings, attention analysis and example outputs for the models described above are available in the Appendix.

In Figure 3, we analyze the word distributions of different system generated summaries with the human annotated reference. The density curves show that most model outputs are shorter than human annotations with the action-only model (6) being the shortest as expected. Interestingly, the two different uni-modal and multimodal systems with ground-truth text and ASR output text features are very similar in length showing that the improvements in Rouge-L and Content-F1 scores stem from the difference in content rather than length. Example presented in Table A2 Section A.3 shows how the outputs vary.

6 Conclusions

We present several baseline models for generating abstractive text summaries for the open-domain videos in How2 data. Our presented models include a video-only summarization model that performs competitively with a text-only model. In the future, we would like to extend this work to generate multi-document (multi-video) summaries and also build end-to-end models directly from audio in the video instead of text-based output from pretrained ASR. We define and show the quality of a new metric, Content F1, for evaluation of the video summaries that are designed as teasers or highlights for viewers, instead of a condensed version of the input like traditional text summaries.

Acknowledgements

This work was mostly conducted at the 2018 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies,[1] hosted and sponsored by Johns Hopkins University. Shruti Palaskar received funding from Facebook and Amazon grants. Jindřich Libovický received funding from the Czech Science Foundation, grant no. 19-26934X. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) supported by NSF grant ACI-1548562 and the Bridges system supported by NSF award ACI-1445606, at the Pittsburgh Supercomputing Center.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. 2017. Nmtpy: A flexible toolkit for advanced

[1] https://www.clsp.jhu.edu/workshops/18-workshop/

neural machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 109:15–28.

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *Twenty-ninth AAAI conference on artificial intelligence*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

P. Das, C. Xu, R. F. Doell, and J. J. Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380. Association for Computational Linguistics.

Spandana Gella, Mike Lewis, and Marcus Rohrbach. 2018. A dataset for telling the stories of social media videos. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 968–974.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*, pages 40–48. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *CoRR*.

Çaglar Gülçehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Volume 1: Long Papers*.

Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *CoRR*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73. ACM.

Adrien Le Franc, Eric Riebling, Julien Karadayi, W Yun, Camila Scaff, Florian Metze, and Alejandrina Cristia. 2018. The aclew divime: An easy-to-use diarization tool. In *Interspeech*, pages 1383–1387. Interspeech, ISCA.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202.

Chin-Yew Lin and Eduard Hovy. 2002. From single to multi-document summarization. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 457–464.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, pages 605–612. Association for Computational Linguistics.

Zheng Lu and Kristen Grauman. 2013. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2714–2721.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Inderjeet Mani. 1999. *Advances in automatic text summarization*. MIT press.

Yajie Miao, Mohammad Gowayyed, and Florian Metze. 2015. Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 167–174. IEEE.

Arthur G Money and Harry Agius. 2008. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, 19(2):121–143.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Ça glar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *CoNLL 2016*, page 280.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. *CoRR*.

Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. 2016. Learning joint representations of videos and sentences with web image search. In *European Conference on Computer Vision*, pages 651–667. Springer.

Vijayaditya Peddinti, Guoguo Chen, Vimal Manohar, Tom Ko, Daniel Povey, and Sanjeev Khudanpur. 2015. Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 539–546. IEEE.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *TACL*, 1:25–36.

Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *Pattern Recognition - 36th German Conference, GCPR 2014*, pages 184–195.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389. Association for Computational Linguistics.

Shagan Sah, Sourabh Kulhare, Allison Gray, Subhashini Venugopalan, Emily Prud'Hommeaux, and Raymond Ptucha. 2017. Semantic text summarization of long videos. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 989–997. IEEE.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NIPS.

Ramon Sanabria, Shruti Palaskar, and Florian Metze. 2019. Cmu sinbad's submission for the dstc7 avsd challenge. In *Proc. 7th Dialog System Technology Challenges Workshop at AAAI*, Honolulu, Hawaii, USA.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68. Association for Computational Linguistics.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*.

Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Richang Hong. 2011. Multiple feature hashing for real-time large scale near-duplicate video retrieval. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 423–432. ACM.

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics*, 2:219–230.

Ilya Sutskever, James Martens, and Geoffrey E Hinton. 2011. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024. JMLR.org.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Atousa Torabi, Niket Tandon, and Leonid Sigal. 2016. Learning language-visual embedding for movie understanding with natural-language. *CoRR*, abs/1609.08124.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700. Curran Associates, Inc.

Meng Wang, Richang Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, and Tat-Seng Chua. 2012. Event driven web video summarization by tag localization and key-shot identification. *IEEE Transactions on Multimedia*, 14(4):975–985.

Kristian Woodsend and Mirella Lapata. 2012. Multiple aspect summarization using integer linear programming. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 233–243.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir R. Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 452–462.

Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun. 2016. Generation for user generated videos. In *European conference on computer vision*, pages 609–625. Springer.

Jianguo Zhang, Pengcheng Zou, Zhao Li, Yao Wan, Ye Liu, Xiuming Pan, Yu Gong, and Philip S Yu. 2018. Product title refinement via multi-modal generative adversarial learning. *arXiv preprint arXiv:1811.04498*.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 7590–7598.

A Appendix

A.1 Experimental Setup

In all our experiments, the text encoder consists of 2 bidirectional layers of the encoder with 256 Gated Recurrent Units (GRU; Cho et al. 2014) and 2 layers of the decoder with Conditional Gated Recurrent Units (CGRU; Sennrich et al. 2017). We optimize the models with the Adam Optimizer (Kingma and Ba, 2014) with learning rate $4 \cdot 10^{-4}$ halved after each epoch when the validation performance does not increase for maximum 50 epochs.

We restrict the input length to 600 tokens for all experiments except the best text-only model in the section Experiments and Results. We use vocabulary the 20,000 most frequently occurring words which showed best results in our experiments, largely outperforming models using subword-based vocabularies. We ran all experiments with the `nmtpytorch` toolkit (Caglayan et al., 2017).

Set	Words
Transcript	the, to, and, you, a, it, that, of, is, i, going, we, in, your, this, 's, so, on
Summary	in, a, this, to, free, the, video, and, learn, from, on, with, how, tips, for, of, expert, an

Table A1: Most frequently occurring words in Transcript and Summaries.

A.2 Frequent Words in Transcripts and Summaries

Table A1 shows the frequent words in transcripts (input) and summaries (output). The words in transcripts reflect conversational and spontaneous speech while words in the summary reflect their descriptive nature.

A.3 Output Examples from Different Models

Table A2 shows example outputs from our different text-only and text-and-video models. The text-only model produces a fluent output which is close to the reference. The action features with the RNN model, which sees no text in the input, produces an in-domain ("fly tying"' and "fishing") abstractive summary that involves more details like "equipment" which is missing from the text-based models but is relevant. The action features without RNN model belongs to the relevant domain but contains fewer details. The nearest neighbor model is related to "knot tying" but not related to "fishing". The scores for each of these models reflect their respective properties. The random baseline output shows the output of sampling from the random language model based baseline. Although it is a fluent output, the content is incorrect. Observing other outputs of the model we noticed that although predictions were usually fluent leading to high scores, there is scope to improve them by predicting all details from the ground truth summary, like the subtle selling point phrases, or by using the visual features in a different adaptation model.

A.4 Attention Analysis

Figure A1 shows an analysis of the attention distributions using the hierarchical attention model in an example video of painting. The vertical axis denotes the output summary of the model, and the horizontal axis denotes the input time-steps (from the transcript). We observe less attention in the first

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.

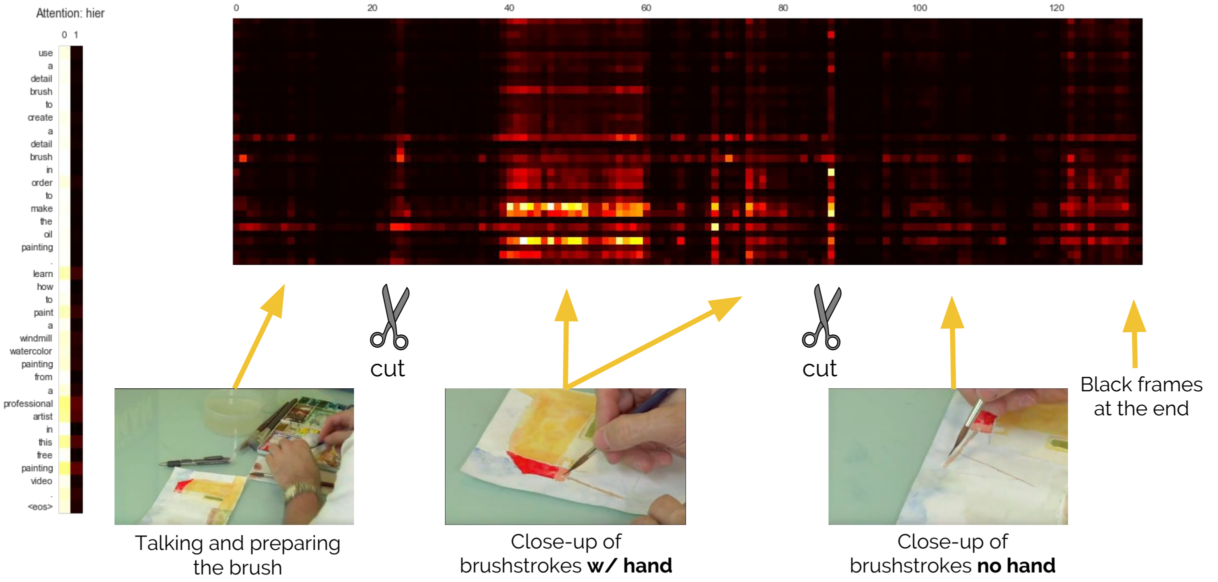


Figure A1: Visualizing Attention over Video Features.

part of the video where the speaker is introducing the task and preparing the brush. In the middle half, the camera focuses on the close-up of brush strokes with hand, to which the model pays higher attention over consecutive frames. Towards the end, the close up does not contain the hand but only the paper and brush, where the model again pays less attention which could be due to unrecognized actions in the close-up. There are black frames in the very end of the video where the model learns not to pay any attention. In the middle of the video, there are two places with a cut in the video when the camera shifts angle. The model has learned to identify these areas and uses it effectively. From this particular example, we see the model using both modalities very effectively in this task of the summarization of open-domain videos.

A.5 Human Evaluation Details

To understand the outputs generated for this task better, we ask workers on Amazon Mechanical Turk to compare outputs of unimodal and multimodal models with the ground-truth summary and assign a score between 1 (lowest) and 5 (highest) for four metrics: informativeness, relevance, coherence and fluency of generated summary. The annotators were shown the ground-truth summary and a candidate summary (without knowledge of the type of modality used to generate it). Each example was annotated by three workers. Annotation was restricted to English speaking countries. 129 annotators participated in this task.