Multimodal Abstractive Summarization for Open-Domain Videos

Jindřich Libovický[1] Shruti Palaskar[2] Spandana Gella[3] Florian Metze[2]
[1]Faculty of Mathematics and Physics, Charles University
[2]School of Computer Science, Carnegie Mellon University
[3]School of Informatics, University of Edinburgh
libovicky@ufal.mff.cuni.cz, spalaska@cs.cmu.edu
spandana.gella@ed.ac.uk, fmetze@cs.cmu.edu

Abstract

Multimodal and abstractive summarization of open-domain videos requires summarizing the contents of an entire video in a few short sentences, while fusing information from multiple modalities, in our case video and audio (or text). Different from traditional news summarization, the goal is less to "compress" text information only, but to provide a fluent textual summary of information that has been collected and fused from different source modalities. In this paper, we introduce the task of abstractive summarization for open-domain videos, we show how a sequence-to-sequence model with hierarchical attention can integrate information from different modalities into a coherent output, and present pilot experiments on the How2 corpus of instructional videos. We also present a new evaluation metric for this task called Content F1 that measures semantic adequacy rather than fluency of the summaries, which is covered by ROUGE and BLEU like metrics.

1 Introduction

There is an abundance of user-generated instructional video content on varied topics available on online platforms like YouTube. Usually, there are multiple videos for a given topic, often with subtle differences. The sheer volume of videos makes it impractical for humans to even "peek" into every one of them. An automatic abstractive summarization model that represents the key features of these videos, along with a short description of its content would be a useful tool for this application.

In the task of summarization for open-domain videos, we propose to generate natural language descriptions for video content using the transcriptions as well as visual features in the input. This task is different from the video captioning where short descriptions or subtitles are generated for each part of the video. Until recently, this task was a difficult problem due to the lack of training data as it would require manually annotated video descriptions for each video. In this work, we introduce this task in detail using the How2 dataset [1] which has such human annotated video descriptions, introduce a new evaluation metric that suits this task and present detailed results that make task understanding better.

2 How2 Dataset for Summarization

The How2 dataset [1] contains about 2,000 hours of short instructional videos, spanning different domains such as cooking, sports, indoor/ outdoor activities, music, etc. A human generated transcript and a 2 to 3 sentence summary is available for every video. The summary is somewhat templated, but it contains abstractive information about the video, that was entered by the video creator to generate

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

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 .

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 online video" is not mentioned in the transcript, and has to be derived from other sources.

interest in a potential viewer. It is a suitable target for the the multi-modal abstractive summarization task as defined above. More details about the dataset and how to download it are available at the website `https://github.com/srvk/how2`.

In Figure 1 we show an example video from the dataset with the transcript, summary and a keyframe from the video. We see that the transcript describes the entire process 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, which when fused together, helps in generating richer and fluent video descriptions. To have bigger validation and testing sets for our summarization experiments, we did not use the original splits of the How2 corpus, but we randomly select 73,993 videos for training, 2,965 for validation and 2,156 for testing. The average length of transcripts is 291 words and summaries is 33 words.

Action Feature Extraction The video features we use in this work are action features of 2048 dimensions extracted every 16 non-overlapping frames using a ResNext-101 3D Convolutional Neural Network [2] trained to recognize 400 different human actions in the Kinetics dataset [3]. This results in a sequence of feature vectors per video rather than a single/global one but we use both of these in our models described in Section 4. In order to obtain the latter, we average pooled the extracted features into a single 2048-dimensional feature vector which will represent all sentences segmented out of a single video.

2.1 Evaluation

Because the summaries in this dataset follow a certain pattern, we analyzed the most frequently occurring words in the source and target distributions, as shown in Table 1. The words in transcript reflect the conversational and spontaneous speech while the words in the summaries reflect their descriptive nature. We evaluate the summaries not only with the ROUGE score, but also introduce the Content F1 metric that fits the template-like structure of the summaries.

Table 1: 20 most frequently occurring words in Transcript and Summaries.

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

ROUGE We evaluate on the standard metric for abstractive summarization ROUGE [4], reporting the ROUGE-L score that measures the longest common sequence between the reference and generated summary. We notice that ROUGE-L prefers style of output over content.

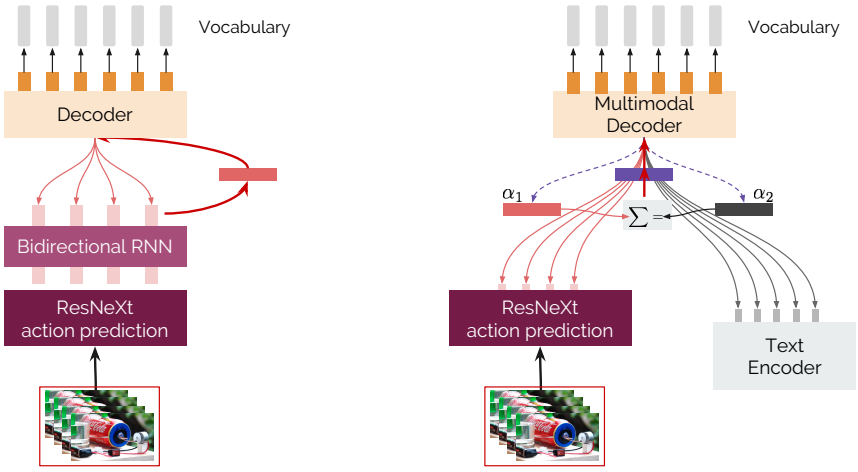

Figure 2: Video-only (left) and Text-and-Video models with Hierarchical Attention (right).

Content F1 This metric is the F1 score of the content words in the summaries based over a monolingual alignment. We use the METEOR toolkit [5, 6] to obtain the alignment and compute the F1 score. We set a zero weight to function words (δ), equal weights to Precision and Recall (α), and no cross-over penalty (γ) for generated words. Additionally, we observe a set of catchphrases from our summaries like the words *in, this, free, video, learn, how, tips, expert* as they appear in most summaries and are act like function words instead of content words. Due to their frequency, they are easy to predict and increase the ROUGE score. To get a fair understanding of the model quality, we remove these words from the reference and hypothesis as a post-processing step while calculating the Content F1 score. Note that Content F1 ignores the fluency of output.

3 Sequence-to-Sequence model with Hierarchical Attention

We use an Recurrent Neural Network (RNN) [7] based Sequence-to-Sequence (S2S) model [8] consisting of 3 main components: the encoder, the decoder and the attention-mechanism [9].

Hierarchical Attention For the multimodal summarization we follow the hierarchical attention approach [10] to combine textual and visual modalities. The model computes the context vector independently for each of the input modalities. In the next step, the context vectors are treated as states of another encoder and a new vector is computed. The hierarchical attention computation is shown in Figure 2, which also shows a video-only model using an RNN. When using a sequence of action features instead of a single averaged vector for a video, the RNN layer helps capture context.

4 Results and Discussion

As a baseline, we train an RNN language model [11] on all the summaries and randomly sampled tokens from it to obtain a prediction. The output obtained is fluent in English leading to a high ROUGE score but the content is poor, leading to a low Content F1 score in Table 2. As another baseline, we replace the target summary with rule-based extracted summary from the transcription itself. We used the sentence containing words "how to" with predicates "learn", "tell", "show", "discuss" or "explain". This was usually the second sentence in the transcript. Our final baseline was a model trained towards the summary of the nearest neighbor of each video instead of its own. This model achieves similar Content F1 score as the rule-based model which shows similarity of content and further demonstrates utility of the Content F1 score.

We use the text and action features to train various models with each and all of the modalities. The text-only model performs best when using the complete transcript in the input. We trained two video-only models: first model uses a single mean-pooled feature for entire video, while the second model applies a single layer RNN over them. The RNN layer helps capture sequence information

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	46.4	36
4	First 200 tokens	40.3	27.5
5	Complete Transcript (650 tokens)	**53.9**	**47.4**
6	Action Features only	38.5	24.8
7	Action Features + RNN	**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

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

in the video features that gives a richer feature representation than mean-pooling. Note that using only the action features in input obtains 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. Model hyper parameter settings, attention analysis and example outputs for the models described above are available in the Appendix.

5 Related Work

Neural abstractive summarization is an emerging field for which some commonly used text-only datasets are CNN/Daily Mail [12, 13], Gigaword [14] and the Document Understanding Conference challenge data [15]. Research in this field has mainly focused on summarization for the news domain, starting with single sentence generation [16], and progressing to multi-sentence [17] and multi-document summarization.

Multimodal abstractive summarization is a much more recent challenge with yet no benchmarking datasets. Li et al. 2017 [18] collected a multimodal corpus of news articles containing 500 videos of English news articles paired with human annotated summaries. The dataset has news articles with audio, video and summaries but there is no human annotated audio-transcript which is the main input. UzZaman et al. 2011 [19] collected a corpus of images, structured text and simplified compressed text for summarization of complex sentences where summarization is aided by these different image and text modalities. They designed methods to select summaries from available data due to lack of annotated training data. Sah et al. 2017 [20] propose methods to generate visual summaries for long videos. They also propose to generate textual summaries by using video captioning and preexisting summarization models as they do not have human annotations.

Summarization for the news domain focuses on condensing information from multiple sentences into one, and reproducing facts as correctly as possible. In our task, summarization focuses on describing the intent of the video and stating the exlusive and unique features of the video, irrespective of modality. This summary can thus be thought of like a textual "teaser" for the video.

6 Conclusion

We present a first multimodal video summarization system, which generates abstractive summaries on the open-domain How2 data. We define and show the quality of a new metric, Content F1, to evaluate the video descriptions that are designed as teasers or highlights for viewers, instead of condensed input like traditional abstractive summaries. We also present a video-only summarization model that performs competitively with a text-only model. In the future, we would like to extend this work to multi-document summarization with this dataset, and also build end-to-end models directly from speech instead of speech transcripts.

Acknowledgments

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. This work was also funded in part by the Amazon Research Awards. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

References

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

[2] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 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, 2018.

[3] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.

[4] Chin-Yew Lin and Franz Josef Och. 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, Barcelona, Spain, July 2004. doi: 10.3115/1218955.1219032.

[5] Satanjeev Banerjee and Alon Lavie. 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, 2005.

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

[7] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In Eleventh Annual Conference of the International Speech Communication Association, 2010.

[8] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems, pages 3104–3112, Montreal, Canada, 2014.

[9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473, 2014. ISSN 2331-8422.

[10] Jindřich Libovický and Jindřich Helcl. 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, Vancouver, Canada, July 2017.

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

[12] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In Advances in Neural Information Processing Systems, pages 1693–1701, 2015.

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

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

[14] Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. Annotated gigaword. In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, pages 95–100, 2012.

[15] Paul Over, Hoa Dang, and Donna Harman. Duc in context. Information Processing & Management, 43(6):1506–1520, 2007.

[16] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685, 2015.

[17] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368, 2017.

[18] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 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, 2017.

[19] Naushad UzZaman, Jeffrey P Bigham, and James F Allen. Multimodal summarization of complex sentences. In Proceedings of the 16th international conference on Intelligent user interfaces, pages 43–52. ACM, 2011.

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

[21] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 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), 2014.

[22] Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch-Mayne, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Miceli Barone, Jozef Mokry, and Maria Nadejde. Nematus: a toolkit for neural machine translation. In Proceedings of the EACL 2017 Software Demonstrations, 2017.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. CoRR, abs/1412.6980, 2014. ISSN 2331-8422.

[24] Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. Nmtpy: A flexible toolkit for advanced neural machine translation systems. The Prague Bulletin of Mathematical Linguistics, 2017.

A Appendix

A.1 Experimental Setup

In all our experiments, the text encoder consists of 2 bidirectional layers of encoder with 256 Gated Recurrent Units (GRU) [21] and 2 layers of decoder with Conditional Gated Recurrent Units (CGRU) [22]. The models are optimized with the Adam Optimizer [23] with learning rate $4 \cdot 10^{-4}$ halved after each epoch when the validation performance does not increase. We restrict the input length to 600 tokens for all experiments except the best text-only model in Section 4. We use vocabulary the 20,000 most frequently occurring words which showed best results in our experiments. We ran all experiments with the `nmtpytorch` toolkit [24].

A.2 Attention Analysis

Figure 3 shows an analysis of the attention distributions using the hierarchical attention model in an example video of painting. The vertical axis is 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 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 a 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 to not pay any attention. In the middle of the video, there are two places where there is 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 multimodal abstractive summarization of open-domain videos.

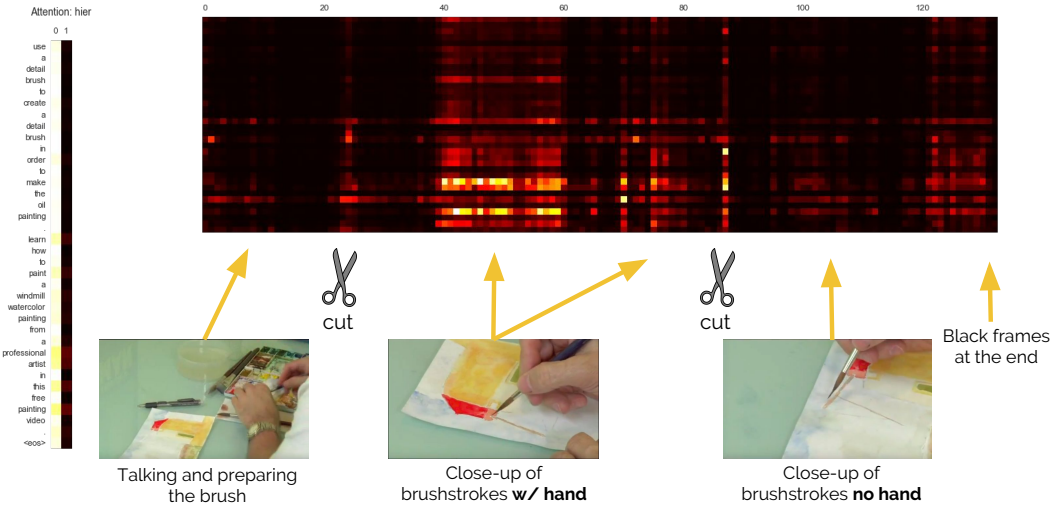

Figure 3: Visualizing Attention over Video Features.

A.3 Output Examples from Different Models

Table 3 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 lesser 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 groundtruth summary, like the subtle selling point phrases, or by using the visual features in a different adaptation model.

Model No.	Model	Content-F1	ROUGE-L	Output
-	Reference	100.0	100.0	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	Hierarchical Attention	48.9	54.9	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 .
5	Text-only	47.4	53.9	learn from our expert how to tie a thread for fly fishing in this free how - to video on fly tying tips and techniques .
7	Action Features + RNN	34.9	46.3	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	24.8	38.5	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	17.9	31.8	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	8.3	27.5	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 3: Example outputs of text-only, action and action-RNN models compared with the reference, and the topic-based next neighbor and random baseline.