AI Previews: A Web-based System for Summarization of live-streamings*

Amir Pouran Ben Veyseh[1,*], Franck Dernoncourt[2] and Thien Huu Nguyen[1]

[1]Dept. of Computer and Information Science, University of Oregon, Eugene, OR, USA
[2]Adobe Research, Seattle, WA, USA

Abstract

With the rapid rise of live-streaming content in recent years, there is a pressing need for developing summarization systems that can produce previews for live-streaming videos. A preview enables users to make quick decisions on whether to watch specific live-streaming content in matter of a 30-60 seconds before committing to watch the entire video. While summarization systems have already achieved promising performance across various domain, summarizing live-streaming content is still challenging due to the specific document characteristics, such as the language informality and the duration of videos which can reach several hours in length. In this paper, we present AI Previews, an easy-to-use web-based live-streaming summarization demo that produces previews of live-streaming videos by summarizing the transcript. AI Previews is an unsupervised extractive summarizer that first identifies the salient utterances within a given transcript, and then extracts those as the final summary to be displayed as a preview.

Keywords

AI Previews, Summarization, Video Transcripts

1. Introduction

In recent years, there has been increasing interest in livestream broadcasting platforms among Internet users [1, 2, 3]. Livestreams are present in various platforms such as YouTube Live, Behance, Vimeo, Instagram Live, Facebook Live, Twitch and TikTok. Livestreams give people an opportunity to be a creator and a presenter by spreading their content to their audience in real-time. Recent research has reported that the global livestreaming market size will expand in upcoming years. This growth can be attributed to rapid digitalization, massive use of smart phones and tablets, and the increasing popularity of online video streaming [4].

Livestreams can be of any variable recording length, ranging from a several seconds to several hours. The language used by livestreamers is mostly informal and unplanned, unlike other video content such as news broadcasts, movies, shows, and other scripted content. For these reasons, livestreams provide a unique challenge and a summarization mechanism can be advantageous for livestream platform users.

VTU'22: The AAAI-2022 Workshop On Video Transcript Understanding, Feb 28, 2022, Virtual

[*]You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

apouranb@cs.uoregon.edu (A. P. B. Veyseh); franck.dernoncourt@adobe.com (F. Dernoncourt); thien@cs.uoregon.edu (T. H. Nguyen)

CEUR Workshop Proceedings (CEUR-WS.org)

In this paper, we present AI Previews, a summarization system with a web-based graphical user interface (GUI) that aims to summarize livestreams videos. The goal of this summarization system is to give the end user a *preview* of the streamed content both visually and textually. The preview allows users to get a quick sense of the posted livestreamed video and then decide if it is in their interest to watch the original video in its entirety.

2. AI Previews

AI Previews is a web-based text summarization system that gives the user automatically generated summaries for each posted live-streaming video. It utilizes an unsupervised extractive summarization model to identify the salient utterances to produce a summary of the live-streaming transcript.

3. Model

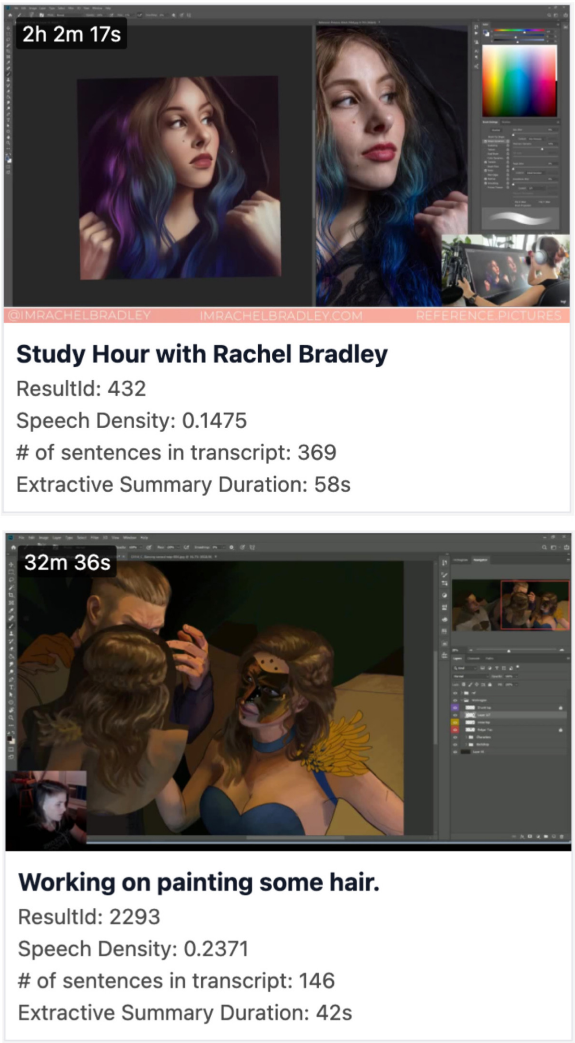
Much of existing research within the area of summarization has focused on written text such as news articles, clinical notes, scientific documents, social media posts, book chapters [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17].

While promising on text document summarization, the applicability of these models remains challenging on spoken text due to the specific characteristics of the verbal content mentioned earlier. AI Previews system employs an unsupervised vector-quantized variational autoencoder [18] to identify the major utterances (i.e., statements) of a given live-streaming transcripts [19]. AI Previews bridges the challenges of sequential prediction models including (1) the incapability of processing extremely long transcripts. Even though neural models that have been proposed to deal with long text documents such as Sparse Transformers [20], Longformer [21], Big Bird [22] and a few other others [23], it is yet challenging to process long transcripts as they are likely to break the maximum allowed length boundaries; (2) limited flexibility to deliver *real-time* summaries while the content is being streamed live. The transcripts can be retrieved from various sources such as interviews, multi-party meetings, podcasts, telephone speeches, and so forth.

4. User Interface

The user interface has thumbnails of live-streaming videos and various meta information such as the video title, video duration, speech density, number of sentences in the transcripts and summary for each posted video. The user can browse the live-streaming videos and if interested select a specific video as demonstrated in Figure 1. After selecting a live-streaming thumbnail, the user will be shown the generated summary by AI Previews below each posted video live-streaming, as shown in Figure 2. Users can also play the visual preview (chunks of the original summary, where the salient utterances occur) by clicking the “Play” (▶) button.

The start time of the utterance within the preview is given in bold text, followed by the start time of the utterance within the original video in parentheses. Users can also use filtering



2h 2m 17s

Study Hour with Rachel Bradley

ResultId: 432

Speech Density: 0.1475

# of sentences in transcript: 369

Extractive Summary Duration: 58s

32m 36s

Working on painting some hair.

ResultId: 2293

Speech Density: 0.2371

# of sentences in transcript: 146

Extractive Summary Duration: 42s

Figure 1: The web-based interface allows the users to browse the posted live-streaming videos. Each video box is expressed via a thumbnail with some meta information such as video title, video, and summary duration, speech density, number of sentences in video transcripts, and the extracted summary.

tools such as video and summary duration, keyword-based title search, and speech density and segments to narrow the list of video thumbnails based off their interest, as shown in Figure 3.

5. Implementation

The summarization model is developed in PyTorch [24], and the web interface is implemented in React and Mobx using TypeScript. The summarization model provides the salient utterances for the preview in JSON format with startTime, duration, and text of the utterance. The web application takes these JSON utterances and seeks the video player to the startTime of the utterance. The web application controls two video players during playback to cross fade the audio and video when the preview reaches the end of an utterance and begins the next utterance in the preview.

6. Conclusion and Future Work

In spite of the massive success of neural text summarization models on written textual content, it is still unclear how they can be scaled up and used in a production setting to summarize the spoken language such as live stream broadcasts. We present AI Previews, a practical and easy-to-use web-based application that aims at summarizing live stream broadcasting videos. The backend summarizer in AI Previews is an unsupervised extractive model that utilizes a variational autoencoder to identify the most important utterances in a given live-streaming transcript. It then delivers the final summary by concatenating the extracted salient utterances. One possible area of improvement could be the incorporation of utterance interrelations into the summarizer or using the video content.

References

[1] K. Pires, G. Simon, Youtube live and twitch: a tour of user-generated live streaming systems, in: Proceedings of the 6th ACM multimedia systems conference, 2015, pp. 225–230.

[2] J. Lin, Z. Lu, The rise and proliferation of live-streaming in china: Insights and lessons, in: International conference on human-computer interaction, Springer, 2017, pp. 632–637.

[3] T. Taylor, Watch me play: Twitch and the rise of game live streaming, Princeton University Press, 2018.

[4] GrandViewResearch, Video Streaming Market Size, Share Trends Analysis Report By Streaming Type, By Solution, By Platform, By Service, By Revenue Model, By Deployment Type, By User, By Region, And Segment Forecasts, 2021 - 2028, https://www.grandviewresearch.com/industry-analysis/video-streaming-market, 2021. [Online; accessed 18-Nov-2021].

[5] A. See, P. J. Liu, C. D. Manning, 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), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1073–1083. URL: https://www.aclweb.org/anthology/P17-1099. doi:10.18653/v1/P17-1099.

[6] R. Nallapati, F. Zhai, B. Zhou, Summarunner: A recurrent neural network based sequence model for extractive summarization of documents, in: S. P. Singh, S. Markovitch (Eds.), Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA, AAAI Press, 2017, pp. 3075–3081. URL: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14636.

[7] F. Dernoncourt, M. Ghassemi, W. Chang, A repository of corpora for summarization, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1509.

[8] S. Gehrmann, S. Layne, F. Dernoncourt, Improving human text comprehension through semi-Markov CRF-based neural section title generation, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1677–1688. URL: https://aclanthology.org/N19-1168. doi:10.18653/v1/N19-1168.

[9] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, N. Goharian, A discourse-aware attention model for abstractive summarization of long documents, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 615–621. URL: https://www.aclweb.org/anthology/N18-2097. doi:10.18653/v1/N18-2097.

[10] S. MacAvaney, S. Sotudeh, A. Cohan, N. Goharian, I. A. Talati, R. W. Filice, Ontology-aware clinical abstractive summarization, in: B. Piwowarski, M. Chevalier, É. Gaussier, Y. Maarek, J. Nie, F. Scholer (Eds.), Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, ACM, 2019, pp. 1013–1016. URL: https://doi.org/10.1145/3331184.3331319. doi:10.1145/3331184.3331319.

[11] Y. Liu, M. Lapata, Text summarization with pretrained encoders, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3730–3740. URL: https://www.aclweb.org/anthology/D19-1387. doi:10.18653/v1/D19-1387.

[12] S. Sotudeh Gharebagh, N. Goharian, R. Filice, Attend to medical ontologies: Content selection for clinical abstractive summarization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1899–1905. URL: https://www.aclweb.org/anthology/2020.acl-main.172. doi:10.18653/v1/2020.acl-main.172.

[13] L. Lebanoff, F. Dernoncourt, D. S. Kim, W. Chang, F. Liu, A cascade approach to neural abstractive summarization with content selection and fusion, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2020, pp. 529–535. URL: https://aclanthology.org/2020.aacl-main.52.

[14] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. URL: https://www.aclweb.org/anthology/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[15] S. Sotudeh, A. Cohan, N. Goharian, On generating extended summaries of long documents, The AAAI-21 Workshop on Scientific Document Understanding (SDU) (2021).

[16] J. Wu, L. Ouyang, D. M. Ziegler, N. Stiennon, R. Lowe, J. Leike, P. F. Christiano, Recursively summarizing books with human feedback, ArXiv abs/2109.10862 (2021).

[17] S. Sotudeh, H. Deilamsalehy, F. Dernoncourt, N. Goharian, TLDR9+: A large scale resource for extreme summarization of social media posts, in: Proceedings of the Third Workshop on New Frontiers in Summarization, Association for Computational Linguistics, Online and in Dominican Republic, 2021, pp. 142–151. URL: https://aclanthology.org/2021.newsum-1.15.

[18] A. van den Oord, O. Vinyals, k. kavukcuoglu, Neural discrete representation learning, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.

[19] S. Cho, F. Dernoncourt, T. Ganter, T. Bui, N. Lipka, W. Chang, H. Jin, J. Brandt, H. Foroosh, F. Liu, StreamHover: Livestream transcript summarization and annotation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6457–6474. URL: https://aclanthology.org/2021.emnlp-main.520.

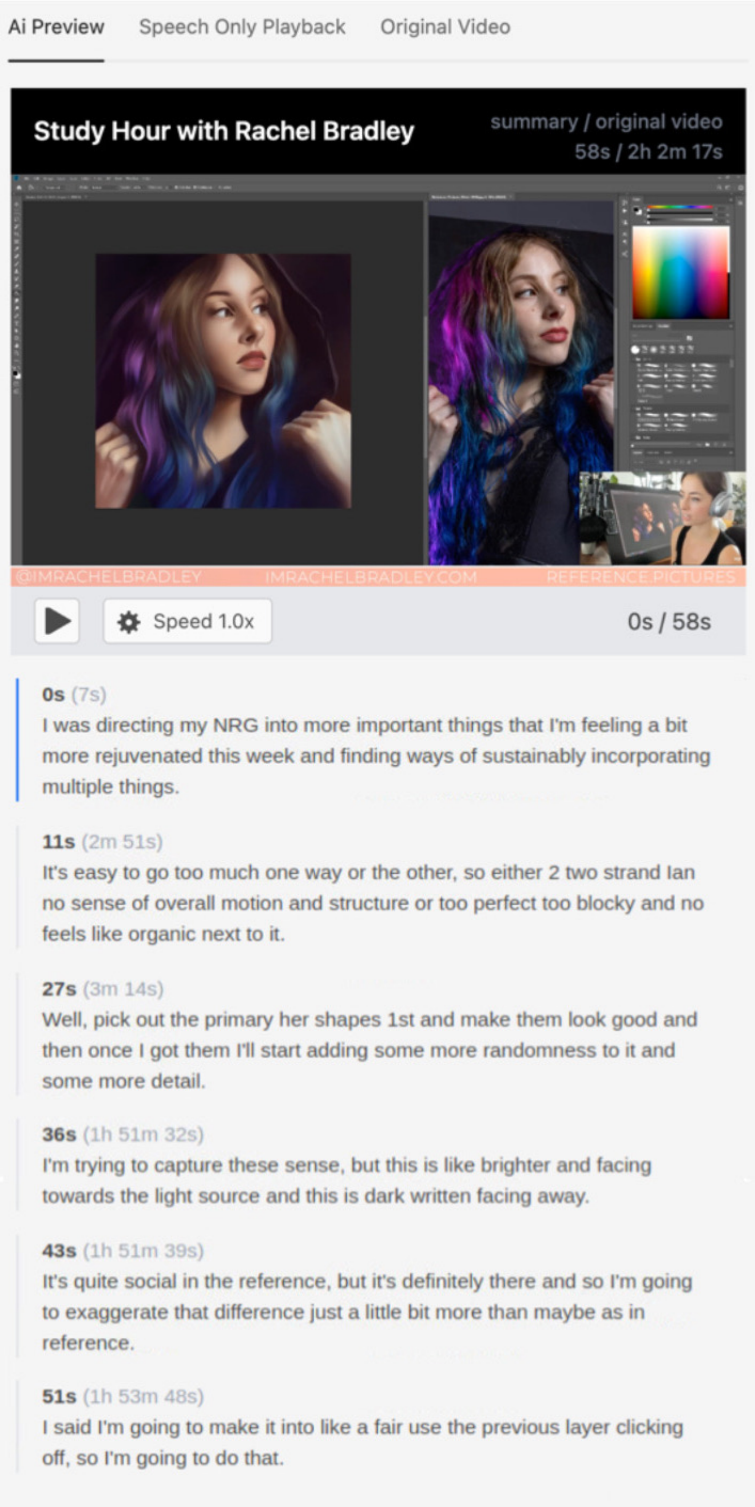
[20] R. Child, S. Gray, A. Radford, I. Sutskever, Generating long sequences with sparse transformers, arXiv preprint arXiv:1904.10509 (2019).

[21] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, ArXiv abs/2004.05150 (2020).

[22] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al., Big bird: Transformers for longer sequences., in: NeurIPS, 2020.

[23] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, D. Metzler, Long range arena : A benchmark for efficient transformers, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=qVyeW-grC2k.

[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019) 8026–8037.



0s (7s)
I was directing my NRG into more important things that I'm feeling a bit more rejuvenated this week and finding ways of sustainably incorporating multiple things.

11s (2m 51s)
It's easy to go too much one way or the other, so either 2 two strand Ian no sense of overall motion and structure or too perfect too blocky and no feels like organic next to it.

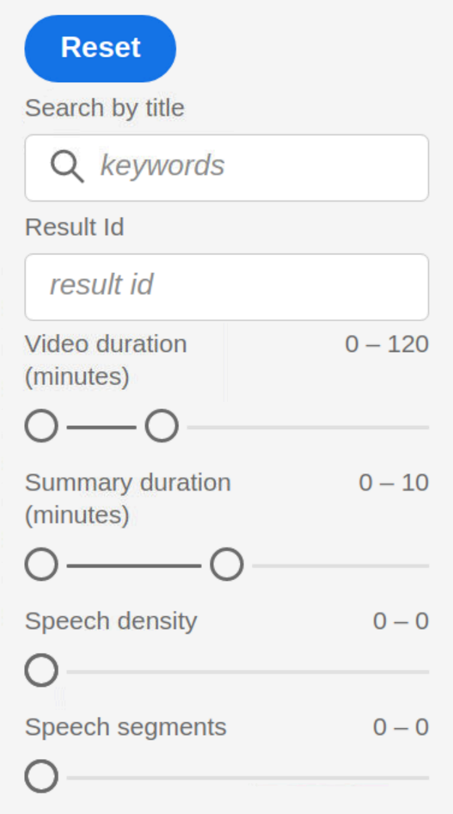
27s (3m 14s)
Well, pick out the primary her shapes 1st and make them look good and then once I got them I'll start adding some more randomness to it and some more detail.

36s (1h 51m 32s)
I'm trying to capture these sense, but this is like brighter and facing towards the light source and this is dark written facing away.

43s (1h 51m 39s)
It's quite social in the reference, but it's definitely there and so I'm going to exaggerate that difference just a little bit more than maybe as in reference.

51s (1h 53m 48s)
I said I'm going to make it into like a fair use the previous layer clicking off, so I'm going to do that.

Figure 2: The web-based interface of AI Previews system. The summary is given underneath the posted video. Users are shown the textual summary along with the visual preview.



Video duration (minutes) 0 – 120

Summary duration (minutes) 0 – 10

Speech density 0 – 0

Speech segments 0 – 0

Figure 3: Filtering functions on various video meta information, provided in the web-based interface, allowing the users to boil the content down according to their interest.