Video Summarization with ChatGPT

Wonho Lee[1]✉, Jungyu Kang[2], Nayoung Seong[3], Suhyeon Cho [4] ,Youngjong Kim[5]
*School of Software, Soongsil University

ABSTRACT

Recently, researchers have shown an increased interest in ChatGPT. ChatGPT is a NLP model, using text. This paper proposes a new way to summarize video efficiently by utilizing ChatGPT. By employing STT, we extract a text file of the caption of a video. ChatGPT summarizes it. In conclusion, it compared original caption text with condensed text by applying COS accuracy. We select parts of the text with higher accuracy to edit the video.

Keywords : Video Summarization, ChatGPT, NLP, STT

ChatGPT를 활용한 영상 요약 모델에 관한 연구

이원호[1], 강준규[2], 성나영[3], 조수현[4], 김영종[5]
*숭실대학교 소프트웨어학부

{hoho0907[1], shinysky5166[2], sna0e[3], parcon99[4]}@soongsil.ac.kr, youngjong@ssu.ac.kr[5]

요 약

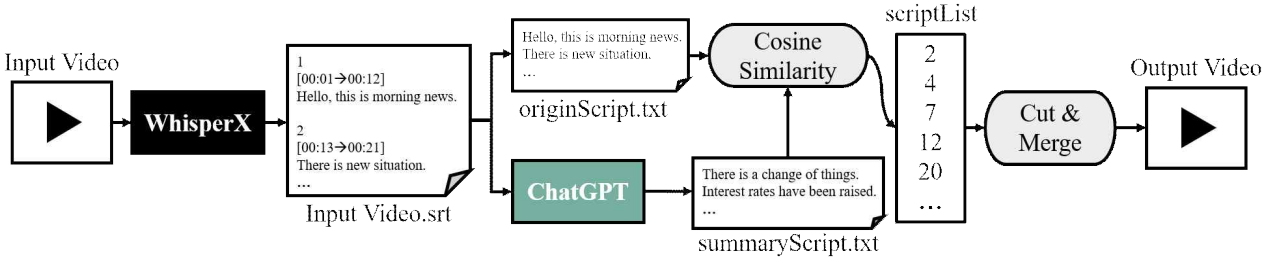
최근 ChatGPT를 각 분야에 활용하는 연구가 활발하게 이루어지고 있다. ChatGPT는 최신 자연어 처리 모델로, 텍스트를 통해 입출력을 진행한다. 본 논문에서는 이러한 ChatGPT를 활용하여 영상을 효과적으로 요약할 수 있는 새로운 접근 방식을 제시한다. STT기술을 사용하여 영상의 자막에 대한 텍스트 파일을 추출하고 이를 ChatGPT로 요약한다. 최종적으로 기존 텍스트와의 유사도 분석을 통해 유사도가 높은 부분을 선택하여 영상을 편집하고 요약한다.

키워드 : 영상 요약, ChatGPT, 자연어처리, 음성-텍스트변환

1. 서 론

ChatGPT는 최근 세계적으로 주목받고 있는 기술 중 하나이다. STATISTA에 따르면 2023년 1월 이용자수만 매일 약 1300만명이 늘어났다고 한다[1]. 틱톡과 인스타그램보다 빠른 성장세를 보인다는 것을 참고하면, ChatGPT의 이용자 수는 앞으로도 계속 늘어날 것이라 전망한다. 이용자의 수만큼 활용 분야 또한 넓어지고 있다. 마케팅, 교육, 가상 어시스턴트, 사용자 데이터 분석 등에 다양한 분야에 적용되고 ChatGPT를 활용한 비즈니스 모델이 점차 많아지고 있다[2]. 본 연구에서는 영상에 ChatGPT를 접목하여 활용성을 넓혔다.

영상은 시각 및 음성 데이터를 가지고 있다. 음성 데이터는 텍스트로 변환할 수 있기에, 영상의 내용을 추출할 수 있다. 이를 위해 STT 기술로 영상을 텍스트화하고, 이를 ChatGPT로 요약한다. 이후 NLP 유사도 분석을 통해 유사한 부분을 편집하여, ChatGPT로 영상을 요약하는 것이 본 연구의 목표이다.



[Fig. 1. Model Architecture]

※ "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음"(2018-0-00209).
✉ Corresponding Author : Lee Won Ho(hoho0907@soongsil.ac.kr)

2. 본 론

2.1 STT(Speech-to-Text)

STT(Speech-to-Text)기술은 음성 데이터를 입력하면 컴퓨터가 이를 해석하여 내용을 텍스트 데이터로 반환한다. 이러한 기술이 활용되고 있는 사례로는 AI 스피커, 음성 검색엔진 등이 있다.

본 연구에서는 영상을 요약하기 위해 ChatGPT를 사용하는데, 영상 내용에 대한 정보를 텍스트 데이터로 입력하기 위해 영상에 STT를 적용하여 자막을 추출한다. 여러 STT 모델 중에서도 자막이 나오는 타임라인이 포함된 srt파일을 반환해주는 whisperX[3]를 활용한다.

이에 더하여 whisperX를 통해 추출한 srt파일을 가공하여 잘못 나누어진 문장을 한 문장으로 합치는 방식으로 후보정을 한 후, 자막에 해당하는 텍스트 부분만을 따로 'originScript.txt'로 저장한다.

2.2 ChatGPT

ChatGPT는 OpenAI가 개발한 대화형 인공지능 챗봇으로 대형 자연어 모델인 GPT-3.5[4]를 기반으로 만들어졌다. 활용되고 있는 사례로는 Bing의 검색엔진 등이 있다.

본 연구에서는 whisperX와 가공을 거쳐 추출한 'originScript.txt'파일을 ChatGPT에 입력하여 이를 요약하도록 한다. 이 과정에서 ChatGPT API[5]를 사용하였으며, 요약한 텍스트를 문장별로 나누는 가공을 거쳐 'summaryScript.txt'파일로 저장한다.

2.3 NLP 유사도 분석

ChatGPT가 요약한 부분에 해당하는 영상을 추출하기 위해 'originScript.txt'와 'summaryScript.txt'파일의 텍스트 유사도를 분석하는 방식을 채택했다. 유사도 함수로는 NLP에서 텍스트의 유사도를 측정하는 코사인 유사도 함수[6]를 사용한다.

사용자가 원하는 만큼의 similarity점수를 threshold로 설정한다. 설정된 threshold를 만족하는'originScript.txt'의 line 인덱스를 집합으로 저장하여 중복값을 없애고, 리스트에 오름차순으로 저장한다.

2.4 Video Edit

유사도 분석을 통해 threshold를 넘는 자막들의 리스트를 받아오면, whisperX 단계에서 후보정한 srt파일에서 해당하는 인덱스의 타임스탬프 값들을 가져온다. 타임스탬프의 형식은 '시작 시간à종료 시간'으로 이루어져 있다. 추출한 타임스탬프들은 Moviepy[7], pysubs2[8]를 통해 subclip을 생성한다. 이후, ffmpeg[9]를 통해 subclip들을 concatenate하는 과정을 거쳐 최종적인 영상을 반환한다.

2.5 모델 구조도

본 논문에서 제안하는 ChatGPT를 이용한 영상 요약 모델은 [Fig. 1]과 같다.

3. 실험 결과

본 논문에서 제시한 ChatGPT의 영상 요약 성능을 측정하였다. 동일한 영상을 각기 다른 옵션으로 요약하였으며 이에 따라 요약된 영상의 길이를 기록하였다.

이 과정에서 사용된 영상은 2023년 4월 22일에 NBC News에서 진행한 Nightly News Full Broadcast영상(https://www.youtube.com/watch?v=Ax8YLYpPX-I)이다. 본 원본 영상의 길이는 총 14분 13초(853초)였으며 이를 각기 다른 옵션에 따라 요약한 결과는 다음 [Table 1.]과 같다.

요약 비율은 요약 영상시간을 원본 영상시간으로 나누고 100을 곱하여 백분율로 나타냈고 이를 100%에서 빼 얼마나 요약했는가를 측정하였다. 표의 결과에서 확인할 수 있듯이 Cosine Threshold값에 따라 요약 비율이 증가하는 것을 확인하였다. 자막 옵션도 요약 비율에 영향을 미쳤으나, 이는 자막에 따른 영향이 아니라 그 당시 ChatGPT가 요약한 정도에 따라 차이가 나는 것으로 확인되었다.

최종적으로 Cosine Threshold를 증가시킬수록 요약 비율이 높아지는 것과 최대 성능은 Cosine Threshold를 0.3으로 적용했을 때라는 것을 확인하였다. 이에 따른 최고 성능은 91.68% 요약한 결과물이었으며 구체적인 수치는 원본 영상에 따라 달라질 가능성이 존재한다. 또한, 매번 모델을 사용할 때 ChatGPT가 요약하는 정도에 따라서도 성능이 달라질 수 있다.

[Table 1. Performance of Model for each option]

Cosine Threshold	자막 옵션	영상 시간	요약 비율(%)
0.1	Add	10분 34초	25.67%
	No	10분 38초	25.21%
0.2	Add	3분 29초	75.50%
	No	4분 12초	70.46%
0.3	Add	1분 11초	91.68%
	No	1분 39초	88.40%

4. 결 론

본 논문에서는 WhisperX의 STT기술을 사용하여 영상을 텍스트화 한 후, ChatGPT와 NLP 유사도 분석을 통해 영상을 요약하는 방식을 제안하였다. STT와 NLP 유사도 분석을 통해 영상을 요약하는 것은 [10]에도 제시된 바가 있다.

본 연구의 취지는 영상을 요약하는 과정에서 텍스트를 주요 처리 데이터로 삼는 ChatGPT를 접목한다는 것에 있다. 현재는 이러한 영상 요약 모델을 사용할 수 있는 웹 인터페이스를 개발하는 과정에 있다.
ChatGPT를 활용한 영상 요약 모델에 대한 성능을 [Table 1.]에서 확인하였지만, 이는 특정 영상에 따른 요약 비율이기에 정확한 성능이라고 판단하기는 섣부른 점이 있다. 그렇기에 추후 일반적인 CNN 기반의 영상 요약 모델[11], 강화학습 기반의 영상 요약 모델[12], NLP 기반의 영상 요약 모델[9], 그리고 ChatGPT를 활용한 본 모델에 대해 성능을 비교하는 연구를 진행할 예정이다.

References

[1] Buchholz, K. (January 24, 2023). ChatGPT Sprints to One Million Users [Digital image]. Retrieved April 16, 2023, from https://www.statista.com/chart/29174/time-to-one-million-users/

[2] AppMagic. (January 12, 2023). Downloads of mobile apps using the keywords "chatbot" and "ChatGPT" in their title or description worldwide between January 1 and 11, 2023 [Graph]. In Statista. Retrieved April 16, 2023, from https://www.statista.com/statistics/1357710/chatbot-chatgpt-keyword-app-downloads/

[3] Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-accurate speech transcription of long-form audio. arXiv preprint arXiv:2303.00747.

[4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[5] OpenAI. (n.d.). ChatGPT. Retrieved [insert date] from https://beta.openai.com/docs/models/gpt-3.

[6] Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval. New York: McGraw-Hill.

[7] https://github.com/Zulko/moviepy

[8] https://github.com/tkarabela/pysubs2

[9] https://ffmpeg.org/

[10] Porwal, K., Srivastava, H., Gupta, R., Pratap Mall, S., & Gupta, N. (2022). Video Transcription and Summarization using NLP. Available at SSRN 4157647.

[11] Hussain, T., Muhammad, K., Ullah, A., Cao, Z., Baik, S. W., & de Albuquerque, V. H. C. (2019). Cloud-assisted multiview video summarization using CNN and bidirectional LSTM. IEEE Transactions on Industrial Informatics, 16(1), 77-86.

[12] Zhou, K., Qiao, Y., & Xiang, T. (2018, April). Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

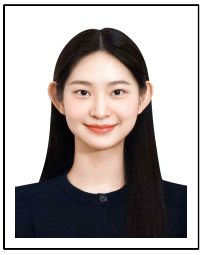
이 원 호

ORCID : 0009-0005-7473-743X
e-mail : hoho0907@soongsil.ac.kr
2021년 숭실대학교 소프트웨어학부 (학부생)
관심분야 : AI Security, Computer Vision

강 준 규

e-mail : shinysky5166@soongsil.ac.kr
2019년 숭실대학교 소프트웨어학부 빅데이터 융합전공 (학부생)
관심분야 : BigData, AI, Computer Vision

성 나 영



e-mail : sna0e@soongsil.ac.kr
2020년 숭실대학교 소프트웨어학부 빅데이터 융합전공 (학부생)
관심분야 : Front-end, Computer Vision, AI

조 수 현

e-mail : parcon99@soongsil.ac.kr
2018년 숭실대학교 소프트웨어학부 빅데이터 융합전공 (학부생)
관심분야 : Back-end, Macine Learning