Project Title

Teaser Generation for Long Documentaries and Educational Videos

Principal Investigator

Hao-Wen Dong, Assistant Professor, University of Michigan

Project Collaborators

Julian McAuley, Professor, University of California San Diego
Taylor Berg-Kirkpatrick, Associate Professor, University of California San Diego

NVIDIA Contact

Ming-Yu Liu (mingyul@nvidia.com) – The PI interned in Ming-Yu's team in Fall 2023.

Abstract

We propose a six-month project to develop new methodology for generating teasers for long videos, with a focus on documentaries and educational videos. Teasers are an effective tool for promoting content in entertainment, commercial and educational fields. However, creating an effective teaser for long videos is challenging for it requires long-range multimodal modeling on the input videos, while necessitating maintaining audiovisual alignments, managing scene changes and preserving factual accuracy for the output teasers. This project aims for the following two research directions: 1) **novel datasets and training methods** and 2) **new application domains and evaluation metrics**.

Our first research direction explores novel deep learning methods for modeling and understanding long-range multimodal data streams of video, audio and narrations. A particular focus is to explore leveraging large language models (LLMs) for generating the scripts for scene transitions and narrations. We also explore end-to-end training for learning the direct mapping from long videos to their teasers. In the second research direction, we apply the proposed method to generate teasers for documentaries, education videos, movies, vlogs and other media. We design new evaluation metrics to better evaluate the new task of teaser generation, where we conduct human evaluations to validate these evaluation metrics.

Project Keywords

Multimodal learning, language-vision model, large language model, intelligent video editing

NVIDIA Platforms

We will use NVIDIA GPUs and CUDA for implementing the deep learning models. The team is proficient in PyTorch programming. We plan to use the following pretrained models on NIM: automatic speech recognition (ASR; "nvidia/parakeet-ctc-1.1b-asr"), LLM ("meta/llama-3.1-405b-instruct"), text-to-speech (TTS; "nvidia/radtts-hifigan-tts"), contrastive language-vision model ("nvidia/nvclip") and multimodal LLM ("meta/llama-3.2-90b-vision-instruct").

Dataset and Model

We have compiled a portion of the dataset for documentaries and old movies in our preliminary study [1]. We are in the progress of scaling up and extending it to cover educational videos and other domains. All datasets will be made publicly available upon paper publications. The datasets include videos with human annotations, along with machine-transcribed narrations and machine-separated audio tracks of speech, music and sound effects. The datasets are expected to be 4-8 TB in total, stored in Ann Arbor, MI and La Jolla, CA.

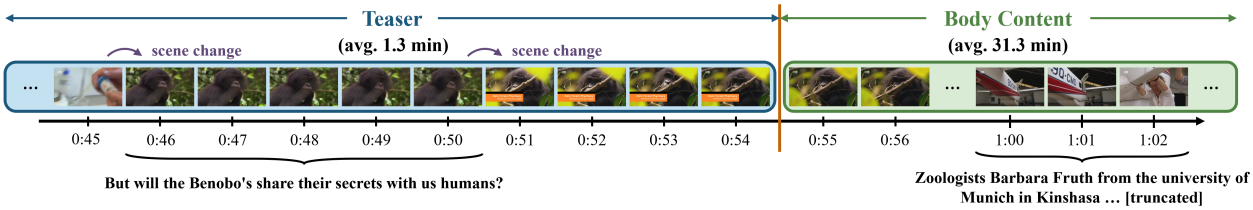

Figure 1: An example of our DocumentaryNet dataset, a collection of 1,269 high-quality documentaries paired with their teasers. We aim to generate teasers for long documentaries and educational videos.

Introduction

Teasers are an effective tool for promoting video contents such as documentaries, movies, vlogs, commercials and educational videos. Crafting a teaser for a documentary requires extensive manual editing, which involves sequencing of video clips, selection of music and sound effects and maintaining a cohesive narrative with audio-visual alignment. In this project, we explore generating teasers for long documentaries and educational videos. We envision it to find applications in both commercial and educational domains, enabling next-generation video editing interface. This project also contribute to the larger effort of the research community towards multimodal generative AI for long videos, where progress has been hindered due to the lack of a publicly-available dataset.

Related work One relevant research direction is movie trailer generation. However, most existing work [2–4] relies on private datasets, creating a barrier for the community to reproduce and follow up their research. Moreover, much prior work on this topic either only produce key frames instead of cohesive trailers with temporal consistency [5–7], or do not include an audio track in the generated trailer [8, 9]. Another relevant line of research is video summaiarization [10–16], where most prior work can only handle shorter videos or did not account for the audio tracks. Instead, teaser generation requires generating both the visuals and audio, handling longer input sequences, and achieving generally higher compression rates. Further, teasers needs an engaging, story-like narrative, necessitating some creativeness on top of preserving the factual accuracy of the content.

Methods

Datasets In our preliminary study [1], we have compiled a new dataset consisting of 1,269 documentaries paired with their teasers, featuring multimodal data streams of video, speech, music, sound effects and narrations (see Fig. 1). We will scale up the dataset to include educational videos and videos in other domains in the next few months.

Two-stage Narration-centered System The first aim of this project is to develop new training methods and generative models for long-range multimodal learning. Unlike movies, documentaries and educational videos rely heavily on narrations to convey information, and thus we will first explore a two-stage narration-centered approach. Given a long input video, we first generate the teaser narration from the transcribed narration, and then select the most relevant visual content from the documentary to accompany the generated teaser narration. We leverage a pretrained LLM with specially-designed prompts to generate teaser narrations, and we plan to finetune the LLM on our datasets to adapt it to better capture the narration style of documentaries and educational videos. For narration-video matching, we explore in our preliminary study training a transformer-based model to learn the mapping between the narrations and visuals [1] (see Fig. 2). We have examined this approach on a smaller dataset, without finetuning the foundational models. Despite the limited training budget,

preliminary study showed that this approach outperforms two baseline models [12, 14] in human evaluation (example results). We will also explore adopting a multimodal LLM to leverage its visual reasoning capability for narration-visual matching.

End-to-end systems We will also explore end-to-end approaches that learn the direct mapping between the full-length videos and their teasers. We will extract the CLIP-image embeddings of the input video and those of the teaser, and train a transformer model to learn the mapping between them. This idea has been explored in [8], but they did not consider the narration of the input video nor output an audio track.

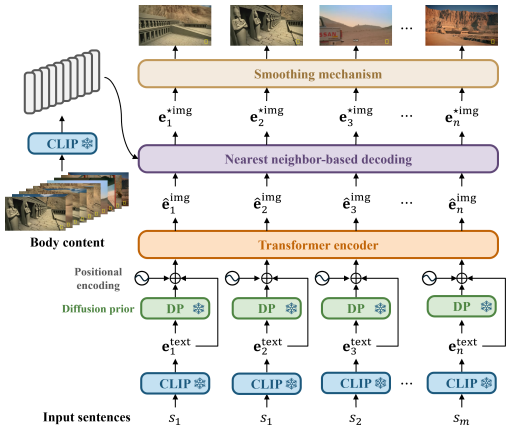

Figure 2: A transformer-based model we explored in our preliminary study that learns the mapping from the narrations to the visuals, in an embedding space.

Evaluation metrics The second aim of this project is to design new evaluation protocol for this new task of teaser generation. While most prior work in movie trailer generation adopts retrieval-based metrics for evaluation [8, 9, 17], our preliminary study showed that these retrieval-based metrics align poorly with human perception [1]. Instead, the two new metrics, repetitiveness and scene change rate, we proposed have shown better alignment with human preference in our human evaluation test. Our goal is to develop new evaluation metrics to evaluate the generated teasers beyond frame-level metrics, e.g, leveraging state-of-the-art scene change detection algorithms [18, 19] to compute scene-level metrics.

Expected Results

- New publicly-available datasets consisting of >2,000 high-quality full-length documentaries and educational videos paired with their teasers
- A new teaser generation system that can effectively compress a >30-min documentary or educational video into a <3-min teaser
- New foundation models finetuned for new application domains
- A new evaluation suite that aligns with human preference for teaser generation
- Conference publications (NeurIPS Dataset & Benchmarks Track, ICLR and CVPR)

Upon paper publications, we will make publicly available all the datasets, source code, pretrained models and hyperparameters on our project website.

Project Support Details

We are requesting a total amount of 32,000 A100 40 GB hours, and we will need at most 16 concurrent GPUs at a time. The requested GPU hours will be used to cover three 2-GPU instances for 6 months for development and one 8-GPU instance for large-scale training sessions for weeks. We will need 10TB of cloud storage for storing datasets and checkpoints.

Cloud Readiness

All developments will be based on Linux-based systems, and the team has experience working on cloud-hosted GPU instances with attached network storage. The team is proficient in “Importing your data,” “Installing dependencies,” “Installing and executing code base at the command line,” “Backing up any intermediate results,” and familiar with “Installing and integrating CUDA and other NVIDIA SDKs” and “Building and running containers.”

Appendix C – References

[1] Weihan Xu, Paul Pu Liang, Haven Kim, Julian McAuley, Taylor Berg-Kirkpatrick, and Hao-Wen Dong. “TeaserGen: Towards Generating Teasers for Long Documentaries.” *submitted to ICLR*. 2024.

[2] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. “MovieNet: A Holistic Dataset for Movie Understanding.” *ECCV*. 2020.

[3] Mattia Soldan et al. “MAD: A Scalable Dataset for Language Grounding in Videos from Movie Audio Descriptions.” *CVPR*. 2022.

[4] Xiaowei Chi et al. *MMTrail: A Multimodal Trailer Video Dataset with Language and Music Descriptions*. 2024.

[5] Hsuan-Wei Chen, Jin-Hau Kuo, Wei-Ta Chu, and Ja-Ling Wu. “Action movies segmentation and summarization based on tempo analysis.” *MIR*. 2004.

[6] Xingchen Liu and Jianming Jiang. “Semi-supervised Learning Towards Computerized Generation of Movie Trailers.” *IEEE SMC*. 2015.

[7] Alan F. Smeaton, Bart Lehane, Noel E. O’Connor, Conor Brady, and Gary Craig. “Automatically selecting shots for action movie trailers.” *MIR*. 2006.

[8] Dawit Mureja Argaw et al. “Towards Automated Movie Trailer Generation.” *CVPR*. 2024.

[9] Lezi Wang, Dong Liu, Rohit Puri, and Dimitris N. Metaxas. “Learning Trailer Moments in Full-Length Movies.” *ECCV*. 2020.

[10] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. “Multi-modal Summarization for Asynchronous Collection of Text, Image, Audio and Video.” *EMNLP*. 2017.

[11] Xiyan Fu, Jun Wang, and Zhenglu Yang. “MM-AVS: A Full-Scale Dataset for Multi-modal Summarization.” *NACCL*. 2021.

[12] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. “CLIP-It! Language-Guided Video Summarization.” *NeurIPS*. 2021.

[13] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. “Align and Attend: Multimodal Summarization with Dual Contrastive Losses.” *CVPR*. 2023.

[14] Kevin Qinghong Lin et al. “UniVTG: Towards Unified Video-Language Temporal Grounding.” *arXiv preprint arXiv:2307.16715* (2023).

[15] Jingyang Lin et al. “VideoXum: Cross-modal Visual and Textural Summarization of Videos.” *TMM* (2023).

[16] Dawit Mureja Argaw et al. “Scaling Up Video Summarization Pretraining with Large Language Models.” *CVPR*. 2024.

[17] Mohammad Hesham, Bishoy Hani, Nour Fouad, and Eslam Amer. “Smart trailer: Automatic generation of movie trailer using only subtitles.” *IWDRL*. 2018.

[18] Md Mohaiminul Islam, Mahmudul Hasan, Kishan Shamsundar Athrey, Tony Braskich, and Gedas Bertasius. “Efficient Movie Scene Detection using State-Space Transformers.” *CVPR*. 2023.

[19] Xi Wei, Zhangxiang Shi, Tianzhu Zhang, Xiaoyuan Yu, and Lei Xiao. “Multimodal High-order Relation Transformer for Scene Boundary Detection.” *CVPR*. 2023.