Joint Moment Retrieval and Highlight Detection Via Natural Language Queries

Richard Luo rluo47@gatech.edu
Austin Peng apeng39@gatech.edu
Koby Beard kbeard8@gatech.edu
Heidi Yap hyap7@gatech.edu

Team Visionary Vids
Georgia Institute of Technology
Atlanta, GA 30332

1. Project Summary

Video content has proliferated in the last few years, now taking up the majority of all internet traffic. However, with increasingly large amounts of video available, there has been a shift towards summarizing or extracting just the highlights of a video. Twitch, for example, has clips where users can splice just an interesting section of hours-long live streams. Similarly, TikTok, Instagram Reels, and YouTube Shorts all move towards short-form content.

In this project, we propose a new method for natural language query based joint video summarization and highlight detection using multi-modal transformers. This approach will use visual and audio cues from a video to generate its retrieval.

2. Related Work and Our Approach

The current approach performs video moment retrieval and highlight detection through a Unified Multi-modal Transformer (UMT) [6], which utilizes text, video, and audio as inputs. Each input is preprocessed by pretrained feature extractors. The model structure consists of five parts: uni-modal encoder, cross-modal encoder, query generator, query decoder, and prediction heads. We will base our general implementation off the UMT codebase [3].

Our team plans on modifying the UMT model to use ConvNeXt v2 model [1, 11] as basis for modifications to the UMT encoder. ConvNeXt v2 builds upon the original ConvNeXt paper [8] by adding a new Global Response Normalization (GRN) layer to enhance inter-channel feature competition. Inspiration behind the ConvNeXt architecture was largely due to the history of using CNNs in computer vision. ConvNeXt takes together several lessons from the increasingly popular vision transformers and modernizes the ConvNet, even outperforming state-of-the-art Swin Transformers [7]. The ConvNeXt paper studied a series of design decisions including macro design (stage compute ratio, patchifying layers), ResNeXt, inverted bottleneck, larger kernel sizes, and micro designs (testing different activation functions and normalization methods, testing number of activation functions and normalizations). ResNeXt designs include using grouped convolutions, depthwise convolutions, 1x1 convolutions to lead to separation of spatial and channel mixing, a property shared by vision transformers (SOTA model for computer vision image recognition tasks).

We will analyze how each design decision in the ConvNeXt paper affects our results for the joint video moment retrieval and highlight detection task. The metrics we will use are mean average precision (mAP), Intersection over Union (IoU), and recall. We will compare our results with the UMT model, Moment-DETR and other related state of the art models for the moment retrieval and highlight detection tasks.

3. Resources and Related Work

The Unified Multimodal Transformers (UMT) approach seems to be the strongest approach to creating video highlight clips [6]. UMT is possibly the first to successfully use multi-modal (visual-audio) learning to detect video highlights with natural language querying (known as video moment retrieval) and without it. The approach involves sending video and audio input through uni-modal encoders which both go to a cross-modal encoder, which connects to a query decoder and a query generator (which can receive text input). The last element are two prediction heads which give results for moment retrieval and highlight detection. Other approaches to highlight clipping problems can be found in the papers below.

- Videogenic [5]
- QVHighlights [4]
- AENet [10]
- Adaptive Video Highlight Detection by Learning from User History [9]

4. Dataset

We will use the QVHighlights [4] dataset which can easily be downloaded from a GitHub repository [2].

5. Ethical Considerations

One ethical consideration is bias with video highlights. The highlights might give a biased representation of the full video, misleading watchers. In order to mitigate this, clips generated should be checked by a human before they are posted anywhere, especially if it is an important video. Another issue is privacy; if our project is used as an app, clips produced may contain private or sensitive information of users. This can be mitigated by making sure the algorithm does not keep any of its user's data in the app.

References

[1] Github: Convnext v2. https://github.com/facebookresearch/ConvNeXt-V2. 1

[2] Github: Qvhighlights dataset files. https://github.com/jayleicn/moment_detr/tree/main/data. 2

[3] Github: Unified multi-modal transformers. https://github.com/tencentarc/umt. 1

[4] Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries, 2021. 1, 2

[5] David Chuan-En Lin, Fabian Caba Heilbron, Joon-Young Lee, Oliver Wang, and Nikolas Martelaro. Videogenic: Video highlights via photogenic moments, 2022. 1

[6] Ye Liu, Siyuan Li, Yang Wu, Chang Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection, 2022. 1

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[8] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. 1

[9] Mrigank Rochan, Mahesh Kumar Krishna Reddy, Linwei Ye, and Yang Wang. Adaptive video highlight detection by learning from user history, 2020. 1

[10] Naoya Takahashi, Michael Gygli, and Luc Van Gool. Aenet: Learning deep audio features for video analysis, 2017. 1

[11] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders, 2023. 1