IE643 Course Project

Non-Audio Video Summarization into Audio

Batclan

Bhavesh Sandbhor (22B2446)
22b2446@iitb.ac.in

Abstract

This project presents a non-audio video summarization system that generates audio summaries in Hindi from video content. Video frames are processed using a pretrained VGG16 model for feature extraction, followed by an encoder-decoder architecture to predict captions. The English captions are translated into Hindi using Google Translator and synthesized into speech using a text-to-speech system. Trained on a dataset of 6513 videos with 20 captions each, the system achieved a BLEU score of 0.54.

Introduction

- The rapid increase in multimedia content has created a pressing demand for efficient video summarization techniques, especially in non-audio scenarios.
- Pretrained models like VGG16 enable precise feature extraction, while encoder-decoder architectures excel in generating sequence-to-sequence outputs such as captions.
- Combining video summarization with language translation and text-to-speech systems makes content more universally accessible.

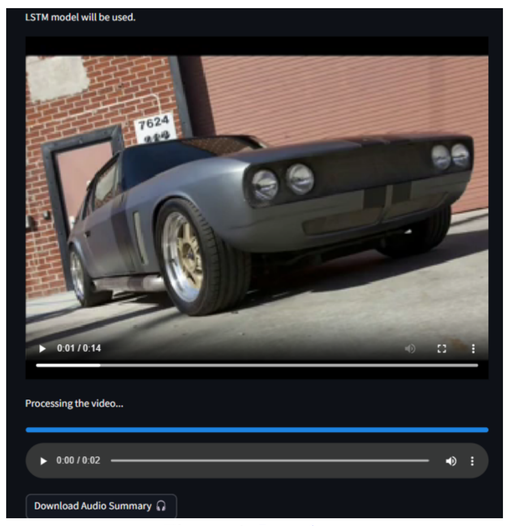

Figure 1: Interface

1. A white car is drifting.
2. Cars racing on a road surrounded by lots of people.
3. Cars are racing down a narrow road.
4. A race car races along a track.
5. A car is drifting in a fast speed.

Figure 2: Dataset

Workflow

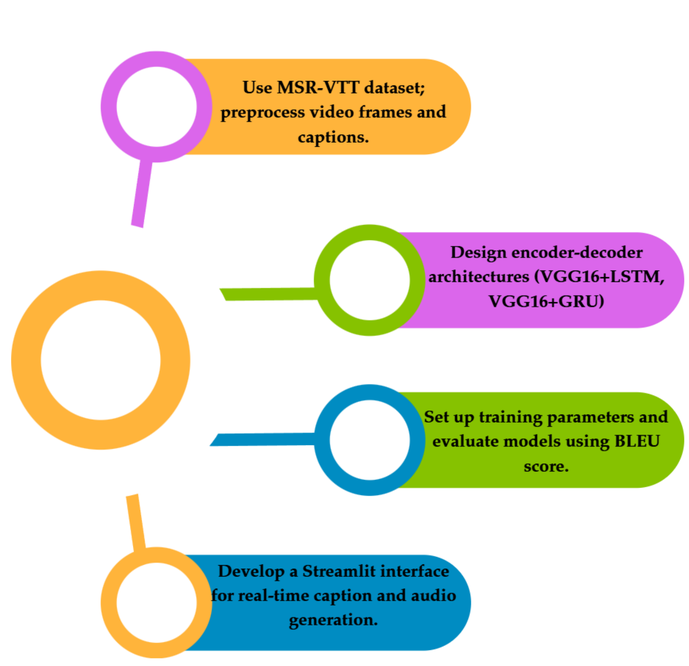

Figure 3: Workflow

Methodology

Figure 4 illustrates the architecture,...

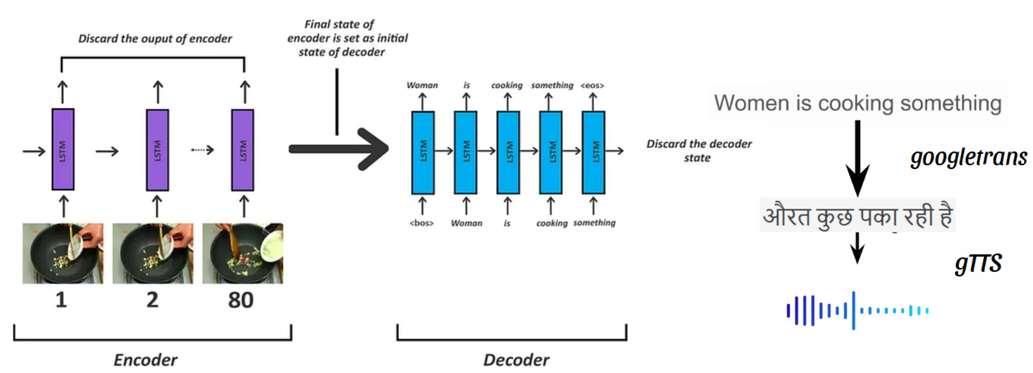

Figure 4: Overall Architecture

- **Feature Extraction (Encoder):** Extracts key features from videos using VGG16+LSTM to represent video content effectively.
- **Sequential Prediction (Decoder):** Predicts textual video summaries using LSTM, GRU, or Attention models.
- **Text-to-Audio Conversion:** Converts the textual summary into Hindi audio using Google Translator and gTTS.

Dataset Details

Category	Details
Video Quantity	7,010 videos
Video Length	10 to 30 seconds per video
Captions per Video	20 captions per video
Vocabulary Size	23,316 unique words
Text Preprocessing	Average of 15 words per sentence
Resolution	320x240 pixels (resized to 224x224 for processing)
Categories	20 categories, including:
	Sports, Music, Cooking, Gaming, News

Figure 5: Dataset Details

Novelty Assessment

- An Attention-based architecture was implemented to improve sequential prediction.
- LSTM-Attention model give better results than only LSTM or GRU

Result

Model	BLEU-1 Score
Baseline VGG16+LSTM	0.56
Our LSTM	0.537
Our GRU	0.543
Our LSTM+Attention	0.60

Table 1: Comparison of BLEU-1 Scores for Different Models

Conclusion

- Developed non-audio video summarization using VGG-16 for feature extraction and LSTM-Attention for sequence prediction
- Created a Streamlit web app for video upload, captioning, and audio playback, achieving competitive BLEU-1 scores.
- Future work includes integrating transformer models to enhance caption quality and scalability.

References

[1] Ramesh Kumar, Suman K. Saha, and Aditi Banerjee. *Attention-Based Video Captioning Framework for Hindi*, Conference on Video Analysis and Processing, 2022. Available: https://paperswithcode.com/paper/attention-based-video-captioning-framework.

[2] MSR-VTT: A Large-Scale Video Description Dataset. Available: https://paperswithcode.com/dataset/msr-vtt,

[3] Harshvardhan S. B. and P. K. Gupta. *Understanding RNNs, LSTMs, and GRUs*. Available: https://towardsdatascience.com/understanding-rnns-lstms-and-grus-ed62eb584d90

[4] Priya Singh. *Video Captioning with Keras*. Available: https://medium.com/analytics-vidhya/video-captioning-with-keras-511984a2cfff

[5] Ananya Tiwari. *RNN vs GRU vs LSTM*. Available: https://medium.com/analytics-vidhya/rnn-vs-gru-vs-lstm-863b0b7b1573

Acknowledgments

I would like to thank my professor P Balamurugan and Teachings Assistant for their continuous support and guidance throughout this project.

Github Link and Demo Video Link

Github Repository

Demo Video Link