Generative AI for Video-to-Text Summarization

Ethan Villalovoz, Fernando Medina, Luke Flock, Tomer Zangi
CPTS 437: Introduction to Machine Learning
School of Electrical Engineering & Computer Science, Washington State University

Project Proposal

***Abstract*—With the exponential growth of video content on platforms like YouTube and Amazon, automatic video summarization is critical for accessibility, search, and content recommendations. This project proposes a Generative AI model combining video, audio, and text data to produce concise and accurate summaries. We utilize machine learning principles such as feature extraction, model training, and evaluation with state-of-the-art models to generate efficient and context-aware summarization.**

I. Introduction

The exponential growth in video content creation has posed significant challenges for platforms like YouTube, Meta, and Amazon. As video consumption increases, providing users with quick access to relevant information from lengthy videos becomes more critical. Manual video summarization is time-intensive and inefficient, especially considering the volume of user-generated content uploaded daily. Automatic video summarization is emerging as a crucial tool for improving accessibility, user experience, and content searchability.

In industries like education, where instructional videos are widely used, or in entertainment, where media consumption is high, the ability to quickly summarize a video for recommendation or search systems is invaluable. Our project focuses on developing an efficient solution to this problem by leveraging state-of-the-art **Generative AI models** combined with machine learning techniques such as feature extraction, supervised learning, and evaluation metrics to generate accurate and context-aware summaries.

II. Methodology

This section details the step-by-step approach we will use for generating video-to-text summaries.

A. Video Feature Extraction

For each video, we will extract visual features using **Convolutional Neural Networks (CNNs)** or **Vision Transformers (ViTs)**. CNNs detect image patterns using convolutional layers focusing on smaller receptive fields, enabling the network to recognize objects, scenes, and interactions in individual video frames. Vision Transformers offer an alternative approach by using self-attention mechanisms to process the entire image, which improves the model's ability to recognize complex interactions between elements in a frame. These models will be pre-trained on large image datasets such as ImageNet and fine-tuned on video-specific data from **HowTo100M**.

B. Text Generation

Once the visual features are extracted, they will be fed into state-of-the-art **Generative AI models** such as **GPT-4** or **FLAN-T5**.

GPT-4 is the latest version of OpenAI's language models, capable of generating fluent, coherent, and contextually accurate text. **FLAN-T5**, a fine-tuned version of the T5 model, excels at text generation tasks, including summarization and translation. These models will be fine-tuned on the video datasets to generate highly coherent summaries based on the extracted visual and audio features. Fine-tuning ensures that the summaries generated align with the key themes and concepts present in the video content.

C. Multimodal Learning

Multimodal learning combines different sources of data to improve task performance. In our project, we will incorporate **audio features** using **Wav2Vec 2.0**, a state-of-the-art speech recognition model. Wav2Vec 2.0 transcribes spoken words into text, adding critical context to conversational or instructional videos. Where available, subtitles will be used as direct mappings to transcriptions, allowing us to align the visual and audio data accurately. By combining these different modalities, our model will generate more comprehensive and context-aware summaries. This multimodal integration is crucial, as videos often rely on both audio and visual cues to convey meaning.

III. Datasets

A. Primary Dataset: HowTo100M

The HowTo100M dataset consists of over 100 million instructional videos, making it an ideal choice for training a multimodal learning model. The dataset provides video and textual data, allowing efficient training on diverse content types. A key challenge in processing this dataset is the alignment of video frames with their corresponding transcriptions. We will preprocess the data by extracting keyframes at one-second intervals, ensuring that crucial visual information is retained while reducing redundancy.

B. Backup Dataset: YouTube8M

In case of scalability issues with **HowTo100M**, we will use the YouTube8M dataset, a large-scale video classification dataset that contains video clips with associated metadata. YouTube8M is well-annotated, and its smaller scale allows for faster iteration and testing. We will use this dataset for preliminary testing and model validation.

C. Custom Dataset

If further testing is required, we will scrape a custom dataset from YouTube, pairing video content with corresponding subtitles.

IV. Execution Plan and Evaluation

Week 1: Dataset Preparation
- Preprocess **HowTo100M** by extracting video frames, synchronizing audio, and pairing with text.
Milestone: Preprocessing complete.

Week 2: Video Feature Extraction
- Extract features using CNNs/Transformers. Validate on a small subset.
Milestone: Features extracted for a subset. Backup Plan: Use fewer frames if necessary.

Week 3: Full Feature Extraction & Audio Processing
- Extract visual and audio features and align them with video frames.
Milestone: Full feature extraction complete. Backup Plan: Prioritize video-to-text summarization.

Week 4: Text Generation
- Use GPT-4/FLAN-T5 to generate summaries based on visual and audio inputs.
Milestone: Text summaries generated for a sample.

Week 5: Multimodal Integration
- Combine visual, audio, and text inputs to enhance summarization quality.
Milestone: Multimodal integration complete.

Week 6: Evaluation and Refinement
- Evaluate the model using ROUGE/BLEU metrics and refine based on results.
Milestone: Final evaluation and refinement complete.

Evaluation Plan: We will use metrics like **ROUGE** and **BLEU** to evaluate text quality by comparing machine-generated summaries to ground truth summaries. Additionally, we will track **precision**, **recall**, and **F1-score** to evaluate the classification accuracy of the generated text. These standard machine learning evaluation metrics will help ensure the model is optimized for both context and coherence in the generated summaries.

V. Labor Division

Person 1: Video feature extraction using CNNs/Transformers, dataset preprocessing.
Person 2: Text generation using GPT-4/FLAN-T5.
Person 3: Audio integration using Wav2Vec 2.0 for multimodal learning.
Person 4: Evaluation, model refinement, and backup dataset management.

VI. Glossary

- **Generative AI**: A subset of AI focused on creating models that generate new content based on input data.
- **Multimodal Learning**: A machine learning approach that combines different data types, such as text, images, and audio, to improve model accuracy.
- **Convolutional Neural Networks (CNNs)**: A class of deep learning models used primarily for image and video recognition tasks.
- **Wav2Vec 2.0**: A self-supervised learning model for speech recognition, optimized for converting speech into text with minimal labeled data.

References

[1] D. Miech, et al., "HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips", *IEEE/CVF International Conference on Computer Vision*, 2019. Available: https://www.di.ens.fr/willow/research/howto100m/

[2] S. Abu-El-Haija, et al., "YouTube-8M: A Large-Scale Video Classification Benchmark", *arXiv*, 2016. Available: https://research.google.com/youtube8m/

[3] A. Baevski, et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations", *NeurIPS*, 2020. Available: https://arxiv.org/abs/2006.11477

[4] OpenAI, "GPT-4 Technical Report", 2023. Available: https://openai.com/research/gpt-4/

[5] Google Research, "FLAN-T5: Scaling Instruction-Tuning," 2022. Available: https://huggingface.co/docs/transformers/model_doc/flan-t5

Appendix

A. Dataset Preprocessing Details

Preprocessing the **HowTo100M** dataset involves extracting video frames at regular intervals and synchronizing them with their corresponding text transcriptions and audio tracks. For each video, frames will be extracted at one-second intervals to capture relevant visual information without overwhelming the model with redundant data.