Enhancing Automatic Metadata Generation for YouTube Videos through Multimodal Contextual Analysis

DRAFT PROPOSAL

IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE DEGREE OF BACHELOR OF SCIENCE IN ENGINEERING

Submitted By:
Kulathunga K.M.P.S (2021/E/078)
Chandrasiri P.G.P.M (2021/E/108)

Supervisor:
DR. Anantharajah Kaneswaran

Co-supervisor:
MR. Y. Pirunthapan

DEPARTMENT OF COMPUTER ENGINEERING
FACULTY OF ENGINEERING
UNIVERSITY OF JAFFNA
APRIL 2025

1. Introduction

Efficient metadata generation is now essential to improve discoverability and engagement due to the exponential growth of video content on platforms like YouTube. Metadata, including titles, tags, and descriptions, plays a key role in optimizing video searchability and viewer retention. However, manual metadata creation is labor-intensive and inconsistent. This research proposes a novel framework for automatic metadata generation using multimodal contextual analysis, leveraging advanced language models like text to text transfer transformer (T5) [3] to process video, audio, and transcript data for generating accurate and contextually relevant metadata.

1.1 Overview

Present metadata procedures on YouTube frequently fail to accurately reflect the full content of videos, mostly depending on manually inserted titles, tags, and descriptions. Because keyword-based algorithms frequently ignore more complex contextual elements in video content, these constraints affect searchability and alignment with user intent. Through multimodal contextual analysis that integrates visual, auditory, and textual features across novel model configurations [1], this research suggests a system that improves metadata generation by learning transformer-based architectures [3] such as T5, generating metadata that is more discoverable and more effectively fits with user search queries.

1.2 Research Gap

Existing metadata practices on YouTube primarily rely on manually entered titles, tags, and descriptions that often fail to comprehensively represent the content of the video. Current algorithms rely heavily on keyword based searches that may not align with user intent. Furthermore, while recent advancements like VATMAN demonstrate multimodal fusion capabilities using BART [1], there remains unexplored potential in adapting state-of-the-art language models like T5 [3] for metadata generation and investigating optimal architectural configurations for integrating emerging modalities. This research bridges the gap by expanding YouTube's search capabilities to focus on content-based metadata through advanced model architectures that systematically combine multiple modalities.

1.3 Aim and Objectives

- **Aim:** To develop a system that generates metadata for YouTube videos what you are going to upload as a content creator.
- **Objectives:**
  1. Decide properly what are the metadata we want to generate.
  2. Prepare a multimodal dataset using MSR-VTT videos [2] with precomputed features, transcripts (via Whisper), and metadata.
  3. Adapt state-of-the-art multimodal architectures (VATMAN) [1] for metadata generation.
  4. Evaluate the system using performance metrics [1].

1.4 Scope

This research works on developing a system that generates metadata for YouTube videos uploaded by content creators, addressing the limitations of manual metadata creation. By leveraging multimodal contextual analysis (video, audio, and transcript data), the system enhances searchability by generating metadata that fits with user intent and expands beyond traditional keyword based mechanisms.

2. Literature Review

Transformer-based models such as BART, T5, and PEGASUS are used in abstractive summarization advancements. Multimodal methods improve summarization by integrating data from voice, videos, and picture sources. VATMAN, our state of the art paper [1], introduces a novel trimodal hierarchical multi head attention mechanism to fuse information from video (extracted using ResNeXt-101), audio (extracted using Kaldi), and text (processed with BERT). For metadata generation research, the MSR-VTT dataset provides a diverse collection of 10,000 video clips sourced from 20 categories, and we used it's 7010 videos to make a custom dataset [2]. The summarization focused architecture of VATMAN offers useful ideas for metadata generation where its attention mechanism can be modified to generate titles, tags, and descriptions for YouTube videos. While VATMAN uses BART, our research uses fine tuned T5 due to its architectural advantages for metadata generation. T5's encoder decoder structure manages a variety of text-to-text operations, allowing a single model to generate titles, tags, and descriptions. Compared to BART's task-specific fine tuning, T5's pre-training on text-to-text translation allows for smooth adaptation and might provide better generalization and efficiency [3].

The quality of metadata can be considerably increased by utilizing automatic speech recognition (ASR) systems. Google's Whisper [4] provides accurate transcripts as valuable textual data for metadata creation. Additionally, integration of chrome extension with YouTube Interface has shown proper results in enhancing user experience. Metrics such as Content F1 [1], BLEU [1], and ROUGE [1] are used to evaluate the quality and relevance of generated metadata, making performance evaluation important. Through multimodal contextual analysis, this study attempts to improve automatic metadata generation for YouTube videos by modifying VATMAN's methodology and adding transcripts produced by ASR [4].

3. Methodology

3.1 Choosing the Model

The model selection is inspired by VATMAN's hierarchical crossmodal attention mechanism [1] due to its proven effectiveness in multimodal tasks:

- **Video Features:** ResNeXt-101 captures spatial-temporal patterns.
- **Audio Features:** Kaldi extracts MFCC features for auditory representation.
- **Text Features:** Whisper-generated transcripts are encoded using BERT for semantic understanding.

Just as VATMAN uses BART for summarization, T5 provides the backbone for metadata generation tasks. T5 is perfect for producing structured outputs such as titles, tags, and descriptions because of its sequence to sequence capabilities.

3.2 Training Process

- **Multimodal Fine-Tuning:** The large language model, T5 is fine tuned on precomputed features from MSR-VTT videos to align metadata effectively [3].
- **Feature Extraction Calibration:** Hierarchical attention layers are calibrated to optimize modality specific contributions during training [1].

3.3 Model Evaluation

Performance Metrics

Evaluation metrics include [1]:

- **Content F1:** Evaluates semantic adequacy by comparing metadata to video and transcript content.
- **BLEU:** Measures fluency and coherence of descriptions based on n-gram overlaps with reference data.
- **ROUGE:** Assesses recall oriented overlap between generated metadata and reference data, focusing on relevance.

3.4 Integration with YouTube Interface

A Chrome extension connects the trained model to YouTube's interface, enabling automatic metadata generation for content creators;

- **Extension Design:** Users upload videos via the extension, and a "Generate Metadata" button triggers API calls to the backend model.
- **Backend Integration:** FastAPI processes video inputs, runs the trained model, and returns metadata (titles, tags, descriptions) as JSON responses.
- **YouTube Metadata Submission:** Generated metadata is automatically populated into YouTube's upload fields (title, description, tags) for seamless integration.

4. Dataset and Architecture

4.1 Dataset Structure

The MSR-VTT dataset [2], consisting of 10,000 video clips, is widely recognized for its diverse and comprehensive video content annotated with textual descriptions. For this research, we utilize the MSR-VTT dataset's videos (7010 .mp4 files) to prepare a custom dataset structured as follows;

1. **Videos:** The original MSR-VTT (.mp4) files are retained as the primary input.
2. **Transcripts:** Speech-to-text models generate transcripts in (.srt) format aligned with video content.
3. **Metadata:** Titles, tags, and descriptions are generated in (.json) format to enhance video discoverability.

4.2 Architecture

The proposed architecture is here;

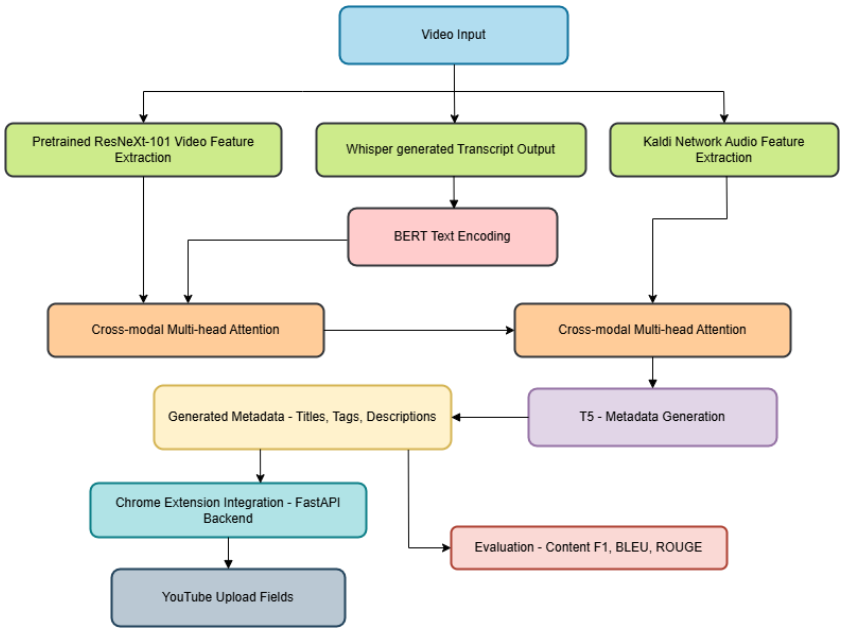

Figure 1: Architecture to generate metadata

1. **Core:** Encoder decoder structure with T5 backbone, inspired by VATMAN's crossmodal attention.
2. **Inputs:**
   - Video: ResNeXt-101 extracts 2048-D features from 16 frames per second.
   - Audio: Kaldi extracts 43-D filter bank and pitch features, normalized with CMVN.
   - Text: Whisper generates transcripts, encoded by BERT.
3. **T5 large language model:** Sequence to sequence model for generating titles, tags, and descriptions.

4. **YouTube Extension:** Connects trained model with YouTube, it takes video as input and generated metadata is populated into YouTube's upload fields

5. **Evaluation:** Content F1, BLEU, and ROUGE metrics are used.

5. Timeline

Semester	6 - SEMESTER				7 - SEMESTER				8 - SEMESTER			
week	1-4	5-8	9-12	13-14	1-4	5-8	9-12	13-14	1-3	4-5	6-8	9-10
Literature Review												
Annotated Bibliography												
Research Proposal												
Data Collection and Preprocessing												
Model Implementation												
Performance Evaluation and Comparison												
Research Project Report Writing												
Final paper & Thesis												

6. References

[1] D. Baek, J. Kim, and H. Lee, "VATMAN: integrating Video-Audio-Text for Multimodal Abstractive summarizatioN via Crossmodal Multi-head Attention Fusion," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3447737.

[2] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Dec. 2016, pp. 5288–5296. doi: 10.1109/CVPR.2016.571.

[3] A. Helwan, D. Azar, and D. U. Ozsahin, "Medical Reports Summarization Using Text-To-Text Transformer," in 2023 Advances in Science and Engineering Technology International Conferences, ASET 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ASET56582.2023.10180671.

[4] K. Penyameen, G. M. Siva Suriya Rajan, A. Arshath Ahamed, S. Yugesh Ram, J. John Shiny, and A. Periya Nayaki, "AI-Based Automated Subtitle Generation System for Multilingual Video Transcription and Embedding," in 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), IEEE, Feb. 2025, pp. 1096–1101. doi: 10.1109/IDCIOT64235.2025.10914946.