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)

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

Content

Contribution

Topics	2021 / E / 078	2021 / E / 108
1. Introduction	✓	
2. Literature Review	✓	✓
3. Methodology	✓	✓
4. Dataset and architecture		✓

1. Introduction

The exponential growth of video content on platforms like YouTube has created a pressing need for effective metadata generation to enhance discoverability and engagement. Metadata, including titles, tags, and descriptions, plays a pivotal 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 video, audio, and transcript data to generate accurate and contextually relevant metadata.

1.1 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 or hashing mechanisms that may not align with user intent. For instance, when a user searches for a video addressing a specific problem, the search results may not accurately reflect the content due to incomplete or irrelevant metadata. By contrast, our approach focuses on generating metadata that captures the full context of the video including visual, auditory, and textual elements thereby enhancing searchability. This research bridges the gap by expanding YouTube's search capabilities to focus on content-based metadata rather than solely relying on human-entered titles or hashtags.

1.2 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 with precomputed features, transcripts (via Whisper), and AI-generated metadata.
  3. Adapt state-of-the-art multimodal architectures (VATMAN) for metadata generation.
  4. Evaluate the system using performance metrics.

1.3 Scope

This research focuses 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 aligns with user intent and expands beyond traditional keyword-based or hashing mechanisms.

1.4 Overview

The proposed system integrates video keyframes (visual features), audio signals (MFCC features), and transcripts (generated via Whisper) to produce metadata through hierarchical attention-based fusion. The model is fine-tuned on dataset's metadata.

2. Literature Review

[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.

VATMAN (Video-Audio-Text Multimodal Abstractive summarizatioN) serves as the foundational work for this research. While VATMAN focuses on video summarization, its hierarchical crossmodal attention mechanism is highly relevant to metadata generation. VATMAN demonstrates:

- Effective fusion of video, audio, and text modalities using attention layers.
- Superior performance in summarization tasks with benchmarks like How2.

Relevance to Our Research:

- The hierarchical crossmodal attention mechanism will be adapted to generate titles, tags, and descriptions instead of summaries.
- VATMAN's use of pretrained models (e.g., ResNeXt-101 for video features) aligns with our objective of leveraging existing architectures for feature extraction.

Performance Insights:
VATMAN achieves high accuracy in multimodal summarization tasks (e.g., BLEU scores exceeding 0.8). These results suggest its potential applicability to metadata generation tasks.

3. Methodology

3.1 Data Preparation

We utilize the MSR-VTT dataset's videos to create a custom dataset with the following components:

1. Video/Audio Features: Extracted using ResNeXt-101 (video) and Kaldi MFCCs (audio).
2. Transcripts: Generated via Whisper speech-to-text models.
3. Metadata: Titles, tags, and descriptions generated by fine-tuned T5-based models.

3.2 Choosing the Model

The model selection is inspired by VATMAN's architecture:

- Pretrained ResNeXt-101 for spatial-temporal video features.
- Kaldi MFCC extractor for audio representation.
- Whisper-generated transcripts encoded using BERT.

3.3 Training Process

- **Multimodal Fine-Tuning:** The model is fine-tuned on precomputed features from the MSR-VTT dataset.
- **Feature Extraction Calibration:** Hierarchical attention layers are calibrated to align modalities effectively during training.

3.4 Model Evaluation

Providing Ground Truth

Since human-generated metadata is unavailable, we use AI-generated metadata as ground truth for evaluation.

Benchmarking Multimodal Integration

The system's ability to integrate video, audio, and text modalities is benchmarked against VATMAN's attention mechanism.

Performance Metrics

Evaluation metrics include:

- BLEU scores for description quality.
- F1-scores for tag prediction accuracy.
- Title similarity using cosine similarity between embeddings.

4. Dataset and Architecture

4.1 Dataset Structure

The custom dataset includes:

1. **Video/Audio Features:** Precomputed ResNeXt-101 and MFCC features stored as .npy files.
2. **Transcripts:** Whisper-generated .srt files aligned with videos.
3. **Metadata:** JSON files containing titles, tags, and descriptions.

4.2 Architecture

The proposed architecture consists of:

1. **Multimodal Feature Extraction**
   - Video: ResNeXt-101 extracts spatial-temporal patterns.
   - Audio: Kaldi extracts MFCC features.
   - Text: Whisper generates transcripts encoded with BERT.

2. **Fusion Mechanism**

   Hierarchical crossmodal attention aligns modalities at timestamp-level granularity:

   $$\text{Fused Features} = \text{Attention}(\text{Video},\text{Audio},\text{Text})$$

3. **Metadata Decoder**

   Three heads generate titles (classification), tags (multi-label classification), and descriptions (sequence generation).

5. Timeline

	Semester 6					Semester 7					Semester 8	
	W1-3	W4-6	W7-9	W10-12	W13-15	W1-3	W4-6	W7-9	W10-12	W13-15	W1-3	W4-6
Literature Review	■	■	■	■	■	■	■	■	■	■	■	■
Annotated Bibliography		■	■									
Dataset preparation			■	■								
Research proposal writing				■	■							
Building the Models						■	■	■				
Experimenting with the model									■	■		
Research Project Report writing											■	■
Research Paper Writing												■

5. Reference

[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.

MSR-VTT dataset - https://www.kaggle.com/datasets/vishnutheepb/msrvtt