

ENHANCING AUTOMATIC METADATA GENERATION FOR YOUTUBE VIDEOS THROUGH MULTIMODAL CONTEXTUAL ANALYSIS

PROGRESS REPORT

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

SEPTEMBER 2025

1. Introduction

YouTube metadata, consisting primarily of titles, descriptions, and tags, plays a crucial role in improving video discoverability and viewer engagement. Metadata helps YouTube's search and recommendation algorithms accurately index and promote videos relevant to users' interests. However, manual creation of comprehensive and contextually relevant metadata is time-consuming and often inconsistent across creators. Automating this process through advanced AI and multimodal analysis techniques can greatly enhance efficiency and content accessibility.

Our research focuses on developing a system that automatically generates accurate, context-aware metadata for YouTube videos by leveraging video, audio, and text features in combination with powerful language models. This enables improved video indexing and user experience while reducing manual effort by content creators.

Research Problem

Current YouTube video metadata generation methods primarily rely on keyword-based or shallow semantic techniques that lack deep multimodal contextual understanding. This limitation reduces the discoverability and engagement potential of videos by failing to represent core content dynamics from visual, audio, and textual modalities. Existing pre-trained models and fine-tuning approaches also face scalability, fluency, and adaptability challenges for dynamic real-time video uploads.

Aim and Objectives

The aim of this research is to develop a scalable, contextually aware system that automatically generates accurate YouTube video metadata (titles, tags, descriptions) by integrating multimodal features and advanced language models.

Objectives:

- Develop a custom multimodal dataset combining video, audio, and text features with human-annotated metadata from the MSR-VTT dataset
- Integrate LLaMA3 7B via Ollama in a Retrieval-Augmented Generation (RAG) architecture with Pinecone vector database for scalable retrieval and generation
- Build a backend API and Chrome extension for seamless metadata generation and insertion during YouTube video upload
- Evaluate initial performance qualitatively and plan improvements via advanced fusion mechanisms

2. Literature Review

Research on YouTube metadata generation has shown that optimizing metadata such as titles, tags, and descriptions can improve video discoverability and engagement by supporting search and recommendation algorithms. Traditional methods largely depend on keyword extraction or shallow semantic analysis lacking robust multimodal contextual understanding.^[4] This is highly relevant to our work as it highlights the limitations that our multimodal fusion approach aims to overcome.^[1]

The VATMAN^[1] model presents a state-of-the-art approach integrating video, audio, and text data through hierarchical crossmodal multi-head attention, fusing ResNeXt-101 video features, Kaldi MFCC audio features, and BERT text embeddings to generate abstractive video metadata.^[3] However, its complex architecture and resource demands limit practical scalability. This informs our choice to adopt scalable architectures and move beyond purely hierarchical attention models by leveraging Retrieval-Augmented Generation^[9] with the Pinecone vector DB.^[10]

Recent advancements in large language models, especially LLaMA3, offer strong multilingual and code reasoning capabilities with multimodal support. Deployment frameworks like Ollama enable local inference with improved privacy and control, while RAG architectures^[5] enhance generation with dynamic retrieval from vector databases such as Pinecone,^{[6][8]} providing enhanced accuracy and adaptability. These aspects are central to our research as we combine LLaMA3 via Ollama^[7] in a RAG framework^[9] to ensure real-time, adaptive metadata generation that scales efficiently.

The MSR-VTT dataset^[2] remains a standard for multimodal video research, providing extensive clip-sentence pairs with video, audio, and text streams, which form a strong base for metadata generation tasks. Its comprehensive coverage and multimodal nature underscore its suitability as the foundation for our custom dataset, enabling rigorous training and evaluation of multimodal metadata generation models.

Research Gaps:

- Existing models inadequately combine scalable large LLMs with vector-based retrieval for dynamic YouTube metadata generation.
- Fine-tuning approaches either lack fluency or adaptability under real time constraints.
- Practical deployment via seamless tools like Chrome extensions integrated with retrieval-based models remains unexplored.

3. Methodology

3.1 Creating the Dataset

The dataset was custom-built around the publicly available MSR-VTT dataset which comprises 8,811 videos covering diverse categories and content. Each video includes video frames, audio, and text transcripts. Audio features were extracted using Librosa MFCC, video features from ResNeXt-101 embeddings (2048 dimensions), and transcripts were generated via Whisper (768 dimensions). Metadata was compiled from two human-written captions per video, serving as ground truth for training and evaluation.

3.2 Architecture & Methods

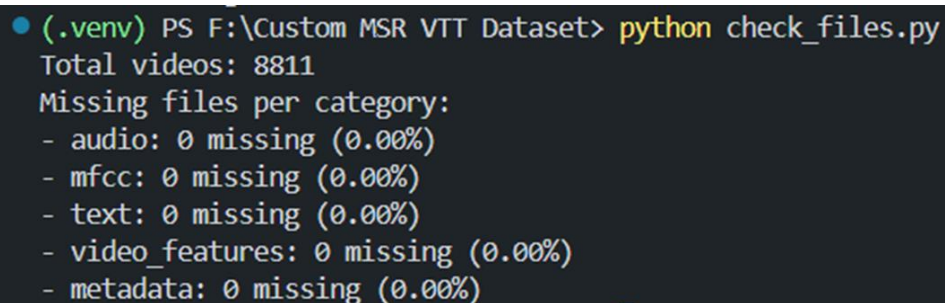
Our system uses the LLaMA3 7B large language model within a Retrieval-Augmented Generation (RAG) setup. Pinecone vector database stores and indexes multimodal embeddings for fast, scalable retrieval of relevant video contexts. The generation pipeline extracts multimodal embeddings, fuses them, retrieves top matches from Pinecone dynamically, and feeds this retrieved context alongside live transcripts into LLaMA3 to generate accurate titles, descriptions, and hashtags.

3.3 Implementation Steps

3.3.1 Data Preprocessing

The backend API is implemented with FastAPI, serving as the intermediate between the Chrome extension and the metadata generation model. The extension sends video data to the backend, where embeddings are extracted and indexed into Pinecone. The MSR-VTT dataset provided a robust source with 8,811 video clips annotated with human captions. We extracted multimodal features accordingly and carefully validated feature integrity (checking dimensions and completeness) before indexing.

The MSR-VTT dataset is characterized by large scale clip sentence pairs, comprehensive video categories ranging from technical demos to conversational English learning videos, and multimodal audio-video-text streams. Our preprocessing pipeline aligns the features and metadata appropriately for retrieval and generation.



```
● (.venv) PS F:\Custom MSR VTT Dataset> python check_files.py
Total videos: 8811
Missing files per category:
- audio: 0 missing (0.00%)
- mfcc: 0 missing (0.00%)
- text: 0 missing (0.00%)
- video_features: 0 missing (0.00%)
- metadata: 0 missing (0.00%)
```

Figure 01: Output of Checking missing values

3.3.2 Model Building

We adopted LLaMA3 7B via the Ollama framework for local, privacy preserving inference. The RAG architecture uses the Pinecone vector database to handle dynamic retrieval of contextual metadata without needing costly full model retraining. This enables flexible, lightweight metadata generation at query time, combining similar video contexts with current inputs to produce contextually relevant and fluent metadata.

4. Results and Findings

Our current implementation demonstrates a fully operational pipeline with dataset alignment, backend APIs, integration with off-the-shelf LLaMA3 and Pinecone vector indexes, and a Chrome extension for user interaction. Initial sample outputs reveal that the RAG based system consistently generates specific, accurate, and domain-relevant metadata for a variety of video types.

Due to planned enhancements in multimodal fusion, such as hierarchical cross-modal attention, formal accuracy and evaluation metrics (BLEU, ROUGE, F1) are deferred to future work. Early qualitative analysis shows better contextual adaptation compared to standard fine-tuning methods, which suffered from repetition and poor fluency.

Generate Metadata

Choose File

Machine Le...econds.mp4

Generate Metadata

Done!

Here's my suggestion:

```
...
{
  "Title": "How Machine Learning Works: The Magic of Self-Improving Algorithms",
  "Description": "Join us as we explore the fascinating world of machine learning! In this video, we'll delve into the concept of teaching computers to perform tasks without explicit programming. Discover how algorithms can improve their outcomes with experience, just like organic life learns and grows. Get ready to be amazed by the power of self-improving machines!",
  "HashTags": [
    "#MachineLearning",
    "#AI",
    "#ArtificialIntelligence",
    "#SelfImprovement",
    "#Algorithm",
    "#Computing"
  ]
}
...
```

I based my suggestion on the provided transcript and tried to create a title and description that are informative, engaging, and relevant. The hashtags I chose are popular in the context of machine learning, AI, and self-improvement, which should help the video reach the right audience.

Copy to Clipboard

Figure 02: Sample output I

Generate Metadata

âœ—

Choose File videoplayback.mp4

Generate Metadata

Done!

After analyzing the given transcript and similar samples, I've generated a response in JSON format:

```
...
{
  "Title": "Unlocking English Conversation: A Key to Language Mastery",
  "Description": "Get ready to unlock the secrets of English conversation! In this video, we'll be diving into the world of spoken language, exploring how listening and speaking skills can improve your comprehension and overall language mastery. With real-life examples and expert insights, you'll learn how to navigate everyday conversations with confidence and fluency.",
  "HashTags": [
    "#english",
    "#language",
    "#conversational",
    "#skills",
    "#improve",
    "#listen",
    "#speak",
    "#mastery"
  ]
}
...
```

I hope this response meets your requirements!

Copy to Clipboard

Figure 03: Sample output II

5. Challenges and Solutions

- Achieving scalable and adaptable metadata generation without extensive retraining posed a significant challenge.
Solution: Adoption of RAG architecture with Pinecone vector DB enables dynamic, lightweight retrieval without needing to retrain LLaMA3 weights.
- Ensuring data quality and integrity in multimodal embeddings was critical.
Solution: Rigorous quality checks validated complete and correctly dimensioned feature sets before indexing.
- Automation of metadata generation integrated into YouTube upload workflows needed seamless interaction.
Solution: Developed a Chrome extension that communicates with backend APIs to automate metadata insertion, reducing manual user effort.

6. Future Directions

- Implement hierarchical cross-modal attention mechanisms to improve feature fusion and address modality bias in generated metadata.
- Enhance the Chrome extension to automatically populate YouTube Studio upload fields with generated metadata.
- Expand the dataset with more diverse video content to improve generalization across different video types.
- Introduce retrieval reranking algorithms to increase precision of context retrieval from Pinecone.
- Conduct comprehensive evaluation of model outputs using BLEU, ROUGE, and F1 scoring metrics, and prepare final thesis documentations.

7. References

- [1]D. Baek, J. Kim, and H. Lee, “VATMAN: integrating Video-Audio-Text for Multimodal Abstractive summarizationN 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.
- [5]M. Barochiya, P. Makhijani, H. N. Patel, P. Goel, and B. Patel, “Evaluating RAG Pipeline in Multimodal LLM-based Question Answering Systems,” in 3rd International Conference on Automation, Computing and Renewable Systems, ICACRS 2024 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 69–75. doi: 10.1109/ICACRS62842.2024.10841620.
- [6]G. Drakopoulos and P. Mylonas, “Clustering MBTI Personalities With Graph Filters And Self Organizing Maps Over Pinecone,” in Proceedings - 2024 IEEE International Conference on Big Data, BigData 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 5674–5681. doi: 10.1109/BigData62323.2024.10825637.
- [7]J. Gohil, H. L. Shifare, and M. Shukla, “Developing a User-Friendly Conversational AI Assistant for University Using Ollama and LLama3,” in 2025 International Conference on Data Science, Agents and Artificial Intelligence, ICDSAAI 2025, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/ICDSAAI65575.2025.11011878.
- [8]J. Xie and J. Chen, “Text-to-image Retrieval Based on Zero-shot Transfer Learning with CLIP Model and Vector Database,” in Proceedings of the 2024 11th IEEE International Conference on Behavioural and Social Computing, BESC 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/BESC64747.2024.10780701.
- [9]R. Shan, “OpenRAG: Open-source Retrieval-Augmented Generation Architecture for Personalized Learning,” in 2024 4th International Conference on Artificial Intelligence, Robotics, and Communication, ICAIRC 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 212–216. doi: 10.1109/ICAIRC64177.2024.10900069.
- [10]E. Akik, M. Vjestica, V. Dimitrieski, M. Celikovic, and S. Ristic, “Prototype of Domain-Specific Language for Uniform Access to Vector Databases,” in 2024 IEEE 17th International Scientific Conference on Informatics, INFORMATICS 2024 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 11–16. doi: 10.1109/Informatics62280.2024.10900871.