Scene-Based Movie Search using Multi-Modal Embeddings

Arun Karthik Sengottuvel, Ashwinkumar Venkatnarayanan, Rohil Wattal,
Nitin Sairaj Paruchuri, Arun Prakash Jayakanthan
University of Southern California

1 Introduction

In recent years, movie search has become an essential tool for navigating the extensive and continually expanding landscape of film content. Think about the last time you couldn't recall a movie title but vividly remembered a specific scene – this is a common experience for many. Traditional movie search engines often rely on actors or titles, failing to capture the rich narrative structure and nuances of movies. We require tools capable of comprehending and responding to natural language descriptions of scenes provided by users, enabling them to search for movie scenes using their own words.

While a few scene-based search functionalities have emerged, such as the approach proposed by Khan et al. (2020), they often struggle with capturing the full context and details of each scene due to limitations in paraphrasing techniques and classification biases (Khan et al., 2020). Paraphrasing techniques employed by these systems may struggle to capture the diversity of user queries, particularly when specific terms are involved. Furthermore, the classification approach often exhibits biases towards movies with fewer scenes or those with similar scenes, impacting the accuracy and relevance of the search results. These challenges underscore the need for more sophisticated approaches, such as leveraging multi-modal embeddings, as introduced by Bain et al. (2020), to overcome the inherent limitations of current scene-based movie search systems.

By leveraging multi-modal information like character recognition, speech analysis, and visual elements, our project strives to enhance the accuracy and depth of scene-based movie searches. Character recognition identifies textual elements in scenes, enabling users to search for movies based on dialogues, quotes, or textual content. Speech analysis enhances our system's ability to interpret spoken scene descriptions, accommodating those who prefer verbal communication. The inclusion of visual elements in our multi-modal approach ensures a more accurate scene understanding by analyzing objects, settings, or visual characteristics, overcoming paraphrasing limitations. Users can now explore movies based on their memorable moments, empowering a more personalized and engaging cinematic experience.

The potential impact of our project extends beyond traditional movie searches, facilitating applications in semantic video summarization, automatic video description for the visually impaired, and intelligent fast-forward functionalities.

2 Related Work

The evolution of scene-based movie search has been significantly influenced by advancements in understanding narrative structures and character dynamics within movies. The work on "Condensed Movies: Story Based Retrieval with Contextual Embeddings" by Bain et al. (2020) represents a pivotal contribution to this field. By focusing on key scenes from over 3,000 movies, the authors have developed a comprehensive dataset that includes high-level semantic descriptions, character face tracks, and other metadata. This approach not only addresses the challenges of semantic understanding in human narratives but also provides a scalable solution obtained from freely available YouTube content.

Parallel to the advancements in movie scene understanding, the field of NLP has seen considerable progress in evaluating machines' Theory of Mind (ToM), particularly through tasks designed to assess the understanding of beliefs, desires, and intentions within language. While these tasks offer insights into LLMs' emerging ToM capabilities, as noted by Kosinski (2023), they often rely on synthetic settings that may not capture the full complexity of character interactions and their development over time.

Furthermore, the exploration of fictional character understanding, as discussed by Brahman et al. (2021) and Sang et al. (2022), highlights the importance of assessing various dimensions of characters' mental states for a comprehensive narrative understanding. These studies underscore the necessity of moving beyond synthetic datasets to include more nuanced representations of character interactions and their evolution within stories.

Our project on Scene-Based Movie Search using Multi-Modal embeddings builds upon these foundational insights by integrating multi-modal information such as character recognition, speech analysis, and visual elements. This approach allows for a more nuanced understanding of movie scenes, enabling users to search for scenes based on a wide range of descriptors that capture the essence of cinematic narratives.

3 Dataset

The Condensed Movie Dataset (CMD) is a repository to facilitate advanced machine comprehension of narrative structures within lengthy cinematic works. Boasting an impressive collection of over 30,000+ clips extracted from approximately 3500+ movies, CMD offers a condensed yet comprehensive glimpse into the pivotal moments of each film. These clips, typically spanning around two minutes each, encapsulate the essence of the storyline, effectively distilling the rich and complex narratives into bite-sized segments.

What sets CMD apart is its meticulous attention to detail in both video content and accompanying textual descriptions. Each clip is not only meticulously selected to represent the most salient parts of the movie but also complemented by high-level descriptions that delve into the character's intentions, emotions, relationships, and overarching thematic elements. Moreover, face-tracks and identity labels for main characters further enrich the dataset, providing valuable insights into character interactions.

One of CMD's standout features is its online longevity and scalability. By sourcing videos from the licensed YouTube channel MovieClips by Fandango, the dataset ensures continuous accessibility and expansion. Unlike many YouTube datasets that suffer from rapid data loss due to videos being taken down by users, CMD is much more stable.

In terms of story coverage, CMD offers a compelling insight into the effectiveness of its condensed clips. Despite representing only 15% of the full-length movie duration, CMD clips cover a remarkable 44% of the full plot sentences, indicating their ability to capture the essence of the storyline. This level of coverage, coupled with the dataset's median span of 85.2% of the plot, underscores the efficacy of CMD in distilling key narrative elements from lengthy films.

Overall, CMD not only serves as a valuable resource for video-text retrieval tasks but also paves the way for broader exploration and understanding of long-range movie narratives.

4 Project Plan

March

Week 1	Finalize the literature review and identify relevant research papers on scene-based movie search and multi-modal embeddings.
Week 2	Gather and preprocess movie datasets for training and testing.
Week 3	Develop and fine-tune the multi-modal embeddings model for scene-based movie search.
Week 4	Implement character recognition, speech analysis, and visual element integration into the model.

Table 1: Project Timeline for March

April

Week 1	Train the model on the collected datasets and validate the model performance using validation datasets.
Week 2	Evaluate the model's accuracy, relevance, and efficiency in scene-based movie searches. Optimize the model based on evaluation results to enhance search capabilities.
Week 3	Prepare the final project report and presentation materials.

Table 2: Project Timeline for April

References

Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. 2020. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision*.

Umair Al Khan, Miguel A. Martinez-Del-Amor, Saleh M. Altowaijri, Adnan Ahmed, Atiq Ur Rahman, Najm Us Sama, Khalid Haseeb, and Naveed Islam. 2020. Movie tags prediction and segmentation using deep learning. *IEEE access*, 8: 6071-6086.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. Let Your Characters Tell Their Story: A Dataset for Character-Centric Narrative Understanding. *arXiv preprint arXiv:2109.05438*.

Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. 2022. Tvshowguess: Character comprehension in stories as speaker guessing. *arXiv preprint arXiv:2204.07721*.