Scene-Based Movie Search using Multi-Modal Embeddings

Arun Karthik Sengottuvel, Ashwinkumar Venkatnarayanan, Rohil Wattal,
Nitin Sairaj Paruchuri, Arun Prakash Jayakanthan
(sengottu, venkatna, rwattal, nparuchu, jayakant)@usc.edu
University of Southern California

Abstract

Searching for movies has become crucial for exploring the vast film content. In this study, we propose a Scene-Based Movie Search framework leveraging Multi-Modal Embeddings, integrating textual data and video embeddings. Traditional movie search engines often struggle to capture the rich narrative structure and nuances of movies, relying primarily on actors or titles. Our project addresses these limitations through advanced methodologies for semantic similarity analysis and ranking. Results show that incorporating video embeddings alongside textual data significantly improves accuracy metrics, providing users with a more immersive and personalized search experience. This approach not only enriches the understanding of plot descriptions but also captures visual elements, actions, and scenes, thereby enhancing the overall search capability. Additionally, our study sheds light on the potential impact of leveraging multi-modal embeddings in scene-based movie search, offering insights into the future of movie exploration and recommendation systems.

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. 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. This study aims to address this need by proposing a Scene-Based Movie Search framework that leverages Multi-Modal Embeddings. By integrating textual data and video embeddings, our objective is to enhance the accuracy and depth of movie searches, offering users a more immersive and personalized experience. Character recognition enables the identification of textual elements within scenes, allowing users to search for movies based on dialogues, quotes, or textual content. The inclusion of visual elements ensures a comprehensive scene understanding by analyzing objects, settings, and visual characteristics, overcoming paraphrasing limitations.

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.

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 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.

3 Dataset

3.1 Data Collection

We obtained textual data by scraping plot summaries of the highest-grossing movies from Wikipedia pages using the Wikipedia API. The scraped data was then stored in JSON format for further processing. Additionally, for visual data, we retrieved video clips from the "HowTo100M" dataset, which consists of 136 million clips from instructional videos, providing a rich pre-trained dataset for various applications.

3.2 Data Preprocessing

In the preprocessing stage, we performed several steps to clean and prepare the data for analysis. First, we implemented sentence tokenization using the NLTK package, dividing paragraphs into individual sentences. Next, we applied lemmatization to normalize the text data, reducing inflected words to their base or dictionary form. Finally, we removed any NaN (Not a Number) values to ensure data integrity and consistency throughout the dataset.

3.3 Data Augmentation

For data augmentation, we employed two techniques to enhance the dataset's diversity and quality. Firstly, we utilized the Pegasus-LARGE model, as proposed by Zhang et al. (2020), for summarization purposes. This involved generating concise summaries of movie plots to address large plot sizes effectively. Secondly, we employed the Parrot Paraphraser to generate query sentences from movie plots, creating diverse query variations for both training and testing purposes.

3.4 Video Retrieval

We retrieved video embeddings using the below feature extraction methods:

a) Spatial Feature Extraction: Converts sampled frames into 2D patches, then into 1D tokens, using multi-head self-attention, MLP, and layer-norm in stacked encoder-decoder blocks.

b) Temporal Feature Extraction with Transformers: Captures temporal relations by modeling interactions between frames, encoding features, and aggregating them into comprehensive representations.

c) Multi-modal Video Feature Extraction: Integrates scene embeddings from DenseNet-161, face features from ResNet50, and motion features from networks like S3D and SlowFast, enabling analysis of scene context, facial expressions, and dynamic movements within videos.

4 Methodology

4.1 Scene Separation

Plots underwent separation into scenes at both paragraph and sentence levels to offer flexibility and granularity in representing the narrative structure of the movies.

4.2 Semantic Similarity Analysis

For semantic similarity analysis, three approaches were utilized:

a) BERT + Cosine Similarity: This method leveraged BERT embeddings and cosine similarity to capture semantic relationships more accurately, benefiting from BERT's contextual understanding of text. The embeddings, with a dimensionality of 768, were summed from the last 4 layers of the BERT model.

b) BERTScore: Semantic similarity was evaluated using F-1 scores, enabling the capture of subtle semantic nuances effectively.

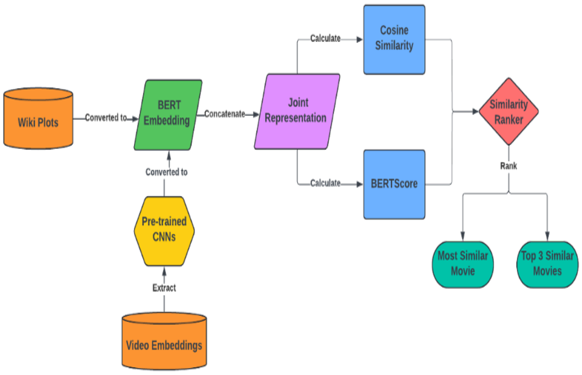

Figure 1: Working of BERT + Cosine Similarity and BERTScore

c) RoBERTa: The similarity problem was reframed as a binary classification task using RoBERTa (Liao et al., 2021) as depicted in Figure 2. The roberta-base variant, a transformer-based model architecture, was employed for training. RoBERTa was fine-tuned for sentence pair classification using a pairwise dataset created through manual annotation and weak supervision. The training involved an 80:20 train/test split over 1000 pairs and was conducted for 100 epochs using the AdamW optimizer.

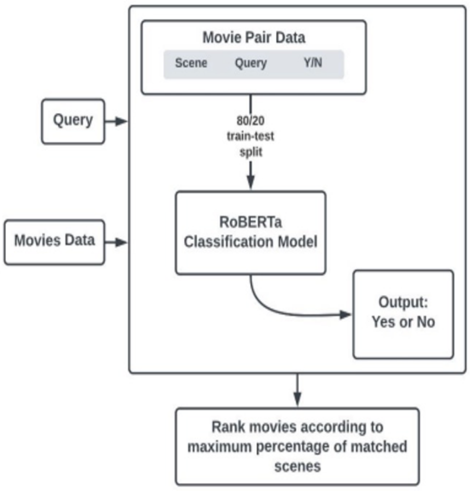
Movie Pair Data
Scene
Query
Y/N
80/20 train-test split
Query
Movies Data
RoBERTa Classification Model
Output: Yes or No
Rank movies according to maximum percentage of matched scenes

Figure 2: Working of RoBERTa Classification Model

4.3 Feature Embedding and Matching

This step aligned video and textual features in a joint embedding space for similarity calculation. It categorized features into Global, Local, and Individual categories to enhance feature matching accuracy. Global matching utilized linear projection and cosine similarity to align coarse-grained features, enabling effective comparison between video and text for tasks like video-text retrieval and understanding. Local matching, on the other hand, divided features into fine-grained hierarchies, allowing for more detailed comparison and improved feature alignment.

4.4 Ranking

Two metrics were employed for ranking:

a) Similarity Score: This metric measured the similarity between an input query and any scene in the dataset, providing a quantitative assessment of how closely a movie matched the query.

b) Percentage of Positive Classifications: Our trained model predicted binary labels (Y/N) for each query-scene pair. The percentage of positive classifications for each movie was aggregated to determine the likelihood of a movie being the best match for the query based on the proportion of positive classifications.

5 Results and Analysis

5.1 Results

Tables 1 & 2 present the accuracy metrics for Semantic Similarity approaches when utilizing only Wiki Plots (Textual Data) and when incorporating Video Embeddings (Visual Data) alongside textual information.

Accuracy	**Paragraph Test Set**		**Sentence Test Set**	
	Top-1 Acc (%)	**Top-3 Acc (%)**	**Top-1 Acc (%)**	**Top-3 Acc (%)**
BERT+Cosine	46.72	59.92	43.39	57.82
BERTScore	51.15	64.78	50.5	61.19
RoBERTa	52.56	68.23	50.15	63.89

Table 1: Accuracies when using only Wiki Plots (Textual Data)

Accuracy	**Paragraph Test Set**		**Sentence Test Set**	
	Top-1 Acc (%)	**Top-3 Acc (%)**	**Top-1 Acc (%)**	**Top-3 Acc (%)**
BERT+Cosine	71.22	79.24	69.13	77.42
BERTScore	76.54	81.32	75.36	80.2
RoBERTa	77.63	84.73	74.3	82.56

Table 2: Accuracies when using Wiki Plots (Textual Data) + Video Embeddings (Visual Data)

From the results, it's evident that incorporating

Video Embeddings alongside textual data significantly improves the accuracy of Semantic Similarity approaches. Across both the Paragraph and Sentence Test Sets, all approaches demonstrated higher Top-1 and Top-3 Accuracy when leveraging multimodal information. Notably, RoBERTa achieved the highest accuracy scores in both scenarios, highlighting its effectiveness in capturing semantic relationships within movie plots. By leveraging multimodal information, our project enhances the accuracy and depth of scene-based movie searches.

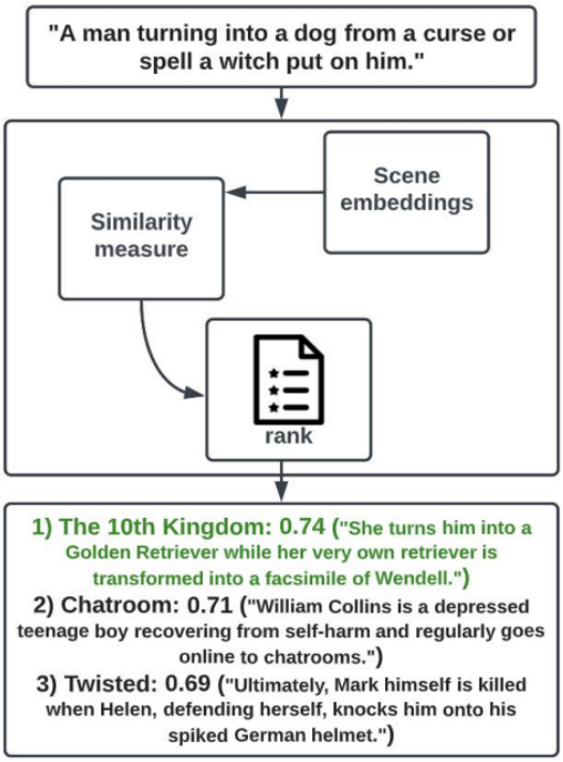
"A man turning into a dog from a curse or spell a witch put on him."
Scene embeddings
Similarity measure
rank
1) The 10th Kingdom: 0.74 ("She turns him into a Golden Retriever while her very own retriever is transformed into a facsimile of Wendell.")
2) Chatroom: 0.71 ("William Collins is a depressed teenage boy recovering from self-harm and regularly goes online to chatrooms.")
3) Twisted: 0.69 ("Ultimately, Mark himself is killed when Helen, defending herself, knocks him onto his spiked German helmet.")

Figure 3: Sample Output

5.2 Analysis

The integration of video embeddings alongside textual data enhances the contextual understanding and depth of scene-based movie searches. Video embeddings provide additional contextual information, enriching the understanding of plot descriptions and capturing visual elements, actions, and scenes not fully represented in textual plots alone. This supplementary information enables models to make more informed similarity judgments, resulting in higher accuracy. Methods like BERT + Cosine and RoBERTa benefit from contextual embeddings derived from both textual and visual inputs, facilitating more accurate semantic similarity assessments by capturing nuanced relationships between words and scenes. Moreover, BERTScore, which evaluates semantic similarity based on F-1 scores considering precision and recall, becomes more adept at discerning semantic nuances when combined with video embeddings, thereby enriching visual content and improving overall accuracy.

6 Conclusion

The project achieves improvements in accuracy and relevance in scene-based movie search systems by leveraging state-of-the-art techniques such as RoBERTa and video embeddings. Our implemented ranking methodology, utilizing BERT embeddings and RoBERTa classifications, enabled the identification of the most likely movies based on input queries. Furthermore, the integration of video embeddings substantially expanded the project's scope, enhancing the capabilities of scene-based movie searches to provide users with a more immersive and personalized experience. Our model achieved an accuracy rate of 52.56% for top-1 accuracy and 68.23% for top-3 accuracy using textual data alone. With the incorporation of video embeddings, these accuracy rates saw significant improvements, soaring to 77.63% for top-1 accuracy and 84.73% for top-3 accuracy.

7 Future Work

Our scene-based movie search using multi-modal embeddings has shown promising results, but there are several areas for future exploration and improvement. Firstly, expanding the size and diversity of the movie database by incorporating a wider range of genres and production years will enhance the robustness and applicability of the model to various user preferences and search scenarios. Additionally, integrating audio embeddings to capture relevant information from movie dialogues, background music, and sound effects can provide a more comprehensive representation of movie scenes and improve the accuracy of the search system. Furthermore, recognizing the global nature of the movie industry, expanding the model's language support beyond English can make the scene-based movie search accessible to a wider international audience. By addressing these future work items, the functionality, accuracy, and user experience of scene-based movie search systems can be enhanced, making them valuable tools for movie enthusiasts and content creators alike.

8 Division of Labor

- Data Collection – Arun Prakash and Nitin Sairaj
- Data Preprocessing – Arun Prakash and Nitin Sairaj
- Data Augmentation – Arun Prakash and Nitin Sairaj
- Video Embeddings Retrieval and Incorporation – Rohil
- Scene Separation – Ashwinkumar and Arun Karthik
- Feature Embedding and Matching – Rohil
- Training Models – Arun Karthik
- Evaluating Semantic Similarity and Ranking – Ashwinkumar
- Documentation – All

We acknowledge the support and guidance provided by the course staff of CSCI 544 throughout our project during the Spring 2024 semester at the University of Southern California.

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.

Arguello, J., Ferguson, A., Fine, E., Mitra, B., Zamani, H. and Diaz, F., 2021, March. Tip of the tongue known-item retrieval: A case study in movie identification. In Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (pp. 5-14).

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.*
Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International conference on machine learning (pp. 11328-11339). PMLR.

Wenxiong Liao, Bi Zeng, Xiuwen Yin, and Pengfei Wei. 2021. An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa. Appl Intell 51, 3522–3533 (2021).