Unsupervised Multi-Modal Video Summarization Report

Papagrigoriou Vasileios Savvas
Auth

Overview

Unsupervised Multi-Modal Video Summarization is a cutting-edge project at the intersection of computer vision and machine learning. Its primary aim is to automate the process of condensing lengthy videos into succinct summaries without the need for pre-labeled training data, leveraging advancements in object detection and unsupervised learning algorithms.

Objectives

- Provide an efficient means of summarizing long videos.
- Utilize state-of-the-art object detection techniques in understanding video content.
- Employ unsupervised learning methods for the automatic generation of video summaries.

Methodology

Object Detection

Tool Used: YoloV3, renowned for its accuracy and speed in real-time object detection.
Process: Analyzes video frames to identify and categorize various objects.
Output: A comprehensive list of detected objects, including their types and coordinates within each frame.

Video Summarization

Dataset: Incorporates insights from the TVsum50 dataset, a widely recognized benchmark in the video summarization field.
Approach: Uses patterns and trends identified in TVsum50 to determine the most critical segments of a video.

Implementation

objectDetection.py

Function: Processes input videos to detect objects using YoloV3.
Location: Script reads videos from input_videos/ and outputs data to be used by the summarization module.

videoSum.ipynb

Core: Central component for the video summarization task.
Integration: Utilizes output from objectDetection.py along with insights from TVsum50 to create video summaries.
Features: Includes interactive elements and visualizations for a better understanding of the summarization process.

Expected Results

- **Efficiency:** Ability to process various video types, providing quick and accurate summaries.
- **Effectiveness:** Summaries that capture the essence of the original content, focusing on key events and objects.
- **User Engagement:** Improved user experience by providing concise yet comprehensive summaries.

Advanced Methodological Insights

This section delves deeper into the core of the project, highlighting the intricate details and the thought process behind various methodological choices.

The Objective of Automated Trailer Generation:

- The primary aim is to automate the generation of a concise trailer from a longer video using unsupervised learning, with a focus on KNN algorithms.
- This process is underpinned by the YoloV3 model for image object detection and the integration of multiple features like visual and audio components.

Integration of Visual and Audio Features:

- Visual Features: Extracted using the VGG16 model for each frame of the video.
- Audio Features: Derived using the MFCC algorithm to capture the audio essence of the video.
- The incorporation of LDA (Linear Discriminant Analysis) was experimented with but yielded inferior results compared to using only PCA (Principal Component Analysis).
- PCA is employed not only for dimensionality reduction but also to visualize data clusters for better understanding and selection.

Object Encoding and Clustering:

- Object Detection: Objects in video frames are detected and encoded using binary encoding (e.g., 001, 010, 100) to facilitate distinction in the KNN algorithm.
- One Hot Encoding was tested alongside binary encoding, resulting in similar outcomes.
- K-Means Clustering: The method of determining the number of clusters (n_clusters) in K-Means is crucial. Currently, the project uses the square root of the square root of the number of features, which, while not the most efficient, provides better performance than methods with larger separations (i.e., more clusters).

Challenges in Frame Selection and Video Summarization:

- The selection of frames for creating a summarized video is still an area of exploration. The current method involves choosing frames from each cluster and creating a 15-second video based on their importance and the knapsack algorithm.
- The best video summary is identified by comparing each generated summary with the ground truth and selecting the one with the highest F-score (macro).
- A challenge arises in videos without ground_truth, as there's no ground truth for comparison, making the selection process more subjective.

Best Results and Performance Metrics

vID	Thres	Precision	Recall	F1	Prop_impo
8	3	0.821229	0.7	0.755784	0.821229
2	3	0.167939	0.104762	0.129032	0.167939
4	2	1	0.0971698	0.177128	1
6	2	0.437778	0.0619497	0.10854	0.437778
9	2	1	0.0327044	0.0633374	1
0	1	1	0.0231481	0.0452489	1
5	2	0.12	0.0169811	0.0297521	0.12
7	1	1	0.0157697	0.0310497	1
10	2	0.192771	0.00503145	0.00980693	0.192771
1	2	0.00888889	0.00125786	0.00220386	0.00888889
3	0	0	0	0	0

TABLE I: Performance Metrics for Various Videos

vID	Thres	Precision	Recall	F1Macro	Prop_importance
8	3	0.821229	0.7	0.87061	0.821229

TABLE II: Performance Metrics using Macro (Only first result)

Update: 27/11/23

- Reinforcement Learning: There is consideration to train a model with the current data using Reinforcement Learning to predict ground truths. However, the limited size of the TVSum dataset (50 videos) may impede the effectiveness of this approach.
- Frame Selection and Cluster Determination: The project is actively exploring efficient techniques for optimal frame selection and cluster determination, especially in the absence of annotations.
- Best Number of Clusters: Finding the ideal number of clusters in K-Means clustering is an ongoing challenge.
- Alternative Models: Exploring other models, such as SVM, for improved results on unknown videos is under consideration.

RECENT DEVELOPMENTS

Update: 11/12/2023

Code Enhancements and Functional Updates: Significant improvements have been made to the codebase since the last documented update. These enhancements aim to refine the model's performance and extend its capabilities. The following list highlights the key developments:

- **Neural Network Integration:** Experimentation with various neural network models has been conducted to improve the summarization process.
- **Dynamic Annotation:** Introduction of a dynamic annotation system to facilitate better labeling of the dataset.
- **Debugging Tools:** Development of tools for saving and previewing data for each video, enhancing the debugging process.
- **Text and Object Encoding:** The integration of tokenization for video titles and one-hot encoding for object detection has been completed to improve feature representation.
- **Object Detection Model Update:** The transition from YOLOv3 to YOLOv5 for object detection aims to utilize advancements in accuracy and speed.

Update: 18/12/2023

Refinements in Video Summary Creation:

- **Importance Calculation:** The significance of each frame is now determined based on the Manhattan distance between the vectors of the video title and detected objects. This approach aims to align the frame's content more closely with the overall theme of the video.
- **Tokenizer Usage:** BERT tokenizer is employed for generating vectors.
- **Reassessment of Clustering Algorithm:** Moved away from Isodata due to its limitations. New clustering techniques are being explored, especially those better suited for handling the dynamic nature of video data.
- **Data Processing Enhancements:** Integration of PCA and LDA for improved feature extraction and representation.
- **Clustering Scope:** Focus on a small number of clusters (3 to 5) for a more targeted summarization approach.
- **Cluster Selection:** Utilizing best score based on changing points for cluster selection, though the effectiveness of this approach is under review.
- **Post-evaluation Concerns:** Unsure if the current post-evaluation approach fully meets the project's needs. Open to revisiting this to ensure it aligns more accurately with our objectives.

Help needed with the following::

- **Evaluation Strategy:** The current strategy heavily depends on selecting the best cluster for video summarization. There's an ongoing discussion on whether a more complex evaluation framework could yield better results, considering aspects such as changing point evaluations and the effectiveness of Isodata hyperparameters.
- **Title-Based Importance Limitations:** The method of determining importance based on the Manhattan distance between titles and objects might not always yield accurate results. For example, if the video title is "Bert's Manscaping" and the detected objects are "person" and "table" the current system may struggle to establish relevant connections.
- **Suggestions for Improvement:** Exploring more advanced techniques or integrating additional layers of

analysis (such as contextual understanding or sentiment analysis) could enhance the evaluation process. Feedback and ideas for refining the methodology are highly appreciated.