Video Search using Natural Language

Badr AlKhamissi, Karim Hasebou, Muhammed Nael, Hussam Ashraf

Computer Science, The American University in Cairo

Abstract

In this research, we propose a framework for searching long untrimmed videos for segments that logically correlate with a natural language query. We develop a new method that exploits state-of-the-art deep learning models on the temporal-action-proposal task and dense captioning of events in videos to be able to retrieve video segments that corresponds to an input query..

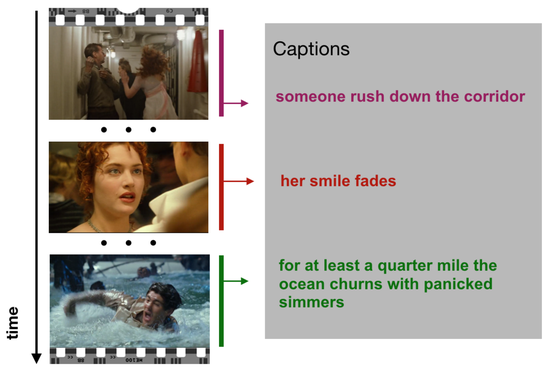

Figure 1: Captions describing scenes from Titanic (1997)

Introduction

Videos are becoming the dominant means of sharing information online; as they are more natural to consume, easier to digest, more diverse in content and more clear in representing it than other recording media like text and audio. However, to this day there has not been a simple way to access and understand videos for quick retrieval. This research aims to address this problem by enabling users to search untrimmed videos for segments semantically correlating with input queries in natural language. To this end, we identify the following sub-problems: Feature Extraction to encode the raw visuals, Temporal Action Proposal to highlight important segments and thereby reducing the search space, Dense-Video Captioning for describing the video and Sentence Matching to measure the semantic similarity with the generated captions in order to retrieve the desired segment(s).

Feature Extraction

Extracting features from the raw visual information by learning the spatiotemporal relationships across the video using both 3D and 2D deep convolutional neural networks as feature extractors, pre-trained on the Sports-1M and the ImageNet datasets respectively, in order to represent motion and action (the temporal aspect) and appearance (the spatial aspect) simultaneously.

Temporal Action Proposal

This task focuses on generating temporal action proposals from long untrimmed video sequences efficiently in order to consider only segments that likely contain significant events, and thereby reducing the overall search space and avoid indexing irrelevant frames.

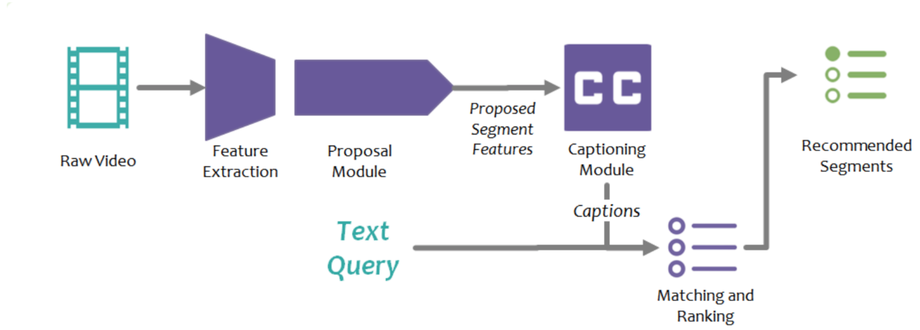

Figure 2: Pipeline

Seq2Seq Model

Inspired from recent success of sequence-to-sequence models in machine translation, this model uses as input a sequence of video frames $\{x_1, ..., x_n\}$ and outputs a sequence of words $\{y_1, ..., y_m\}$. The model estimates the conditional probability of the output sequence given the input sequence:

$$\mathbb{P}(y_1, ..., y_m \mid x_1, ..., x_n) \quad (1)$$

This is done by using a stack of two LSTMs, where the first layer encodes the input sequence to fixed length vector, and the second map the vector to a sequence of outputs.

SoftCap Model

Let $\{v_1, ..., v_n\}$ be the set of visual features in different timesteps. The features are weighted by different attention scores $\alpha_i^{(t)}$ to obtain relevant visual context feature $\phi_t(v)$ when predicting the t-th word:

$$e_i^{(t)} = w^T tanh(W_a h_{t-1} + U_a v_i + b_a) \quad (2)$$

$$\alpha_i^{(t)} = exp\{e_i^{(t)}\} / \sum_{j=1}^{n} exp\{e_j^{(t)}\} \quad (3)$$

$$\phi_t(v) = \sum_{i=1}^{N} \alpha_i^{(t)} \quad (4)$$

$\phi_t(v)$ is then concatenated with the previous word embedding as the input of the LSTM decoder.

Video Indexing/Event Description

Describing videos by densely captioning the events in a given segment into natural language descriptions in order to have a common space between the original visuals and search queries. We experimented with two different models; a sequence-to-sequence model based on S2VT and a model with soft-attention mechanism to attend to relevant temporal information when generating different words.

Matching

Matching and ranking video segments that semantically correlate with search queries. We used a pre-trained Combine-Skip thought model to encode the captions generated by the captioning module and the user's input into vectors, in order to find the semantic similarity between them.

Results

We were able to achieve a CIDEr score of 30 on the ActivityNet Captions dataset using the Seq2Seq model. The proposal module achieved a recall of 0.59 at tIOU of 0.5. We evaluated the results of the SoftCap model on movies from the LSMDC dataset qualitatively and it achieved desirable results.

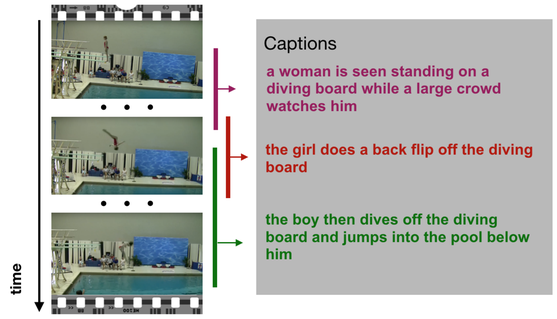

Figure 3: Dense-captions generated using our model

Further Work

Incorporating context from neighboring proposals while captioning each event to capture the correlations between events. Using other methods for describing the video content that might prove desirable for better search, such as considering the video's audio alongside the visual information or identifying the top-k objects within a sequence of frames. Employ different techniques for matching and ranking depending on the structure of event descriptions we generate when indexing. Proposing an evaluation metric for overall search accuracy.

References

Acknowledgements

We would like to thank our supervisor Dr. Mohamed Moustafa for his support during this project.