Semantic Understanding for Contextual In-Video Advertising

Rishi Madhok
Delhi Technological University, India
rishi_bt2k14@dtu.ac.in, +91-9811481780

Shashank Mujumdar, Nitin Gupta, Sameep Mehta
IBM Research, India
{shamujum, ngupta47, sameepmehta}@in.ibm.com

Abstract

With the increasing consumer base of online video content, it is important for advertisers to understand the video context when targeting video ads to consumers. To improve the consumer experience and quality of ads, key factors need to be considered such as (i) ad relevance to video content (ii) where and how video ads are placed, and (iii) non-intrusive user experience. We propose a framework to semantically understand the video content for better ad recommendation that ensure these criteria.

Introduction

More often than not, while watching a video online, viewers may be interrupted by untimely as well as irrelevant advertisements that hamper their viewing experience. To account for this problem, different methods to recommend context-aware ads have been proposed (Mei, Hua, and Li 2009). However, "context" is generally inferred through text associated with the video or w.r.t pre-defined concepts (Car, Food, etc.) identified in the video. Other methods to recommend video ads based on biddings (Zhang, Yuan, and Wang 2014), sentiments (Vedula et al. 2017), and user's past history (De Bock and Van den Poel 2010) have also been explored. However, such information is not sufficient nor necessary to insert context aware ads in videos.

We define context to be the semantic understanding of the different events in the video through analysis of the video frames, text extracted from speech and associated meta data, and audio extracted from the video. Given a video of a person driving a car that meets an accident, a semantic understanding of the event would suggest a car/life insurance ad as opposed to a generic advertisement of a car. At the same time, semantic understanding of the video also helps to decide where and how an ad is inserted in the video to ensure a non-intrusive user experience.

We consider the following three scenarios to understand the context in videos and recommend/insert relevant ads: (a) concept identification - identify the central entity (object, celebrity, location etc.) in the video scene, (b) concept + event understanding - identify the event associated with the central concept, and (c) theme understanding - identify the theme of the video scene (happy, adventure, death, crime etc.). A given video can be segmented into meaningful chunks with standard scene detection methods. In this paper, we propose a new method which uses multi-modal information to semantically understand the video content, and recommend/insert ads in a video such that the inserted ads are context-aware and do not interrupt the user experience.

The problem of inserting context-aware ads in videos has been explored extensively in literature. (Mei, Hua, and Li 2009) proposed a framework to insert ads in videos based on global textual relevance gathered from video meta data and local visual-aural relevance through low-level image and audio features. Similar approach is utilized by (Xiang, Nguyen, and Kankanhalli 2015) where the context is established through textual meta data and visual saliency is maintained through high-level features extracted from deep neural networks. (Sengamedu, Sawant, and Wadhwa 2007) propose a method to identify appropriate ad-insertion points within a video and take into account face/object detection to associate context relevant ads.

The major limitation of the methods in prior art is the lack of analysis of the video content to identify the context for inserting better ads. A semantic understanding obtained through multi-modal analysis of the textual meta data and the video content is necessary to recommend/insert the most relevant ads for a non-intrusive user experience and to the best of our knowledge it has not been addressed in prior art.

Approach

In order to identify the semantic meaning from different sources of information associated with the video, we extract the following features from the video (a) video concepts - detect salient objects in a video, (b) video clip summarization - to understand context/event of a video, (c) speech to text - to understand the theme of a video, and (d) embedded text and other meta data - to augment the feature set.

For the video, we use the pre-trained *darknet* imagenet model for object detection. We identify meaningful clips through the video using the standard python library *pyscenedetect*. We extract key frames from each of the clips and generate captions (Karpathy and Fei-Fei 2015) which are fed into the sequence-to-sequence learning model (Sutskever, Vinyals, and Le 2014) to generate the text summary of the clip. Relevant keywords are extracted using Rapid Automatic Keyword Extraction algorithm (concepts,

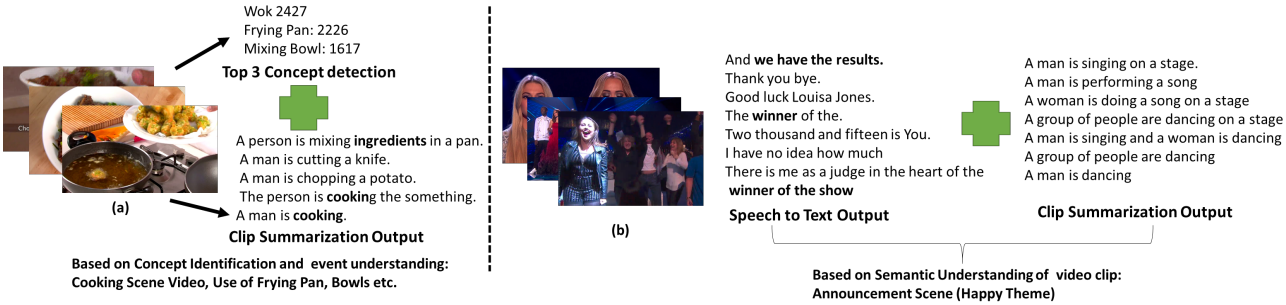

Figure 1: Illustration of the approach for semantic understanding of videos.

entities, locations, events etc.) through text obtained from video meta data, speech-to-text and clip summaries. The keyword set is expanded by fetching synonyms from external data source (For e.g. Thesaurus). Events are specifically identified through clip summaries and themes are identified through speech-to-text. Similar processing is being done on the video ads with the events and theme being provided by advertiser. Based on the frequency count or the keyword match using a global dictionary, between a video ad and a video clip, relevant ads are recommended.

For ad insertion point localization, the candidate insertion points, (say 'n' positions) are detected using the python *pyscenedetect* technique wherein it finds the insertion points based on the difference between two subsequent frames. The most suitable candidate insertion point is chosen for the ad, based on the frequency count of the concept in the subsequent clips detected based on insertion points.

Results and Discussions

For the evaluation of the proposed framework, we have created a database of 50 videos, from YouTube, which are relevant from an advertising point of view. We have manually analyzed the videos and created the list of ads that can be recommended/inserted, based on the presence of salient objects in video frames from categories such as umbrella, shoes, laptop, cab, table, monitor etc.

Given a video, our algorithm first detects the top 3 concepts present in each frame of a video and then based on the frequency count, relevant ads are recommended. We manually annotated the 50 videos with identified objects and ad insertion points w.r.t. advertisements. The performance of our advertisement recommendation system on the dataset of 50 videos is 96%. Our model was able to correctly suggest ads for 48 videos based on the frequency count of the identified objects present in the video, output of clip summarization from video scenes and identification of themes from the speech-to-text output.

Some of the videos in the dataset related to the following two scenarios: concept + event understanding, and theme understanding. The frames in Fig. 1 (a), and (b) are related to cooking, and reality show result announcement videos respectively. Here, the keywords extracted from video summarization and speech to text features help to understand: (i) the event, like in Fig. 1 (a) cooking scene, and in Fig 1. (b) reality show; and (ii) the theme, like in Fig 1. (b) happy theme. On the other hand, the concept detection features helps to extract potential concepts from a video, like in Fig 1. (a) the salient concepts are wok, frying pane, and mixing bowl. The understanding of events in a video frames also helps to associate the ads of concept(s) which are not present directly in a video. For example, cooking event helps to recommend ads related to utensils like microwave etc. The event combined with theme provides more deeper insights about a video, and helps to recommend more relevant ads like chocolate/cold-drinks ads in a happy themed video.

Conclusion and Future Work

In this work, we proposed a context aware ad recommendation/insertion system using multi-modal analytics through semantic understanding of video content. In future, we would like to combine all the proposed feature set in an optimized way to understand semantics of video in an automatic fashion for ad recommendation and validate our results on a larger database of ads and videos.

References

De Bock, K., and Van den Poel, D. 2010. Predicting website audience demographics for web advertising targeting using multi-website click stream data. *Fundamenta Informaticae*.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Mei, T.; Hua, X.-S.; and Li, S. 2009. Videosense: A contextual in-video advertising system. *CSVT*.

Sengamedu, S. H.; Sawant, N.; and Wadhwa, S. 2007. vadeo: video advertising system. In *ACMM*. ACM.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Vedula, N.; Sun, W.; Lee, H.; Gupta, H.; Ogihara, M.; Johnson, J.; Ren, G.; and Parthasarathy, S. 2017. Multimodal content analysis for effective advertisements on youtube. *arXiv*.

Xiang, C.; Nguyen, T. V.; and Kankanhalli, M. 2015. Salad: a multimodal approach for contextual video advertising. In *ISM*. IEEE.

Zhang, W.; Yuan, S.; and Wang, J. 2014. Optimal real-time bidding for display advertising. In *SIGKDD*. ACM.