Webly-Supervised Learning of Multimodal Video Detectors

Junwei Liang, Lu Jiang and Alexander Hauptmann
Carnegie Mellon University
{junweil, lujiang, alex}@cs.cmu.edu

Abstract

Given any complicated or specialized video content search query, e.g. "Batkid (a kid in batman costume)" or "destroyed buildings", existing methods require manually labeled data to build detectors for searching. We present a demonstration of an artificial intelligence application, Webly-labeled Learning (WELL) that enables learning of ad-hoc concept detectors over unlimited Internet videos without any manual annotations. A considerable number of videos on the web are associated with rich but noisy contextual information, such as the title, which provides a type of weak annotations or labels of the video content. To leverage this information, our system employs state-of-the-art webly-supervised learning (WELL) (Liang et al.). WELL considers multi-modal information including deep learning visual, audio and speech features, to automatically learn accurate video detectors based on the user query. The learned detectors from a large number of web videos allow users to search relevant videos over their personal video archives, not requiring any textual metadata, but as convenient as searching on Youtube.

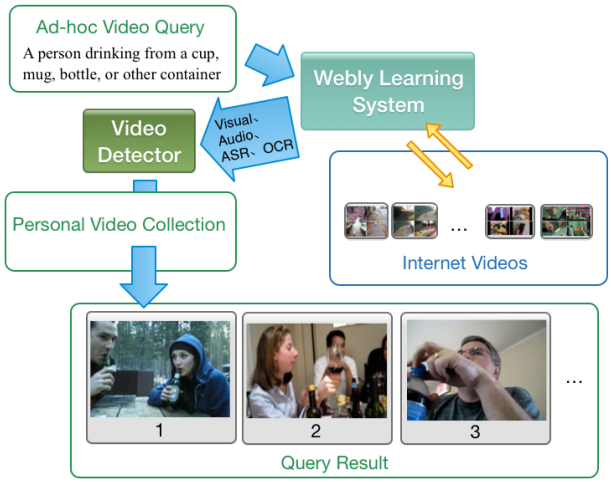

Figure 1: The workflow of WELL

Introduction

Nowadays, millions of videos are being uploaded to the Internet every day. These explosively growing user generated content videos online are becoming an crucial source of video data. Automatically categorizing these web videos into concepts, such as people actions, objects, etc., has become an important research topic. More interestingly, it would be even more valuable to learn detectors from the web videos and apply to personal video collections. Unlike web videos that are usually accompanied with textual labels like titles and descriptions, personal videos or surveillance videos are often stored with no label at all. Therefore traditional text-based indexing methods would not work in such video collections. Recently many work have been proposed to tackle with building concept detectors both in image domain and video domain (Deng et al. 2009; Liang et al. 2015; Karpathy et al. 2014; Jiang et al. 2015b). However, the need for manual labels by human annotators has become one of the major important limitations for large-scale concept learning. In addition, it is difficult for common datasets to include concepts like "Batkid" (a kid in batman costume), which makes it even harder to search these concepts in person collections. Meanwhile, videos are available on the web and contain rich contextual information with a weak annotation about their content, such as their titles, descriptions and surrounding text. These webly-labeled data are orders of magnitude larger than that of any manually-labeled collections. Moreover, automatically extracted features from multiple modalities such as existing still image classification models, automatic speech recognition and optical character recognition tools can be useful additional information for the content of the video.

In this demo, we present our learning framework called **WEbly-Labeled Learning (WELL)** (Liang et al.), extended by utilizing multi-modal representations. The learning framework is motivated by human learning, where people generally start learning easier aspects of a concept, and then gradually take more complex examples into the learning process(Bengio et al. 2009; Kumar, Packer, and Koller 2010; Jiang et al. 2015a). Such learning paradigm is proven to be efficient to deal with noise and outliers.

System Framework

In our demo system, users can provide arbitrary text to query for videos in a target video collection. Figure 1 shows the