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.

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.

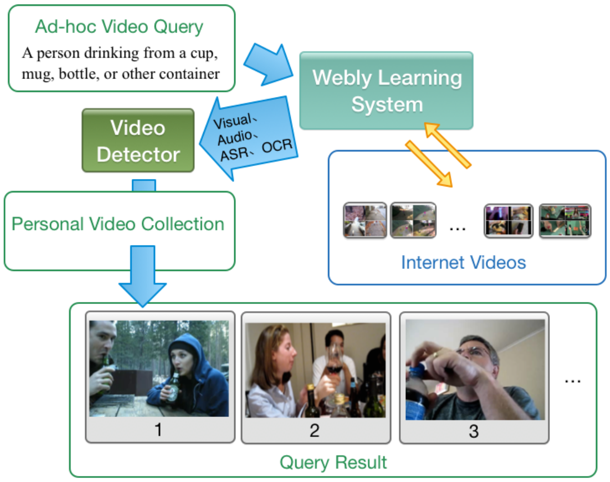
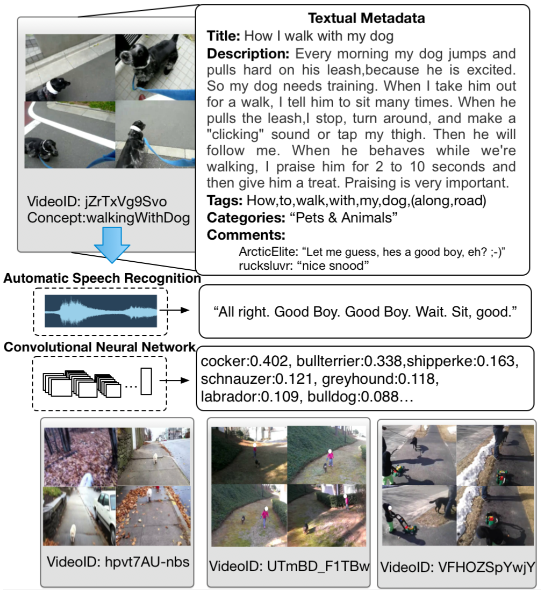


Figure 1: The workflow of WELL

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



Title: How I walk with my dog

Description: Every morning my dog jumps and pulls hard on his leash,because he is excited. So my dog needs training. When I take him out for a walk, I tell him to sit many times. When he pulls the leash,I stop, turn around, and make a "clicking" sound or tap my thigh. Then he will follow me. When he behaves while we're walking, I praise him for 2 to 10 seconds and then give him a treat. Praising is very important.

Tags: How,to,walk,with,my,dog,(along,road)

Categories: “Pets & Animals”

Comments:

ArcticElite: “Let me guess, hes a good boy, eh? ;-)”
rucksluvr: “nice snood”

“All right. Good Boy. Good Boy. Wait. Sit, good.”

cocker:0.402, bullterrier:0.338,shipperke:0.163, schnauzer:0.121, greyhound:0.118, labrador:0.109, bulldog:0.088…

Figure 2: Multi-modal prior knowledge from web videos.

workflow of our system. Our system will first collect related videos from popular video hosting sites like Youtube by using their text search engine and combine these videos to our own video dataset. Then it mainly involves two procedures, curriculum design and model training, to learn multimodal video detectors.

Curriculum Design

In curriculum design, our system extract the multi-modal prior knowledge from the videos to decide which videos are more confident to be positive examples in model training. Figure 2 shows an example of a high-confidence webly-labeled video for the query ”walking with a dog”. As we see, the textual metadata we get from the web videos contain useful but very noisy information. The multi-modal prior information we get is correlated across modalities. Our system first extract bag-of-words features from different modalities and then match them to the query words. Each video will then come with a matching score to the query.

Model Training

After curriculum is extracted, we train detectors using multi-modal webly-labeled learning(Liang et al.). Specifically, we train models based on Convolutional Neural Network (CNN) features, Motion features and MFCC features with late (average) fusion. We use a pre-trained convolutional neural network as the low-level features (VGG network (Simonyan and Zisserman 2014)). We extract the key-frame level features and create a video feature by the average pooling. We automatically generate curriculum labels based on the video metadata, ASR, OCR and VGG net 1,000 classification results using latent topic modeling with word embedding matching. During each iteration of the training, the system combines the multi-modal curriculum with the dynamic information learned from the statistical model (SVM model) to determine which video samples to learn in the next iteration.

Experiments

We compare our method with the state-of-the-art method trained using ground truth labels on FCVID (rDNN) (Jiang et al. 2015b). We compare WELL trained using a single the static CNN features (WELL-CNN), the standard features provided by the authors (Jiang et al. 2015b), and we also compare WELL to the method that achieves the best result on FCVID trained using the same multi-modal features, rDNN-F (Jiang et al. 2015b). Noted that the state-of-the-art method uses the ground truth labels to train models, which includes 42,223 videos with manual labels, while WELL uses none of the human annotation into training but still be able to outperform one of the state-of-the-art results.

Table 1: Ground-truth Training Comparison on FCVID. The methods with * are trained using human annotated labels.

Method	P@5	P@10	mAP
WELL-CNN	**0.918**	**0.906**	**0.615**
Static CNN(Jiang et al. 2015b)*	-	-	0.638
WELL	**0.930**	**0.918**	**0.697**
rDNN-F(Jiang et al. 2015b)*	-	-	0.754

Acknowledgement

This paper was partially supported by the National Science Foundation under Grant No. IIS-1251187. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *ICML*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Jiang, L.; Meng, D.; Zhao, Q.; Shan, S.; and Hauptmann, A. G. 2015a. Self-paced curriculum learning. In *AAAI*.

Jiang, Y.-G.; Wu, Z.; Wang, J.; Xue, X.; and Chang, S.-F. 2015b. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *arXiv preprint arXiv:1502.07209*.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*.

Kumar, M. P.; Packer, B.; and Koller, D. 2010. Self-paced learning for latent variable models. In *NIPS*.

Liang, J.; Jiang, L.; Meng, D.; and Hauptmann, A. Learning to detect concepts from webly-labeled video data.

Liang, X.; Liu, S.; Wei, Y.; Liu, L.; Lin, L.; and Yan, S. 2015. Towards computational baby learning: A weakly-supervised approach for object detection. In *ICCV*.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.