Application and Extensions of CNN-L for Near Duplicate Video Retrieval

Anuja Golechha, Shivanee Nagarajan, Varun Ramesh

February 16, 2019

1) Briefly explain what problem you are trying to solve.

With the advent of social media, growth of videos being shared across the internet has risen exponentially. Current forecast predict that 80 percent of the internet traffic this year will be dominated by videos. A core problem in such a scenario is to deduct duplicate videos from a dataset when a video is passed in as a query. In general duplicate video detection covers the class of problems where we are supposed to find copies of a video that is either transformed photometrically or geometrically from the original video. The current state-of-the-art method is to derive signature from the videos (temporal and spatial statistics) and compare them across videos with a threshold to denote duplication.

A larger class of problem that envelopes the same is to find videos which have the same visual content but with some slight changes - either encoding or color or scaling. The aforementioned class of problems are known as near duplicate video retrieval (NDVR). In this project, we attempt to break down this class of problems by following the model of querying a video database. Given a dataset of videos and a query from a new video, we aim to retrieve a ranked list of near-duplicate videos. Further, we will be attempting a novel extension of this class of problems on low-quality videos/modifying input key-frame extractors in phase 2 and 3 of the project.

2) Why is this problem important? Why are you interested in it?

There is a tremendous interest in applications such as video copy infringement, real-time video search, and recommendations given the growth of video traffic over the last decade or so. Coupled with this growth is the generation of adversarial networks that tweak existing input to fool current state-of-the-art detection techniques. Another possibility is to get away with state-of-the-art techniques by generating low-quality query videos that either impact keyframe extraction or similarity matching step. Hence, the aforementioned challenges and the parallel growth of video traffic leads to near-duplicate video detection as a core problem in today's internet era.
A lot of our interest is generated towards this topic due to two reasons - First, the opportunity to explore applications of core machine learning algorithms such as CNN, DNN in

the field of computer vision evokes great interest in this line of field. Second, the possibility of novel extensions, either by applying new key-frame extraction methods before passing as an input layer to the ML algorithm or by applying new feature extraction methods for low-quality videos provides us with multiple lines of exploration in the future.

3) What is the current state-of-the-art?

Recent years have shown tremendous increase in interest towards Near Duplicate Video Retrieval (NDVR). Multiple state-of-the-art approaches for NDVR are classified based on their level of matching i.e video-level matching, frame-level matching and hybrid-level matching.

Following is a brief summary of these techniques:

Auto Color Correlograms (ACC): [1] This method extracts one frame per second from the original video and ACC of each such frame is computed and aggregated based on visual codebook. The near-duplicate video is identified using TF-IDF weighted cosine similarity over the visual word histograms of a query and a dataset video.

Stochastic Multi-view Hashing (SMVH): [2] It combines multiple keyframe features to learn a group of mapping functions that project video keyframes into the Hamming space. The combination of keyframe hash codes generates a video signature that constitutes the final video representation. A composite Kullback-Leibler (KL) divergence measure is used to compute similarity scores.

Pattern-based approach (PPT): [3] Pattern based indexing tree is build based on the encoding from keyframes. M-pattern- based dynamic programming (mPDP) and time-shift m-pattern similarity (TPS) to determine video similarity.

Layer-wise Convolutional Neural Networks (CNN-L): [4] Frame descriptors are extracted using GoogleNet. A video-level histogram representation derives from the aggregation of the layer vectors to visual words. The similarity between two videos is computed as the tf-idf weighted cosine similarity over the video-level histograms.

Deep Metric Learning (DML): [5] This method leverages Convolutional Neural Network (CNN) features from intermediate layers to generate discriminative global video representations in tandem with a DML framework with two fusion variations, trained to approximate an embedding function for accurate distance calculation between two near-duplicate videos.

From the aforementioned methods, we will be basing our approach on CNN-L. We will tackle NDVR using this approach and take this forward in the next phases in two novel directions. The modifications have been detailed under section 4.

Timeline

Before we detail on how our approach bases/differs from the exiting CNN-L model, we will be describing our timeline in this section.

1. *Completion Date - March 11th*
   Implement existing current state-of-the-art (CNN-L) for Near-Duplicate Video Retrieval using shot boundary detection method as the key frame extractor.

2. *Completion Date - April 3rd*
   In this phase, we will take two diverging approaches before we explore one/both of them in detail. One line of approach is to explore current state-of-the-art keyframe extractors by applying video summarization techniques as a scalable option. The second line of work will involve modifying the module implemented in phase 1 to extract near-video duplicates from low-quality videos.

3. *Completion Date - April 22nd*
   Based on results obtained in this section, phase 3 will build on one/both of the two aforementioned options. In either case, we will attempt to challenge current-state-of-the-art methods using a CNN-L approach with a modified/existing keyframe extractor for low-quality video scenarios. To the best of our knowledge, this is the first attempt of NDVR on low-quality videos.

4) Are you planning on re-implementing an existing solution, or propose a new approach?

Piggybacking on the previous section phase-1 involves re-implementing an existing state of the art approach for NDVR. Phase-2 will involve a novel extension either in terms of generating key frames for NDVR or adapting this approach on low-quality videos. Therefore, our project builds on both of these - we implement an existing solution and then modify/adapt it in different scenarios.

5) If you are proposing your own approach, why do you think existing approaches cannot adequately solve this problem? Why do you think your solution will work better?

Existing approaches rely on using keyframes of the videos for feature extraction. The CC WEB VIDEO [6] dataset contains keyframes which were generated using the shot boundary detection method [7].
The focus of Phase 2 of our project is on using a video summarization technique to generate these keyframes. Since keyframe generation is an integral part of video summarization, we expect those techniques will be highly applicable here and will result in a concise and accurate set of keyframes.
Existing approaches do not address the case of NDVR for low quality videos. In Phase 3, we aim to extend our approach (CNN-L with a customised keyframe generator) to build an NDVR system for low quality videos. The challenge here lies in modifying the existing keyframe generator for low quality videos.

6) How will you evaluate the performance of your solution? What results and comparisons are you eventually planning to show?

We plan to use the CC WEB VIDEO [6] dataset and the mAP (mean Average Precision) metric for evaluation. This is a popular dataset for near-duplicate video retrieval applications.

We plan to compare the result of using CNN-L on the existing keyframe dataset vs using CNN-L on keyframes generated by our chosen video summarization algorithm.

References

[1] Yang Cai, Linjun Yang, Wei Ping, Fei Wang, Tao Mei, Xian-Sheng Hua, and Shipeng Li. Million-scale near-duplicate video retrieval system. In *ACM Multimedia*, pages 837–838, 2011.

[2] Yanbin Hao, Tingting Mu, Richang Hong, Meng Wang, Ning An, and John Y Goulermas. Stochastic multiview hashing for large-scale near-duplicate video retrieval. *IEEE Transactions on Multimedia*, 19(1):1–14, 2017.

[3] Chien-Li Chou, Hua-Tsung Chen, and Suh-Yin Lee. Pattern-based near-duplicate video retrieval and localization on web-scale videos. *IEEE Transactions on Multimedia*, 17(3):382–395, 2015.

[4] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Yiannis Kompatsiaris. Near-duplicate video retrieval by aggregating intermediate cnn layers. In *International conference on multimedia modeling*, pages 251–263. Springer, 2017.

[5] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Yiannis Kompatsiaris. Near-duplicate video retrieval with deep metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 347–356, 2017.

[6] Video dataset cc web video. http://vireo.cs.cityu.edu.hk/webvideo/.

[7] Gautam Pal, Dwijen Rudrapaul, Suvojit Acharjee, Ruben Ray, Sayan Chakraborty, and Nilanjan Dey. Video shot boundary detection: A review. In Suresh Chandra Satapathy, A. Govardhan, K. Srujan Raju, and J. K. Mandal, editors, *Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2*, pages 119–127, Cham, 2015. Springer International Publishing.