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

Anuja Golechha, Shivanee Nagarajan, Varun Ramesh

March 25, 2019

In this project we are performing near duplicate video retrieval/detection (NDVD) using a CNN-L based approach [3]. With several competing approaches available such as Multi-view hashing, Pattern-based approach, ACC, CNN-based models and deep metric learning, we decided to move forward with the CNN-based model as it is currently state-of-the-art for NDVD, based on the mAP (mean Average Precision) metric.

Our project can be broadly segregated into the following modules:

1. CNN Based feature extraction for key frames
2. Feature Aggregation (Vector based)
3. Generating Frame & Video level histograms
4. Video Querying (Similarity Module)
5. Modified Key frame extractors and impact on CNN-model
6. Low-resolution video results using CNN-model

1 Current Progress

The CNN-based model works on top of a pre-trained neural network (AlexNet [4]) which has previously provided state-of-the-art results for image processing. **We have completed implementation of modules 1-4 (CNN-model). We have also performed preliminary checks on the results using a small sample keyframe and query video dataset [1] using cosine similarity between video-level aggregations**. At the time of writing of this report, we are currently working on porting this algorithm for the CC_WEB_VIDEO dataset [2] which has been previously used by several top publications in this domain.

As stated above, most of our focus in the first half of our project has been geared towards implementation of the CNN model. Our novel improvements that include modifying keyframe generations and its impact on the CNN model is the next module we will be working on. Bounded by

time constraints, we will also generate keyframes for low-resolution videos and run them through the developed CNN model to understand results and interpret specific requirements in case of low-resolution videos.

The rest of the report is organized as follows. In section 2, we have described a high-level overview of the CNN-based approach. In section 3, each of the modules of the CNN-model is described in further detail. Challenges that we faced during this implementation, preliminary results and updates to our schedule have been provided in section 4.

2 Overall Idea for NDVD

The CNN-L based approach is a well defined model in the space of NDVD algorithms where features are extracted from intermediate CNN layers. The following is a pictorial representation of the CNN-model architecture -

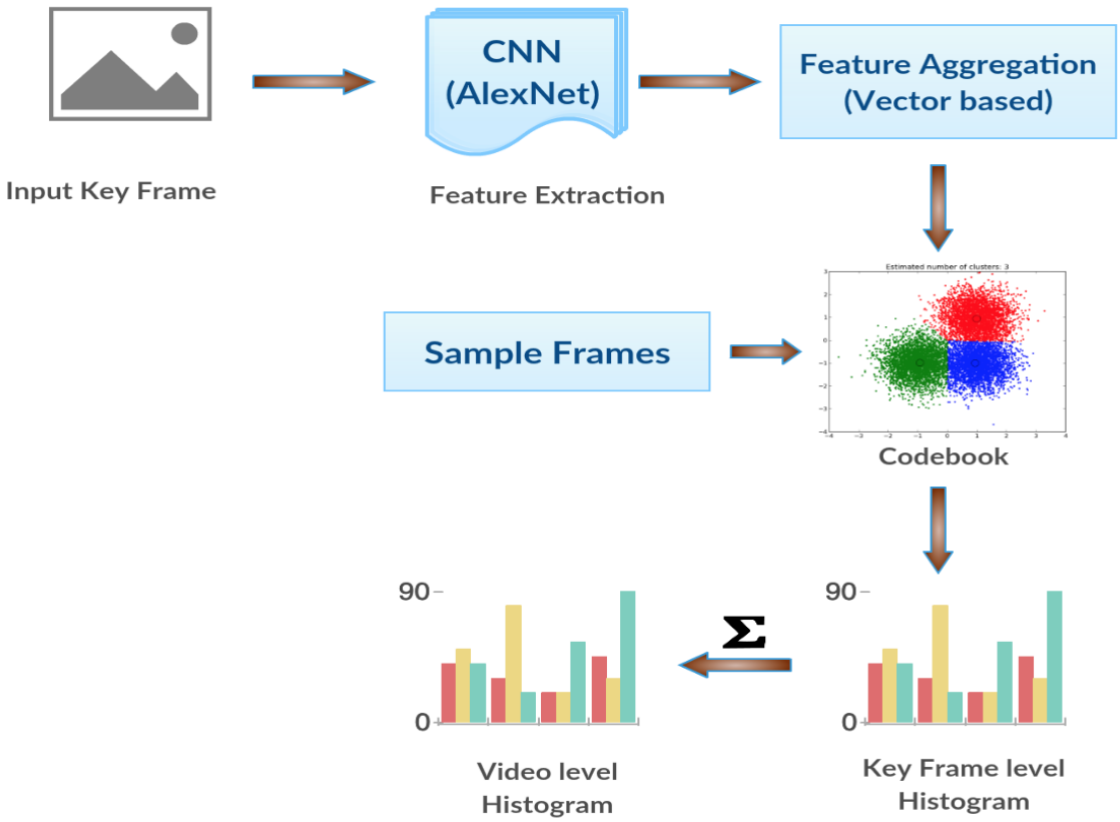

Figure 1: CNN-L model architecture

A bag-of-words representation or visual-word representation is generated using a vector-based aggregation method. As done in recent works in the NDVD space, a pre-trained CNN model has been adopted to extract these features from intermediate CNN layers. The main objective is to push an image through the CNN network and use a max pool function on each of the layers to extract features. Next, these feature vectors are concatenated to generate a video-level representation which is then followed by aggregation and query similarity module.

3 NDVD modules

3.1 CNN (pre-trained using AlexNet)

In this step, we extract the features from the input key frame using CNNs. The feature vector is obtained by the forward propagation of the input image through the CNN and applying max-pooling on every convolutional layer. In our implementation, we used AlexNet deep network architecture [4]. AlexNet consists of 5 convolutional layers and 3 fully connected layers. Relu function is applied after every convolutional and fully connected layer. Dropout is applied before the first and the second fully connected year.

3.2 Keyframe to Video Level Histograms

A vector based aggregation scheme is used to generate a key frame level histogram. A bag-of-words scheme is applied on the vector V_c resulting from the concatenation of individual layer features to generate a codebook of K visual words using clustering. After generating the visual codebook, every video keyframe is assigned to the nearest visual word. The final video level histogram (bag-of-words representation) H_v is derived by summing the histogram vectors corresponding to its keyframes.

3.3 Video Querying (Similarity Module)

Cosine Similarity and tf-idf weighting is used to calculate the similarity between two video histograms. The tf-idf weights are computed for every visual word in every video in the collection. Weight computation takes place online as it is not recalculated for every new query. The feature extraction and aggregation steps for a query video q are the same as the ones described above. Once the video-level histogram H_v is generated, the above mentioned inverted index structure is used to efficiently get videos that have atleast one common visual word. The videos are then sorted in descending order of rank based on their cosine similarity with the query video.

4 Challenges, Results, and Next Steps

4.1 Challenges and Results

The initial challenges faced during this phase was interpretation of Alexnet's architecture and understanding its context in the NDVD space. The CNN-L model develops the overall feature vector

by manipulating results obtained from intermediate layers and hence, the initial implementation and feature vector interpretation was a challenging but great experience.

Subsequently, implementation of this deepnet on a large dataset without appropriate GPU hardware was an initial hiccup. We worked around this by deploying the deepnet on a smaller keyframe and query dataset that helped to both understand and fine-tune the implementation. In terms of results, we observed a very high mAP (mean Average Precision) for the smaller keyframe and query dataset [1]. Given the skew, it is natural for a model to overfit the given data and hence, a more accurate and precise measurement of this metric will be presented in subsequent reports.

4.2 Using CNN-L model for CC_WEB_VIDEO dataset

Our immediate next step is to run our model for the much larger and established CC_WEB_VIDEO dataset [2]. The current model has been run on part of a sample dataset [1] which consists of labeled traffic videos and their keyframes.

4.3 Keyframe generator algorithm and impact on CNN model

Upon completion of the previous step, our next objective is to adapt an existing video summarization technique for keyframe generation from videos. We plan to use the keyframes generated using this technique as inputs for our above CNN-L model. We will explore the impact of using this technique instead of using the dataset's keyframes generated by shot boundary detection.

4.4 Low-resolution video accuracy using CNN model

Finally, we plan to run our model for a low resolution video dataset. Our aim is to understand the results and interpret specific requirements in case of low-resolution videos. This part of the project is subject to time constraints.

4.5 Updated Schedule

1. *Base Model Implementation - Modules 1 - 4 -* ***Completed March 20th***
   In this phase, we implemented current state-of-the-art (CNN-L) for Near-Duplicate Video Retrieval using shot boundary detection method as the key frame extractor. The model was trained on a sample mock dataset to alleviate large training makespan and fine-tuning issues. This phase is completed and preliminary results project high mAP.

2. *Training model on CC_WEB_VIDEO dataset -* ***Planned Completion - April 5th***
   In this phase, we will be porting the aforementioned model onto the much larger CC_WEB_VIDEO dataset which has been used in several other works in the NDVD space. This will help us benchmark our implementation of CNN-L against existing state-of-the-art before we delve into the keyframe modification phase.

3. *Modified Key Frame Extractor and impact on CNN model -* ***Planned Completion - April 17th***
   In this module, we will be exploring state-of-the-art keyframe extraction methods and video

summarization methods to replace the existing keyframe dataset. After this replacement, we will be inspecting the metrics of our CNN-L model for NDVD and contrast the two methodologies.

4. *Low Resolution Videos NDVD using this CNN model* - ***Stretch Goal - April 26th***
   Upon completion of the previous three models we will have a trained CNN-L model for NDVD with multiple keyframe extractors. In this stretch goal module we will attempt to perform NDVD for low-quality video scenarios using the output from first three modules.

References

[1] Ua-detrac benchmark suite. http://detrac-db.rit.albany.edu/download.

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

[3] 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.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.