ASSIGNMENT 2 – IR Group 046

Video Retrieval System Based on Deep Learning

PAPER - 1

Topic: Video retrieval based on deep convolutional neural network

Authors: Yajiao Dong and Jianguo Li

Publish year: April 2018

Conference: ICMSSP '18: Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing

Problem addressed

Conventional hashing algorithms learn binary hashing codes whose distance is correlated to the similarity relationship of the original input data but these cannot represent the input data accurately. These traditional methods all employ hand-crafted features to compare to what was originally there, which is not effective due to the differences between the high-level semantic similarity that humans can observe and the low-level visual similarity that machines can learn.

Contributions

This paper proposed a deep neural network based hashing mechanism through which we can generate binary codes which can be stored and used when we search for any videos on the platform.

The Architecture of the system having three parts:

1) Features extraction layers using Convolution neural network

2) FC layer followed by sigmoid layer to learn similarity preserving binary codes and second FC having k nodes which can also the number of categories

3) Combination of triplets and classification loss.

They are proposing a new variant of triplet loss function which contains relative similarity.

$l1 = \max(\|F(X) - F(X+)\|2\ 2 - \|F(X) - F(X-)\|2\ 2 + m, 0)$

(X,X+,X-) : X is more similar to X+ than X-

They have one more loss function which is classification loss function and final loss function as combination of these two with some weights.

Datasets: To evaluate the performance they used UCF101 and HMDB51 datasets.

UCF101 contains 13320 videos divided into 101 categories and HMDB51 has 7000 videos divided into 51 categories.

To compare results, they have compared their model performance with LSH, ITQ, SGH, Spectral hashing (SH), AGH, and Deep Hashing (DH), PCA-RR and SELVE hashing mechanism.

Test Results and Conclusion

Performance comparison of different video retrieval

algorithms on the UCF101 dataset. This table shows the mean Average Precision (mAP) of top10.

method	512bits	256bits	128bits	64bits
AGH	0.449	0.491	0.515	0.495
PCA-RR	0.724	0.717	0.689	0.65
LSH	0.736	0.71	0.671	0.605
SH	0.616	0.641	0.644	0.619
ITQ	0.757	0.75	0.735	0.701
SGH	0.697	0.685	0.491	0.323
SELVE	0.683	0.665	0.683	0.66
DH	0.79	0.778	0.759	0.723
ours	0.8	0.796	0.783	0.747

They used Deep convolutional neural networks to extract high-level semantic features of input videos and map real-valued representations into binary hash codes to simplify complexity.

They Proposed two loss functions which can help to optimize the NN through minimization.

PAPER - 2

Topic: Video Retrieval System Based on Deep Learning

Title: Semantic concept-based video retrieval using convolutional neural network

Authors: Nitin Janwe , Kishor Bhoyar

Publication details: Received: 16 November 2019 / Accepted: 7 December 2019 / Published online: 14 December 2019

Problem Addressed:

Due to the advancement in technology, there was an exponential growth in video collection on the internet. Finding the required videos among them will be quite challenging as we can't directly break into labels/tokens and can't classify based on rank directly. Among those, classifying multi-concept video was very difficult.

Contributions:

The Architecture divide the framework into two modules. Training and testing modules, In the training module we implement a classifier for key-frames dataset and in the testing module we will get the frames/scenes for relevant key-frame pairs. Here we use CNN's with Asymmetric training for classification of videos and FDCCM(Foreground Driven Concept Co Occurrence Matrix) matrix for score refinement and RIF(Ranked Intersection Filtering) for short listing common concepts. Add the common concepts with key-frame pairs in the database.

Test results and conclusion:

Precision will be calculated based on the key-frame Hits and Miss.

Precision = Hit/(Hit+Miss)

key-frame performance on sample test:

S. no.	Key-frame/s Id	Key concept	H	F	Precision (P)
1.	9199	Car and person	5	1	0.83 (83%)
2.	5311, 5312	Police_Security, Person, Face, Crowd, Outdoor	5	1	0.83 (83%)

Results retrieved with proposed Algorithm:

Key-frame	Ground-truth concepts	Detected concepts	Ranked Detected Concepts	Retrieved Key-frames for shots					
	4,6,21,23,32,34	21,32,4,27,6	21,32,4,27,6						
9199	6	4		7437	7645	8010	8092	8106	8114
	Key Fr-1: 17,21,23,24,33	6,26,32,31,24	24						
5311	5								
	Key Fr-2 6,17,21,23,33,34	23,24,10		10526	10531	10536	10537	10562	10566
5312	6								

Concepts:
4: Building
6: Car
10: Crowd
17: Military
21: Outdoor
23: Person
24: Police_Security
26: Road
27: Sky
32: Urban
33: Vegetation
34: Walking_Running

Comparison with existing methods:

Method	Dataset used	Precision (MAP)
Proposed method	TRECVID 07	0.544
Statastical_Active_Learning [30]	TRECVID 07	0.235
CRMActive [31]	TRECVID 07	0.260
pLSA [32]	TRECVID 07	0.390

Conclusion:

Based on the above results the proposed algorithm for Multi semantic concept-based video indexing was working better than other existing methods. And we can expand this work further using Fast R-CNN.

PAPER – 3

- **Title Of The Paper**

  Deep Learning Based Semantic Video Indexing and Retrieval

- **Authors**

  Anna Podlesnaya and Sergey Podlesnyy

- **Publications Details**

  September 21-22 2016(Intellisys)

- **Problem Addressed**

  1. **Huge Archives-**

     Archives are increasing on a daily basis . For example - the Russian Archives has 250k items(tv) and 100k items(movies) .

     Youtube - unlimited videos are uploaded by various users.

  2. **Modern Requirements-**

     Movies , documentaries require a lot of analysis so searching is necessary.

  3. **Speech Recognition**

     Speech Recognition limitation has been addressed.

     Querying by media (either by sample image or by sample video clip) is not possible using text-based indexing. Spatial querying would be very much limited as well. One needs to index video by visual content in addition to speech content .

- **Contribution**

  1. **Video Indexing**

     a. **Feature Extraction And Film Segmenting**

        GoogLeNet network structure has been used as the primary source of semantic feature extraction. It has been established by this paper that one time operation of CNN calculation per frame is enough to build a powerful video indexing and retrieval system.

     b. **Graph Oriented Indexing**

        Single CNN is not the correct way to represent a frame in video processing .Each frame can be classified with a different classifier.

        If we apply all these classifiers ,we can obtain other tags for the frame .

        So it is good to present a film as a graph.

2. **Video Retrieval**

   a. **Searching By Structured Queries**

      Basic keywords-based search in our graph index can be implemented with Cypher statement (3). It accounts for minimum confidence level of shot tags, and sorts the search results by shot duration descending.

   b. **Searching By Sample Videos**

      Video retrieval by sample clip is important in content production (finding footage in archives) and in duplicates finding (for legal purposes and for archives deduplication). In our setting the sample video is limited to a single shot discussed above, and the goal is to find semantically close shots. This differs from many existing solutions based on e.g. HSV histograms or SIFT/SURF descriptors.

      We found that feature vector fv $\mathcal{R}$ 1024 extracted in Algorithm 1 contains enough semantic information for retrieving video shots having similar content with the sample clip. A brute force solution involves comparing distance between sample clip feature vector and every other shot's feature vector with some threshold, and including the shots having smaller distance to the sample into the search results. We compared Euclidean distance and cosine distance metrics of vector distance and selected the cosine distance as preferred one (6). ddot(x, y) Where x - sample clip feature vector, y - other clip feature vector.

   c. **Searching By Sample Images**

      In order to extend possibilities for video retrieval beyond the scope of pre-set nomenclature of categories we explored on-line training of linear classifiers over feature vectors extracted by CNN

- **Conclusion**

  We showed in this work that feature vector fv $\mathcal{R}$ 1024 extracted by CNN [6] contains enough semantic information for segmenting'

  - raw video into shots with 0.92 precision.
  - retrieving video shots by keywords with 0.84 precision;
  - retrieving videos by sample video clip with 0.86 precision
  - retrieving videos by online learning with 0.64 precision.

  All that is needed for indexing is a single pass of feature vector extraction and storing into the database. This is the only time when expensive GPU-enabled hardware is needed. All video retrieval operations may run in commodity servers e.g. in cloud-based settings. However more efforts are necessary to increase the performance of samples-based video retrieval. While lexical pruning of search space helps to limit the scope for brute force algorithm it scales linearly with the data amount. We plan to explore several approaches for lowering the feature vector dimensionality in order to search in log time scale, e.g. random projections and compact binary descriptors.