Unsupervised Teacher-Student Model for Large-scale Video Retrieval

Dong Liang[1,2,*] , Lanfen Lin[1], Rui Wang[2], Jie Shao[2], Changhu Wang[2], Yen-Wei Chen[3]
[1]Collage of Computer Science and Technology, Zhejiang University, China
[2]AI Lab, ByteDance Inc., China
[3]College of Information Science and Engineering, Ritsumeikan University, Japan

Abstract

With the growth of video-sharing platforms and social media applications, video retrieval plays an import role in many aspects, such as copyright infringement detection, event classification, personalized recommendation, and etc. The content-based video retrieval presents the following two main challenges: (i) Distribution inconsistency for feature representation from the source domain to the target domain. (ii) Difficulty of video aggregation by sufficiently incorporating frame-based information. In this paper, we propose an unsupervised teacher-student model (UTS Net) to improve the performance of the content-based video retrieval tasks: (i) A teacher-student model maintaining the global consistency for feature representation from different domains and retaining the local inconsistency within the intra-batch data; (ii) A simple but effective video retrieval pipeline integrating the frame-level binarized feature. Our proposed framework experimentally outperforms the state-of-the-art approach on the DSVR, CSVR, and ISVR tasks in the FIVR datasets, and achieves a mean average precision of 76%, 72%, and 61%, respectively.

1. Introduction

With the explosive growth of online video sharing and consumption, content-based video retrieval technique is desired in a wide range of internet applications areas, such as incident classification, copyright infringement detection, personalized recommendation etc. Video retrieval concern the fact of indexing similar video scenes to a given video query, and most studies focus on design of the image feature representation and the video feature aggregation.

In the recent works on content-based image retrieval, the feature extracted from the activations of pre-trained CNN network is used as off-the-shelf image representation [1-5] and surpasses the conventional hand-crafted features such as SIFT [6]. Maximum activation of convolution (MAC) or region maximum activation of convolution (R-MAC) [1] are well-exhibited unsupervised image representation from the activations of intermediate convolution layers via global or regional pooling layer. To generate more discriminative image representations, a supervised framework of deep metric learning (DML) [7] is conducted via the pairwise or the triplet-wise constraints, aiming at maximizing similarity between the relevant contents and minimizing the similarity between the irrelevant contents. Nevertheless, few methods attempt to minimize the distribution inconsistency between the feature representations obtained from the source domain (pre-trained dataset) and target domain (unlabeled dataset).

On the other hand, video aggregation integrating frame-level feature plays an import role in the content-based video retrieval. A commonly used aggregation technique is to generate a global video descriptor by averaging the frame-level features while ignoring the frame-level differences [8]. To compact more visual information of video shots into the final aggregated feature, the bag of words (BoW) technique incorporating the *tf-idf* weighting is employed in [9-11]. The drawback is that feature representation of visual words is prone to be influenced by vocabulary size and may not preserve some relevant contents [12].

These methods mentioned above present two main challenges: (i) Distribution inconsistency: the distribution gap for feature representation exists between the source domain (e.g., pre-trained dataset) and the target domain (e.g. unlabeled dataset). How to map feature representation into a shared domain space retraining the local inconsistency? (ii) Video feature aggregation: how to design an effective video retrieval pipeline sufficiently incorporating the frame-level information? The two inter-related challenges mentioned above are addressed by the following proposed contributions:

(1) Teacher-student model

An unsupervised teacher-student model is proposed to tackle the problem of distribution inconsistency for feature representation from the different domains. Specifically, the global consistency is maintained by minimizing the distance between feature spaces in respect to the teacher model and the student model; the local inconsistency within the intra-batch data is retained by maximizing the distance between the dissimilar pair in the same batch.

(2) Retrieval pipeline

A frame-level retrieval pipeline on the basis of the hash binarized feature is developed to perform video retrieval, which effectively preserves the critical contents of the key-frames without increasing extra burden of computation.

The evaluation of our proposed method achieves the best performance on duplicate scene video retrieval (DSVR), complementary scene video retrieval (CSVR), and incident

[*] Work done while Dong Liang was an intern at ByteDance AI Lab.

scene video retrieval (ISVR) tasks in respect to the FIVR dataset [9]. The rest of paper is organized as follows. Section 2 describes the proposed methodology in detail. Section 3 shows the implementation detail, the experiment results as well as a brief discussion. Section 4 draws a conclusion on our proposed retrieval framework.

2. Methods

In this section, the motivation perspective within the proposed unsupervised teacher-student model is firstly discussed. Then the framework is illustrated with two key steps, namely the teacher-student model training and the retrieval pipeline.

2.1. Motivation

Though the off-the-shelf features proposed in [1-5] exceeds the performance of the conventional geometric feature on the retrieval, the further promotion of feature representation is inhibited by the distribution gap existing between the target domain and the source domain. For a video-retrieval task with large scale gallery, we can explicitly make the following hypotheses that the number of hit videos meeting the condition of content similarity for a specific query is small. The proposed hypotheses assert that frames in a batch present high probability of contents irrelevance. Under the above assumption, the distance between the irrelevant pair within frames in a batch should be maximized for a trained model (called the student model). The optimization of the student model is easily trapped into a local optimum without a teaching model maintaining the global consistency. Thus, a teacher-student model training strategery is proposed to maintain global consistency for data from different domains and retain intra-inconsistency within the same batch.

2.2. Teacher-Student Model

Fig. 1 presents the block diagram of the teacher-student model. The student model should be designed highly efficient to fulfil the requirement of real-time feature extraction, whereas the teacher model is not limited to the same backbone as the student model. Notably, the teacher model needs no training to avoid global consistency changing between the inter-batch data. Considering the ease of measuring correlation between the student model and the teacher model, a fully-connected (fc) layer realized PCA-module is added to the backbone of the teacher model for the feature dimension consistency. On the other hand, given that the PCA-module in the teacher model achieves a zero-mean output, the last block of VGG [13] (student model) is replaced by a Conv-BN-ReLU×2 + Conv-BN [14] module to scale the outputs to zero mean.

A fc layer is adopted to implement the PCA in the teacher model. Mathematically, the $\boldsymbol{f}$ and $\boldsymbol{f}^*$ are used to denote the global feature vector of the official Resnet50 [15] and the designed teacher-model, respectively. The parameters in respect to the fc layer are computed as follows:

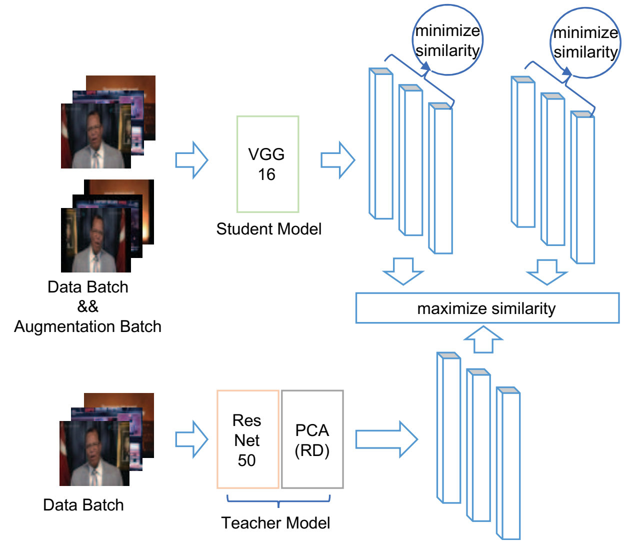

Fig. 1. The architecture of proposed unsupervised teacher-student model.

$$\boldsymbol{\Sigma} = \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{f}^{(i)} - \overline{\boldsymbol{f}})(\boldsymbol{f}^{(i)} - \overline{\boldsymbol{f}})^T \tag{1}$$

$$\boldsymbol{U}, \boldsymbol{V} = eig(\boldsymbol{\Sigma}) \tag{2}$$

$$\boldsymbol{f}^* = \frac{(\boldsymbol{U}[:,:D^*])^T \otimes (\boldsymbol{f} - \overline{\boldsymbol{f}})}{\sqrt{\boldsymbol{V}[:D^*]}} = \frac{(\boldsymbol{U}[:,:D^*])^T}{\sqrt{\boldsymbol{V}[:D^*]}}\boldsymbol{f} - \frac{(\boldsymbol{U}[:,:D^*])^T}{\sqrt{\boldsymbol{V}[:D^*]}}\overline{\boldsymbol{f}} \tag{3}$$

where m and $\overline{\boldsymbol{f}}$ indicate the number of frames and the average vector of frame feature representations. The eigenvectors $\boldsymbol{U} \in \mathbb{R}^{D \times D}$ and the eigenvalue $\boldsymbol{V} \in \mathbb{R}^{D}$ are computed via the function (2), respectively, where D and D^* represent the dimension in respect to $\boldsymbol{f}$ and $\boldsymbol{f}^*$.

2.3. Training Strategy

The concept of the inter-batch can be understood as a batch input of a pseudo-siamese network with two subnetworks: the student network and the teacher network which are mentioned above. To maintain the global consistency between the feature representations acquired from the teacher-model and the student-model, the optimized function is formulated as follows:

$$\mathcal{L}(\{x_i\}, \{y_i\}) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, y_i) \tag{4}$$

$$d(x_i, y_i) = norm(x_i) * norm(y_i) \tag{5}$$

where n is batch size, and $\{x_i\}, \{y_i\}$ denotes the feature representations of the teacher model and the student model, respectively. The distance between x_i and y_j is measured by cosine similarity as shown in equation (5), and *norm* denotes *l2_norm*.

On the other hand, the intra-batch inconsistency is retained via minimizing the similarity within the batch frames irrelevant to each other. The optimized function is described as follows:

$$\mathcal{L}(\{x_i\}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(\beta-\alpha)*n}\sum_{j=\alpha*n}^{\beta*n}\mathcal{L}_i^s[j] \tag{6}$$

where $\mathcal{L}(\{x_i\})$ represent the average similarities between the *i*-th image and some other images in the intra-batch, α and β are two hyper-parameters indicating the fraction of similar images used to calculate $\mathcal{L}(\{x_i\})$ in a batch. $\mathcal{L}_i^s$ is a sorted list of cosine similarity measured on pairwise images sampled from the same intra-batch, as is defined as follows.

$$\mathcal{L}_i^s = sorted(\{d(x_i, x_j)|\forall j \in \{1,2\ldots,n\}\}) \tag{7}$$

Finally, joint optimization for the ***global consistency*** and ***local inconsistency*** is performed via minimizing the loss function given in (8). The former attempts to maintain the inter-batch consistency by maximizing similarity between the outputs of the student model and the teacher model; the latter minimize correlation between irrelevant frame contents among the batch with and without augmentation.

$$\mathcal{L} = \lambda_0(1-\mathcal{L}(\{s_i\},\{t_i\})) + \lambda_1(1-\mathcal{L}(\{s_i^*\},\{t_i\})) + \lambda_2\mathcal{L}(\{s_i\}) + \lambda_3\mathcal{L}(\{s_i^*\}) \tag{8}$$

where s_i and s_i^* are the feature representation of the student model obtained from data with and without augmentation; and t_i represents the output of teacher model from data without augmentation.

2.4. The frame-level retrieval pipeline

The commonly used video aggregation scheme is to generate a global vector (GV) by averaging the frame-level features. Though the video aggregation is efficient and achieve a reasonable result, the frame-level features are not considered fully to promote the retrieval performance. The frame-level retrieval based on real-value feature (i.e. the output of student model) often suffers from heavy computation and high memory storage. Hence, hash-based methods are encouraged to generate binary codes for video frames with less burden of computation. In this paper, the locality-sensitive hashing (LSH) algorithm [16] is employed to binarize the frame-level feature. As to a specific video query task, keyframes are extracted and frame-level retrieval is performed to construct video-level similarity. The complete frame-level retrieval pipeline is built via Algorithm 1.

Algorithm 1 Aggregation of the video similarity

```
Input: binary frame-level feature for a query video
Definition: qf_id, query frame id; gf_i, gallery frame id; gvid$^{(i)}$, vid for the i-th frame in gallery
Global Initialize: all-zero dict S to store the scores of the gallery videos.
For qf_id = 0; qf_id < qf_num; qf_id++ do
    Find a sorted list of topN (N = 1000) similar frames in the gallery: gf = {gf_1,gf_2,…,gf_N}
    Local Initialize: all-zero dict X
    For gf_i in {gf_1,gf_2,…,gf_N} do
        Find gvid$^{(i)}$,
        If X[gvid$^{(i)}$] == 0 do
            S[gvid$^{(i)}$] += frame-level similarity
            X[gvid$^{(i)}$] = 1
        End if
    End for
End for
The scores in S are divided by qf_num.
Output: the normalized score array S.
```

3. Experiments

3.1. Experimental setup

Dataset: FIVR [10] is a large-scale dataset collected for the fine-grained incident video retrieval problem, which consists 225,960 videos and 100 queries. According to the level of content association, three types of similar videos are classified, namely the duplicate scene video (DSV), the complementary scene video (CSV), and the incident scene video (ISV). The duplicate scene video retrieval (DSVR) focus on DSV sharing at least one scene from a unique camera source; the complementary scene video retrieval (CSVR) processes CSV pair of two videos containing an overlapping segment on the same incident as well as DSV; the incident scene videos (ISV) in respect to the same incident and ignoring the time overlapping is processed by incident scene video Retrieval (ISVR). Thus, we can conclude the relationship between these three tasks as follows: DSVR ∈ CSVR ∈ ISVR.

Implementation details: the teacher-student model is trained on 8 Tesla-V100 32GB GPUs (batch size: 320) with adam optimizer [17] (learning rate: 0.0001). Online augmentation (such as rotation, random cropping, distorted color operation, and etc.) are utilized during the model training. Keyframes of all videos are extracted via FFMPEG command. Hyper-parameters: α = 0.1, β = 0.6, λ_0, λ_1, λ_2 and λ_3 are 2.0, 2.0, 1.0, 1.0, respectively. Faiss [18] in combination with the LSH technique are employed to train the hash mapping function and each of the frame feature is binarized to 512 bits.

3.2. Ablation study

For better understanding of the significance of the teacher-student model and the proposed pipeline, a series of ablations are conducted, and the results are summarized in Table 1. The UTS V2 model employs the ResNet50 and VGG16 as the backbone of the teacher model and the student model, respectively. The results indicate a fact that the teacher-student model achieves a better ability of feature representation compared with the official VGG16.

Table 1 Ablation results, mAP(%).

	Method	DSVR	CSVR	ISVR
Global Vector (GV)	VGG 16	28.80	27.31	22.35
	ResNet 50	32.65	31.64	26.88
	UTS V1	28.51	27.74	23.26
	UTS V2	**41.45**	**40.78**	**35.36**
RMAC + GV	VGG 16	44.48	42.50	35.47
	ResNet 50	**51.82**	**50.74**	**43.80**
	UTS V1	48.74	47.31	40.24
	UTS V2	50.94	49.83	43.18
RMAC + Frame-level retrieval pipeline (FRP)	VGG 16	68.32	63.43	52.09
	ResNet 50	76.83	**72.47**	**61.27**
	UTS V1	74.98	70.34	58.63
	UTS V2	**76.86**	72.39	61.07

Furthermore, both of the student model and the teacher model in UTS V1 use VGG16 as the backbone, and the results indicates the student model surpass the teacher model via minimizing the intra-batch similarity. Besides, the frame-level retrieval pipeline (FRP) proposed in this work brings significant promotion of the retrieval performance, as shown in the last cell in Table 1.

3.3. Comparison with the state-of-the-art.

Our proposed method is compared with the state-of-the-art methods, and the results are given in Table 2. For commonly used global average aggregation scheme, the performance of our proposed model is superior to those of other methods. Besides, the proposed frame-level video retrieval pipeline achieves the state-of-the-art on the DSVR, CSVR, and ISVR tasks.

3.4. Discussion

Computation cost: The computation cost of methods mentioned above are investigated and the results is shown in Table 3. Both the GV-based and FRP-based methods cost approximately 2KB memory to store the video representation. The same brute-force search engine is utilized to compare the index time of GV, LBoW, and our proposed FRP. With the FPR approach, the retrieval time is controlled in about 210ms which is one-third index time of the LBoW method. Thus, the proposed FRP is time and memory efficient to meet large scale video retrieval tasks.

Table 2 mAP (%) of the different feature extractors and feature aggregation schemes for three tasks.

Method	GV	LBoW[10]	FRP
HSV*	16.5	N/A	-
VLCD*	29.4	N/A	-
VGG*	36.6	71.0	-
RES*	35.0	59.6	-
C3D_fc*	24.4	N/A	-
I3D_fc*	32.1	N/A	-
R-VGG	44.48	-	68.32
R-ResNet	51.82	-	**76.83**
R-UTS-V1	48.74	-	74.98
R-UTS-V2	50.94	**-**	**76.86**

Method	GV	LBoW[10]	FRP
HSV*	15.3	N/A	-
VLCD*	27.5	N/A	-
VGG*	34.7	67.5	-
RES*	33.3	57.2	-
C3D_fc*	23.1	N/A	-
I3D_fc*	31.2	N/A	-
R-VGG	42.50	-	63.43
R-ResNet	50.74	-	**72.47**
R-UTS-V1	47.31	-	70.34
R-UTS-V2	49.83	**-**	**72.39**

Method	GV	LBoW[10]	FRP
HSV*	11.8	N/A	-
VLCD*	21.4	N/A	-
VGG*	28.1	57.2	-
RES*	27.4	48.8	-
C3D_fc*	17.6	N/A	-
I3D_fc*	25.3	N/A	-
R-VGG	35.47	-	52.09
R-ResNet	43.80	-	**61.27**
R-UTS-V1	40.24	-	58.63
R-UTS-V2	43.18	**-**	**61.07**

Notes: * represents the results reported in [9]

Table 3 Storage and computation per video

Method	**GV**	**LBoW[10]**	**FRP**
Storage (B)	2048	3050	2075
Index time (ms)	301.8	774.9	210.3

Difficulty and challenge: The challenge and difficulty in our video retrieval task can be summarized as the local patch retrieval problem. Although RMAC provides a compact image representation incorporating with multiple region-level information, regional importance is not considered and the impact of local patch with crucial information is attenuated. For instance, the query (vid=wrC_Uqk3juY[1]) failed to hit its candidates (vid: wmFQfvg1OYA; vid: uimXpGbHYPQ) in the gallery with a relative low-ranking index of 112 and 3331. Local descriptor and attention mechanism [19-22] may provide an opportunity to alleviate the patch retrieval problem.

4. Conclusion

In this paper, we proposed an unsupervised teacher-student model and a frame-level retrieval pipeline for large-scale video retrieval. The aim was to find an approach to acquire discriminative feature representation and afford a pipeline to unleash the potential of the frame-level feature in the video retrieval task. We establish experimentally that our proposed method achieves the state-of-the-art performance in large scale video retrieval task.

[1] You can visit the video by: https://www.youtube.com/watch?v=vid

References

[1] Tolias G, Sicre R, Jégou H. Particular object retrieval with integral max-pooling of CNN activations[J]. arXiv preprint arXiv:1511.05879, 2015.

[2] Radenović F, Tolias G, Chum O. Fine-tuning CNN image retrieval with no human annotation[J]. IEEE transactions on pattern analysis and machine intelligence, 41(7): 1655-1668, 2018.

[3] Li Y, Xu Y, Wang J, et al. Ms-rmac: Multiscale regional maximum activation of convolutions for image retrieval [J]. IEEE Signal Processing Letters, 24(5): 609-613, 2017.

[4] Babenko A, Lempitsky V. Aggregating local deep features for image retrieval[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1269-1277.

[5] Wu X, Irie G, Hiramatsu K, et al. Weighted generalized mean pooling for deep image retrieval[C]//2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018: 495-499.

[6] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 60(2): 91-110, 2004.

[7] Kordopatis-Zilos G, Papadopoulos S, Patras I, et al. Near-duplicate video retrieval with deep metric learning [C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 347-356.

[8] Wu X, Hauptmann A G, Ngo C W. Practical elimination of near-duplicates from web video search [C]//Proceedings of the 15th ACM international conference on Multimedia. ACM, 2007: 218-227.

[9] Jiang Y G, Ngo C W, Yang J. Towards optimal bag-of-features for object categorization and semantic video retrieval[C]//Proceedings of the 6th ACM international conference on Image and video retrieval. ACM, 2007: 494-501.

[10] Kordopatis-Zilos G, Papadopoulos S, Patras I, et al. Near-duplicate video retrieval by aggregating intermediate cnn layers[C]//International conference on multimedia modeling. Springer, Cham, 2017: 251-263.

[11] Zhao W L, Ngo C W. Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection[J]. IEEE Transactions on Image Processing, 2009, 18(2): 412-423.

[12] Kordopatis-Zilos G, Papadopoulos S, Patras I, et al. FIVR: Fine-grained Incident Video Retrieval[J]. IEEE Transactions on Multimedia, 2019.

[13] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[14] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [J]. arXiv preprint arXiv:1502.03167, 2015.

[15] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[16] Datar M, Immorlica N, Indyk P, et al. Locality-sensitive hashing scheme based on p-stable distributions [C] // Proceedings of the twentieth annual symposium on Computational geometry. ACM, 2004: 253-262.

[17] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[18] Johnson J, Douze M, Jégou H. Billion-scale similarity search with GPUs[J]. IEEE Transactions on Big Data, 2019.

[19] Siméoni O, Avrithis Y, Chum O. Local Features and Visual Words Emerge in Activations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 11651-11660.

[20] Mukundan A, Tolias G, Chum O. Explicit Spatial Encoding for Deep Local Descriptors[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 9394-9403.

[21] Li W, Zhu X, Gong S. Harmonious attention network for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2285-2294.

[22] Dai Z, Chen M, Zhu S, et al. Batch feature erasing for person re-identification and beyond[J]. arXiv preprint arXiv:1811.07130, 2018.