Unsupervised Learning of Visual Representations by Solving Shuffled Long Video-Frames Temporal Order Prediction

Fatemeh Siar
fatemeh.siar@aut.ac.ir
Amirkabir University of Technology

Amin Gheibi
amin.gheibi@aut.ac.ir
Amirkabir University of Technology

Ali Mohades
mohades@aut.ac.ir
Amirkabir University of Technology

Figure 1: Left: Shuffled frames of a video. Right: Successfully ordered frames.

ABSTRACT

There is lots of hidden information behind the sequential data and their sequences. We proposed a model for learning visual representation by solving order prediction task. We concatenated the frame pairs, instead of concatenating the feature pairs. This concatenation makes it possible to apply a 3D-CNN to extract features from the frame pairs. Also, we proposed a new grouping, which have achieved 80 percent accuracy on average. We have modified the shuffled video clips order prediction task to the shuffled frame order prediction, by selecting a frame from each clip, by random. Then this task was solved by applying our model.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Computer vision**.

ACM Reference Format:
Fatemeh Siar, Amin Gheibi, and Ali Mohades. 2020. Unsupervised Learning of Visual Representations by Solving Shuffled Long Video-Frames Temporal Order Prediction. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters (SIGGRAPH '20 Posters), August 17, 2020.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3388770.3407409

1 INTRODUCTION

Recently CNN has widely used in learning visual representations. In some cases, a model trained to predict the temporal order [Lee et al. 2017; Misra et al. 2016; Santa Cruz et al. 2017; Xu et al. 2019]. In [Misra et al. 2016], the authors applied simple concatenation. They extracted features from the video frames and concatenated all extracted features to detect whether the video is shuffled. In [Lee et al. 2017], a 2D-CNN, such as AlexNet or its modified version CaffeNet, is applied to extract features from the frames. In [Xu et al. 2019] a 3D-CNN is used to extract features from the shuffled video clips. Both these papers [Lee et al. 2017; Xu et al. 2019], conducted pairwise feature concatenation. Then by comparing feature pairs, acquired the temporal order. We proposed a model for a shuffled video frames order prediction task. A sample result is shown in Fig.1. Also we applied our model for the shuffled video clip temporal order prediction task, which yielded significant results.

2 OUR APPROACH

In 2017 [Lee et al. 2017] have proposed the pairwise feature comparison model for sequence sorting tasks (SS). It later was used by [Xu et al. 2019] for video clip order prediction (VCOP). Requirements for high computing resources are a VCOP [Xu et al. 2019] disadvantage. A pre-processing is a requirement of SS to prevent the network from learning low-value information [Lee et al. 2017]. In the test set, where the frames are shuffled, extracting important parts of the frames and any pre-processing that uses the actual order could results big differences between the test data and training set. So we have decided to eliminate this phase by using 3D-CNN. In all experiments recurrent 3D convolutional neural networks (R3D-CNN) is used. The other contribution of the proposed model is in its grouping. This change has a huge impact on our results. Two different networks trained, using all classes (no-grouping(–)) and the grouping used by SS [Lee et al. 2017]. The accuracy went from 40 to 60 percent, shown in Table.1.

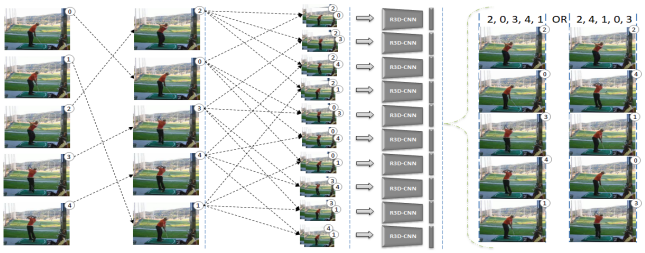
By considering the pairwise comparison and existence of the shuffled frames, we have proposed a new grouping algorithm (Algorithm.1) instead of commonly used horizontal flipping for images [Lee et al. 2017]. We considered the coherency between the frames pairs instead of the whole clip. The proposed grouping brought the results close to 80 percent. We reviewed the grouping for AlexNet. Using a 2D-CNN like AlexNet requires a pre-processing. Without a pre-processing, the accuracy for both SS and our grouping are under 40 percent. It is almost a random result for a 3-class task. Therefore, only the results of R3D-CNN are reported in Table.1.

2.1 Shuffled Video-Frame Order Prediction

R3D-CNN is used to extracting features from the original video frame pairs. As shown in Fig.2, all frames are concatenated in pairwise. They became a three-dimensional structure so a R3D-CNN can be applied. Then a softmax layer used to classify the extracted features. As explained earlier, we have changed the grouping. Table.1 shows the effect of the proposed grouping on the UCF101

SIGGRAPH '20 Posters, August 17, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7973-1/20/08.
https://doi.org/10.1145/3388770.3407409



(a). Data sampling (b). Concatenation (c). Feature extraction (d). Classification

Figure 2: (a). Frames are selected and shuffled by random. (b). Frame are resized and cropped. All two frame pairs are concatenated. (c). R3D-CNN is applied to extract features. (d). Classification to N!/2 groups by using softmax layer.

database [Soomro et al. 2012]. According to the comparison table, the proposed grouping achieved 88.1 percent accuracy, which is a significant improvement over 63 percent of the previous method SS [Lee et al. 2017]. Similar to [Xu et al. 2019], we are reporting the result per model with the lowest loss in the validation set during the training phase. The data set is divided into train and test set. Learning rate, momentum, dropout, and weight decay are 0.001, 0.9, 0.5, and 0.0005, respectively (same as VCOP [Xu et al. 2019]).

While an increase in length dramatically increases the number of classes, the Batch-Size (BS) decreased due to hardware resources. As can be seen in Table.2, the BS is much smaller than the number of classes in 7 tuple frames network. Also, the number of classes will increase exponentially while the training data is still constant and is the same. However, the obtained results are acceptable and significant in comparison with SS [Lee et al. 2017] and VCOP [Xu et al. 2019]. It provided on comparison table from Table.1. Our full results are provided in Table.2.

2.2 Shuffled Video-Clip Order Prediction

Our model can be applied for the shuffled video clips order prediction task. As shown in Fig.3, video frames randomly selected from the non-overlapped shuffled video clips, and then our model is applied to these shuffled frames. For a robust result, this experiment

Table 1: Merged grouping and comparison tables.

Grouping					
length	**Group**	**Class**	**CNN**	**Validation**	**Test**
3-tuple	–	6	R3D	49.7%	48.7%
3-tuple	SS	3	R3D	60.3%	58.5%
3-tuple	Ours	3	R3D	90.4%	88.1%
Comparison					
length	**method**		**CNN**	**Validation**	**Test**
3-tuple	Ours		R3D	90.4%	88.1%
3-tuple	SS		Alexnet	–	63%
4-tuple	Ours		R3D	84.4%	82.3%
4-tuple	SS		Alexnet	–	41%
3-clip	Ours		R3D	90.4%	80%
3-clip	VCOP		R3D	–	68.4%

Video clip sampling and shuffling Select a frame by random Input frames to model

Figure 3: Shuffled video clip sequencing: Frames extracted from the video clips by random. Then our model is applied.

conducted multiple times, and the class with the maximum repetition reported in Table.2. For details, consider three video-clips, with a length of 16 frames per each clip. There is a possibility of 16 * 16 * 16 for selecting three frames. We can eliminate frames that are slightly different in the clip and make a smarter choice, but the frames selection in Table.2 is random and blind. It should be noted that in the trained network similar to [Xu et al. 2019] at the time of sampling, the distance between the samples was 8 (which is constant). To improve the model's performance, we have trained the second model with 3 different sampling intervals of 8-16-32. Its results are shown in the "3-v2" row of Table.2. According to this table, this model with various intervals is better for this task.

Algorithm 1 Grouping

```
procedure CLASS GROUPING
    tuple_length ← number of frames; class ← current class
    if class >= (tuple_length)!/2 then
        class ← (tuple_length!) − class − 1
    return class    %class index start from 0
```

3 CONCLUSION

In this paper, we proposed a model for the shuffled video frame temporal order prediction task. We have applied the proposed model to the shuffled video clip order prediction. By applying the proposed grouping and frame concatenation together, we have achieved a significant improvement compared to the previous works.

REFERENCES

Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2017. Unsupervised representation learning by sorting sequences. *ICCV* (2017), 667–676.

Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. 2016. Shuffle and Learn: Unsupervised Learning Using Temporal Order Verification. *ECCV 2016* (2016), 527–544.

Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, and Stephen Gould. 2017. Deeppermnet: Visual permutation learning. *CVPR* (2017).

Khurram Soomro, Amir Roshan Zamir, and M Shah. 2012. A dataset of 101 human action classes from videos in the wild. *CRCV* (2012).

Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. 2019. Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction. *CVPR* (2019).

Table 2: Shuffled video frame/clip order prediction results.

Tuple of N frames					Video-clip per repetition				
Len	$\frac{N!}{2}$	BS	**Val**	**Test**	1	20	50	80	100
3-tpl	3	256	90.4%	88.1%	70.9%	78.4%	78.5%	78.7%	78.6%
3-v2	3	256	81.4%	88.4%	73.4%	80%	80.7%	80.5%	80.7%
4-tpl	12	130	84.4%	82.3%	55.6%	65%	66.8%	65.9%	65.6%
5-tpl	60	70	81.6%	77.3%	46.1%	57.8%	58.7%	59%	58.4%
6-tpl	360	48	86.8%	77.3%	41%	54.6%	56.6%	56.9%	56.8%
7-tpl	2050	32	56.7%	43.3%					