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