Unsupervised Feature Learning from Videos for Discovering and Recognizing Actions

Carolina Redondo-Cabrera
carolina.redondoc@edu.uah.es

Roberto J. López-Sastre
robertoj.lopez@uah.es

GRAM
Department of Signal Theory and Communications
University of Alcalá, Spain

Abstract

In this work, we evaluate different unsupervised feature learning approaches using Convolutional Neural Networks (CNNs) and videos. Essentially, we experiment with a CNN feature learning model based on the Contrastive Loss function. We propose a novel Siamese-tuple network architecture trained with a new loss function which benefits from the temporal coherence present in videos as the unique form of *free* supervision. Technically, our approach learns a feature representation where the temporal coherence in contiguous video frames is kept, while it encourages that the distance between frames extracted from the same video is smaller than the distance between frames of different videos. Experiments show the impact of our solution on two different tasks: unsupervised action discovery in videos, and action recognition in still images.

1 Introduction

What is a good visual representation? At the beginning of the century, most computer vision research focused on this "what" and used hand-crafted features as the underlying visual representation. However, the last years have seen the resurgence of learning visual representations directly from pixels themselves using CNNs [5]. These deep learning models designed for object or action recognition have millions of parameters, necessitating enormous manually annotated datasets for training. Therefor, a question still remains: Is a strong supervision necessary for learning a good visual representation?

In this work we explore unsupervised feature learning models for CNNs which simply use unlabeled video for training. We also propose a Siamese-tuple network architecture with a new loss function. Our loss benefits from the temporal coherence present in videos as the unique form of *free* supervision. Technically, it enforces that in the learned feature representation the temporal coherence of contiguous video frames is kept, while the distance between frames of different videos is always greater than the distance between frames of the same video. We validate our approach proposing a novel benchmark: the fully unsupervised discovery of actions in videos. Essentially, given a set of unlabeled videos, the goal is to separate the different classes. This idea is motivated by the work of Tuytelaars *et al.* [8] who discover object classes from collections of unlabeled images. We propose now to use the UCF101 video dataset [7] to discover the different actions. Moreover, we also explore whether our unsupervised feature learning solutions can even surpass a strongly supervised pipeline, with pre-training and fine-tuning, for the problem of action recognition using the PASCAL VOC 2012 dataset [1].

2 Unsupervised Learning from Videos

Our goal is to train CNNs using unlabeled videos. However, since we do not have labels, it is not clear what the loss function should be, and how we should optimize it. But videos come with a *free* form of supervision: temporal coherence. In other words, we all know that, generally, contiguous video frames do not drastically change. In [9] the authors follow this principle, and present the concept of *slow feature analysis* (SFA).

SFA encourages the following property: in a learned feature space, temporally nearby frames should lie close to each other. For a learned representation Ψ, and adjacent video frames V_t and V_{t+1}, one would like $\Psi(V_t) \approx \Psi(V_{t+1})$. Following [2], we go one step further and present a contrastive version of the loss function, which can also exploit negative (non-neighbor) pairs of frames. As it is shown in Equation 1, the contrastive loss $L(\Psi(V_t), \Psi(V_{t+1}), \Psi(V_{t+n}))$ penalizes the distance between pairs when they are neighbors ($Y = 1$), and encourages the distance between them when they are not ($Y = 0$).

$$\begin{split}\mathcal{L}(\Psi(V_t), \Psi(V_{t+1}), \Psi(V_{t+n})) = \sum_{t \in \mathcal{V}} Y \cdot d(\Psi(V_t), \Psi(V_{t+1})) + \\ (1-Y) \cdot \max\{0, \delta - d(\Psi(V_t), \Psi(V_{t+n}))\}.\end{split} \tag{1}$$

Inspired by SFA [9] and the Contrastive Loss [2], we first propose the following idea. A query frame and its adjacent video frame should lie close in the feature space Ψ, while they lie far from each other when considering our previous query frame and any other randomly extracted from a different video. So, the Eq. 1 can be re-written as,

$$\begin{split}\mathcal{L}_1(\Psi(V_t), \Psi(V_{t+1}), \Psi(V'_{t'})) = \sum_{t \in \mathcal{V}} Y \cdot d(\Psi(V_t), \Psi(V_{t+1})) + \\ (1-Y) \cdot \max\{0, \delta - d(\Psi(V_t), \Psi(V'_{t'}))\},\end{split} \tag{2}$$

being ${V'}_{t'}$ any other frame randomly extracted from a different video. In the feature space $\Psi(\cdot)$, we use the standard euclidean distance.

However, it is still possible to go one step further. We also propose an unsupervised learning model which jointly exploits the "slowness" and the image similarity present in video sequences, inspired by [3]. Technically, we enhance our loss function defined in Equation 2 as follows. The key idea is that a query frame and its adjacent frame should lie close in the feature space, while this query frame and a not neighboring frame, which belongs to the same video, should share a more similar representation than with a frame extracted from a different video.

Formally, given a set of N unlabeled videos $\mathcal{S} = \{V_1, V_2, \ldots, V_N\}$, for each video V_i, one can define $V_{i,t}$ as the query frame, $V_{i,t+1}$ as its neighboring frame, and $V_{i,t+n}$ with $n \neq 1$ as the not neighboring frame. We then define $V_{j,t'}$ as a frame randomly extracted from a different video V_j, which has been randomly selected from $\mathcal{S}$ too. Technically, we want to enforce that $d(\Psi(V_{i,t}), \Psi(V_{i,t+1})) \approx 0$ and $d(\Psi(V_{i,t}), \Psi(V_{i,t+n})) < d(\Psi(V_{i,t}), \Psi(V_{j,t'}))$. Therefore, we define our new loss as follows,

$$\begin{split}\mathcal{L}_2(\Psi(V_{i,t}), \Psi(V^{+}_{i,t+1}), \Psi(V_{i,t+n}), \Psi(V_{j,t'})) = \sum_{t \in \mathcal{S}} d(\Psi(V_{i,t}), \Psi(V_{i,t+1})) + \\ \max\{0, d(\Psi(V_{i,t}), \Psi(V_{i,t+n})) - d(\Psi(V_{i,t}), \Psi(V_{j,t'})) + \delta\},\end{split} \tag{3}$$

where δ represents the gap between two distances. For all experiments, we set the leap between frames $n = 20$, $\delta = 1$ for Siamese networks trained with $\mathcal{L}_1$ and $\delta = 0.5$ for Siamese networks trained with $\mathcal{L}_2$.

3 Experimental results

3.1 Experimental setup

We evaluate the unsupervised learning approaches for action discovery (see Section 3.2) using the UCF101 dataset [7]. It consists of over 12.000 videos categorized into 101 human action classes. The dataset is divided in three splits. For our experiments we use the split-1. Note that we train our models *without* the class labels provided. For the problem of action recognition in still images (see Section 3.3), we use the the PASCAL VOC 2012 dataset [1], following the experimental setup proposed in [3]. This dataset offers 10 action categories. The images are cropped using the action annotations provided. Only 50 images (10 classes *x* 5 images per class) are used during the fine-tuning stage. We test on 20000 images randomly extracted from the validation set.

For our experiments, we scale the video frames to a size of 227 × 227. The base network of our Siamese architecture is based on the AlexNet model [5] for the convolutional layers. Then we stack two fully connected layers on the pool5 outputs, whose neuron numbers are 4096 and 1024 for the model trained with $\mathcal{L}_2$, and 4096 and 4096 for the architecture trained with $\mathcal{L}_1$. We apply mini-batch SGD in training, using the Caffe framework. As the Siamese networks share the same parameters, we perform the forward propagation for the whole batch by a single network and calculate the loss based on the output feature. To train our Siamese and Siamese-tuple networks, we set the batch size to 120 pairs and 40 tuples of images, respectively. For both networks, the learning rate starts with $\varepsilon_0 = 0.001$.

Table 1: Comparison of different architectures on the UCF101 dataset. Measured as CE (lower is better) and Purity (higher is better).

Methods	CE	Purity (%)
Baseline RANDOM	6.31	3.33
AlexNet with $\mathcal{L}_1$ fc6	3.89	25.15
AlexNet with $\mathcal{L}_1$ fc7	3.95	24.84
AlexNet with $\mathcal{L}_2$ fc6	3.55	28.40
AlexNet with $\mathcal{L}_2$ fc7	3.57	28.12
AlexNet trained on ImageNet fc6	**2.68**	**41.51**
AlexNet trained on ImageNet fc7	2.89	39.76

3.2 Evaluation on Unsupervised Action Discovery

We here introduce a novel problem: unsupervised action discovery in videos. Essentially, given a set of unlabeled videos, the goal is to separate the different classes. We here evaluate how the features learned following our unsupervised approaches perform with a clustering algorithm. This idea is motivated by the work of Tuytelaars *et al.* [8] who discover object classes from collections of unlabeled images with partition methods too. We start characterizing the video frames with the learned feature representation Ψ, and then we run the clustering algorithm (K-means) to finally measure the quality of the clusters discovered. Like in [8], we use *Purity* and *Conditional Entropy* (CE) as the evaluation metrics for the discovery quality.

We compare the performance of the two models described in Section 2: Siamese network learned with loss $\mathcal{L}_1$ and loss $\mathcal{L}_2$. We also evaluate the performance in the same task, when the features for the clusterings are obtained using the fully supervised AlexNet model trained on ImageNet [5]. For all the methods, we show the results obtained performing k-means (with $k = 101$) over the features extracted from the fully connected layers 6 and 7 (fc6 and fc7). Results are summarized in Table 1. First, note that randomly assigning images to clusters results in a CE of 6.31. This value is close to the maximum $log_2(101) = 6.66$, giving this fact an idea of the difficulty of the proposed problem.

The best performing method is the fully supervised ImageNet model, using features extracted from layer 6. Our architecture (AlexNet with $\mathcal{L}_2$) works better than the Siamese network based on the Contrastive Loss (AlexNet with $\mathcal{L}_1$). We achieve a CE of 3.55 working with features extracted from layer 6. In other words, by applying our unsupervised procedure, the remaining uncertainty on the true object category has been reduced from $2^{6.31} = 79.34$ out of 101 for the random assignment, down to $2^{3.55} = 11.63$ out of 101 for our Siamese network trained with $\mathcal{L}_2$. If we now observe the *Purity* results, again, our proposal outperforms the Siamese network trained with $\mathcal{L}_1$.

Considering the Purity of the fully supervised AlexNet model trained with ImageNet as an upper bound for the performance, note that our unsupervised solution obtains a normalized purity of $\frac{28.40}{41.51} = 0.68$. We show qualitative results in Figure 1 for the actions discovered.

3.3 Unlabeled video as a prior for supervised action recognition

As a final experiment, we follow the experiment detailed in [3] for action recognition. We examine here how the proposed unsupervised feature learning model competes with the popular supervised pre-training plus fine-tuning paradigm. We believe that the unsupervised feature learning approaches have an important advantage: they can leverage essentially infinite unlabeled data without requiring expensive human labeling effort. The question is: can they compete with the fully supervised models?

Table 2 summarizes the main results. For the first group of rows we show the performance obtained by following the setup proposed in [3], *i.e.* a CIFAR CNN architecture [4] initialized with the features learned following our unsupervised model on 1000 videos clips randomly extracted from the HMDB51 dataset [6]. The second part of the table presents the performance achieved by the AlexNet based architectures, but using the UCF101 for the unsupervised learning of features.

First, one notices for this experiment that the network architecture seems to have little influence on the performance, except for the fully supervised pre-training models (CIFAR 20.22 vs. AlexNet 28.45). Note that this fact might be due to the large number of images with which the AlexNet model is trained on Imagenet.

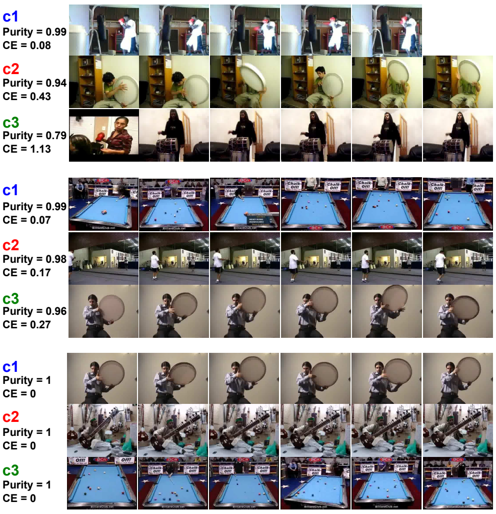

Figure 1: Qualitative results for K-mean clustering on the UCF101 dataset. 6 random images for the top three clusters discovered, in terms of purity. The first three rows correspond to the Siamese network trained with $\mathcal{L}_1$, the second group of three rows belong to the Siamese network trained with $\mathcal{L}_2$, and the third group correspond to the AlexNet trained on ImageNet.

Table 2: Results on PASCAL VOC 2012 Action dataset.

HMDB51 → PASCAL VOC 2012	Accuracy (%)
Baseline RANDOM	9.6
CIFAR net Init Random	15.20
CIFAR net initialized with $\mathcal{L}_2$	18.75
CIFAR net presented in [3]	**20.95**
CIFAR net supervised pre-trained on CIFAR-100	20.22
UCF101 → PASCAL VOC 2012	Accuracy (%)
AlexNet Init Random	15.10
AlexNet initialized with $\mathcal{L}_2$	18.15
AlexNet supervised pre-trained on ImageNet	**28.45**

Our Siamese network model outperforms the fully supervised CIFAR and AlexNet with random initialization! This means that our unsupervised learning strategy can be effectively used for the initialization of the networks. For the CIFAR network, our proposal even outperforms the fully supervised pre-training model when ≈ 17.400 training images are used (Our 18.75 vs. 17.35[1]). Our approach also obtains competitive numbers with respect to the best unsupervised model presented by [3] (20.95) and the supervised model pre-trained with all 50.000 images of CIFAR-100 (20.22). Finally, if the AlexNet model is used, our performance is far from the fully supervised approach (18.15 vs. 28.25). But the results seem promising, indicating that our models are able to learn effective visual representations for the action recognition task in a fully unsupervised manner.

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.

[2] R. Hadsell, S. Chopra, and Y. Lecun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

[3] D. Jayaraman and K. Grauman. Slow and steady feature analysis: Higher order temporal coherence in video. In *CVPR*, 2016.

[4] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[6] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011.

[7] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, 2012.

[8] T. Tuytelaars, C. H. Lampert, M. B. Blaschko, and W. Buntine. Unsupervised object discovery: A comparison. *IJCV*, 2010.

[9] L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 2002.

[1] This number has been provided by the authors of [3]