Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

A Review of Unsupervised Learning Architectures and Framework for Visual Data

V.S. Tondre
Brijlal Biyani Science College, Amravati.

Abstract: The Supervised learning is rapidly growing, in artificial intelligence and machine learning has been applied in data processing research. Many researchers have paid attention to supervised learning. In the last few years, there has been a growth towards to keep hold on unsupervised learning in research to improve performance in video pose estimation, detection, segmentation, sequencing images, and classification. It retains great success; when applying unsupervised learning for Computer Vision, Natural Language Processing, Networking, Visual data representation, Image Processing etc. The focus of this paper is to provide an overview of the architecture and framework of unsupervised learning in the domain of visual data in previous published papers and their benefits.

Keywords: Unsupervised Learning, Visual data, human pose estimation, human pose detection, pose segmentation.

1. Introduction:

Many important problems, such as speech recognition, machine translation and caption generation for images are solved by understanding temporal sequences of the AI-set. It is also applied on videos for recognition of action and generating natural language description. It uses either supervised or **unsupervised** learning. Supervised learning has been successful in learning good visual representations it produces good result at the task they are trained for and transfer well to other tasks and datasets[3].

The rapidly becoming grown-up field is supervised body pose estimation. Although, the huge training datasets are available, the research in this filed is restricted since it cannot be guaranteed for many kinds of human motions. The solution to this is to use unsupervised data to learn a low dimensional representation of poses. Because, it has good unsupervised learning objective, for which many existing techniques are available for multi-view footage.

Visual representations like videos are much higher dimensional entities as compared to single images. It requires costly and difficult efforts to do credit assignment and learn long range structure and to keep low dimensionality. The costly work of collecting more labeled data and tedious work to make clever engineering needs much time to solve particular problems. Therefore, the new learning method must be used to find and represent structures in videos i.e. unsupervised learning. Lots of structures are in videos, which makes them well suited as a domain n for building unsupervised learning models. It may perform better and useful in representations.

2. Architecture and Framework Of Unsupervised Learning For Visual Data

In this section, some widely used architecture is discussed for visual data:

2.1 Event Classes from Video:

Events are the planned human activities that are structured in terms of units. These events are generated according to the following three steps generative process.

In a domain of interest, it is supposed that for activities, there is an underlying prior probability distribution over sets of event classes. For a certain set of event classes, each class of event in this set is itself a probability distribution over a finite set of qualitative spatio-temporal graphs referred to as event graphs [1].

1. According to the probability distribution, Event graphs are sampled from the event classes.
2. A single structure is called as the activity graph. It is constructed by combining all the event graphs. It is

also specifying the spatio-temporal relationships between objects across different event graphs. The activity graph captures the spatio-temporal relations between all the objects that constitute the activities. The generation of activity graphs is influenced by a conditional distribution that favours certain activity graphs over others, given a set of event graphs.

3. The activity graph is embedded as tracks in space and time. Tracks are with concrete objects, spatial positions and temporal intervals.

In this method, the generative model provides a probabilistic framework which is used in the unsupervised event learning for the task of actually generating a set of tracks is called as setting. In this method, a set of tracks is observed for a video from a certain domain. The main goal behind this is to find the most likely event classes, event graphs. The activity graph could have generated the observed tracks. The posterior probability for any candidate interpretation is a measure of how it is that the candidate interpretation could have generated the observed set of tracks. By efficiently sampling the space of posterior distribution of candidate interpretations using MCMC a Maximum a Posterior (MAP) solution is found.

2.2 Image Manifolds By Semidefinite Programming:

In this paper authors K.Q. Weinberger and L.K. Saul proposed a new algorithm [2], for manifold learning based on semidefinite programming. It relies on the efficient and tractable optimizations, which is not overwhelmed by false local minima. It is based on completely different geometrical perception. A simple algorithm leads for computing low dimensional mappings that preserve the distance between nearby data points. In that, the inputs are pulled apart, maximize their total variance subject to the constraints imposed by the rigid rods.

2.3 Video Representations Using LSTMS Architecture:

To learn the video representation authors [3] uses the LSTMs Encoder–Decoder framework. It selects the right inductive biases and selects right objective function for learning useful features. The Encoder LSTMs runs through a sequence of frames to come up with a representation. This representation is then decoded through another LSTM to produce a target sequence[3]. Select some target sequences. One choice is to predict the same sequence as the input. It will collect those things that is needed to reproduces the input but at the same time go through the inductive bias considered by the model. Another way to predict the future frames is to learn a representation the extracts all that is needed to extrapolate the motion and appearance further than what has been observed. These two things can be combined. The input to the model may be any representation of individual video frames. It uses natural image patches as well as a dataset of moving MNIST digits as input. The second is high level percepts extracted by applying a conventional net trained on ImageNet.

2.4 Supervoxel Embedding For Video Segmentation:

The authors [4] influence the unsupervised learning technique to learn better feature representation for segmentation. This framework is divided into two sections. First is video segmentation and second deep embedding method. In video segmentation, it develops a principled framework for learning contextually- aware embedding video segments in an unsupervised setting. Here, video segments with the same context, means those which are in the same neighborhood. This is used to group small video segments into larger groups that represent either semantically consistent objects or actions that spontaneously share a common context. In this way, it groups supervoxels based on embeddings trained by framework. It uses a graph based partitioning framework using the embedded features to produce the final segmentation.

2.5 Ten-Layer Deep Neural Network For Video To Detect Foreground Objects In Single Images:

In computer vision, unsupervised learning is the most challenging task. The authors in [5] achieve the task of unsupervised learning to detect and segment foreground objects in single image. Here, it uses ten layer deep neural network i.e. a novel student-teacher architecture.

In this architecture, first it learns to detect and segment foreground objects in images in a completely unsupervised fashion. It didn't use any pre-trained features or manual labeling of single image for test. Secondly, unsupervised learning in video, it uses two processing pathways, with complementary functions and properties. The first

pathway discovers the foreground objects in video in an unsupervised manner. It has access to all the video frames. It acts as a teacher. The second is student pathway. When having a access to a single input image, it is a very difficult to learn to predict the teachers output for each frame.

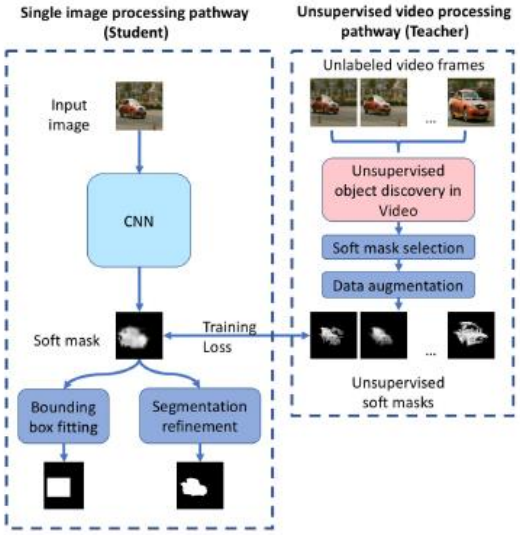

Fig.1. Student Teacher architecture

2.6 Model Architecture Constructive Self-Supervised (Css) Learning For Monocular Videos For 3d Human Pose Estimation:

The authors [6], presents a novel idea of constructive self-supervised (CSS) learning alternative unsupervised learning technique for human pose estimation using videos obtained by single RGB camera.

The main aim of the technique is without any supervision, to extract rich latent features from single-camera videos featuring people. It only captures the information only about the foreground subjects and to force the similarity or dissimilarity of frames given their distance in time[6]. It trains a given video database without footnote, which splits latent features into time-variant and time-invariant components. The previous capture features remain consistent over time. For ex. person's clothing and appearance and on the other hand time-changing elements in each frame i.e. latter models such as body poses. It facilitate to determine a similarities measure, which come across for the different behavior of these two components and to implement an effective contrastive-learning procedures. That may produces latent vectors useful for human pose estimation. Otherwise it would unable to reliably differentiate vectors for temporally away frames. Since no matter how distant they are, they still share information.

3 Analysis and Benefits Of The Methods:

In this section, all the above architectures and frameworks are discussed, analyzed, and list benefits of unsupervised learning for visual images:

3.1 On the predefined event classes, the proposed framework [1] gives a promising performance. It is new and promising direction for discovering semantically meaningful events on challenging videos with complex events, in spite of visual noise in the tracked input.

3.2 The distance preserving constraints in maximum variance unfolding can be relaxed to encourage more aggressive solutions for dimensionality reduction[2]. It can express and enforce distance-preserving constraints, such constraints can be tailored to particular applications of nonlinear dimensionality reduction is the main advantage of the use of semidefinite programming. The disadvantage of this method is the required amount of computation.

3.3 It uses a subset of the Sports-1M dataset and UCF-101 and HMDB-51 for unsupervised models. Author trained a two layer composite model adding depth helps the model make better predictions. It can change the future

predictor by making it conditional. It helps to improve classification accuracy, when there are a few training examples, it also helps in action recognition performance with even models on unrelated datasets[4].

3.4 For a given video, embedding provides a discriminative mapping of supervoxel feature representation[4]. This learning framework has the potential to help in separating a supervoxel's neighbor feature representation from the feature representation of a distant supervoxel.

3.5 By using relatively simple methods in [5], it is possible to detect and segmentation of object in single image for unsupervised object discovery in video to train a powerful deep neural network.

A system leran general visual characteristics that predict well the presence and shape of objects in images. The network effectively discovers appearance object fetures from single image, at different level of abstraction. During the unsupervised training phase, appreciably, the student network is able to outperforms its teacher by learning general objectiveness characterstics that are well beyound the capabilities of its teacher. It includes good form, closure, smooth, contours, as well as contrast with its background.

3.6 The framework in [6] is used to design to extract features from the foreground object. It is compatible to 3D human pose estimation. It is not limited for human pose estimation. This framework verified on three benchmark datasets. It is proved that it outperforms other single-view self-supervised learning strategies and it also matches the performance of multi-view ones.

4. Conclusion:

In this paper, verity of architectures and frameworks are studied. These are essential and very important for feature extraction, segmentation, detection and estimation in videos. Using the framework and approaches of unsupervised learning discussed in this paper are outperforms, improves the performance over the state-of-art on standard datasets. This study can be a focus for the researcher's attention toward unsupervised learning for more accurate prediction and analysis.

References:

[1] K. Q. Weinberger And L. K. Saul, "Unsupervised Learning of Image Manifolds by Semidefinite Programming", Springer Science, International Journal of Computer Vision 70(1), 77–90, May 2006.
[2] N. Srivastava, E. Mansimov, R. Salakhutdinov, "Unsupervised Learning of Video Representations using LSTMs", International Conference on Machine Learning, arXiv:1502.04681 **[cs.LG]** Mar. 2015.
[3] M. Khodabandeh, S. Murlidharan, A. Vahdat, N. Mehrasa E.M. Pereira, S. Satoh, and G. Mori, "Unsupervised Learning of Supervoxel Embeddings for Video Segmentation", 23rd International Conference on Pattern Recognition (ICPR), 2392-2397, 2016.
[4] I. Croitoru, S. Bogolin, M. Leordeanu, "Unsupervised learning from video to detect foreground objects in single images", in ICCV, 31 Mar. 2017.
[5] S. Honari, V. Costantin, H.Rhodin, M. Salzmann, P. Fua, "Unsupervised Learning on Monocular Videos for 3D Human Pose Estimation", *arXiv preprint arXiv:2012.01511*, 2021. Dec 2020.
[6] C.Redondo-Cabreara and R. J. Lopez-Sastre, "Unsupervised Feature Learning from Videos for Discovering and Recognizing Actions ", 1st Workshop on Action and Anticipation for Visual Learning (ECCV) Oct. 2016.
[7] R. Raina, A. Battle, H. Lee, B.Packer and A. Y. Ng, "Self Taught Learning: Transfer learning from unlabeled data" ICML, 2007
[8] X. Wang and A. Gupta, "Unsupervised learning of visual representation using videos", CVPR, 2015
[9] H. Lee, J. Huang, M. Singh, M. Yang, "Unsupervised Representation Learning by Sorting Sequences", IEEE International Conference on Computer Vision, pp. 667-676, 2017.
[10] C.Doersch, A. Gupta, and A.A. Efros, "Unsupervised visual representation learning by context prediction", in ICCV, 2015.
[11] J. Xie, X. Zhan, Z. Liu, Y. Ong, C, Loy, "Unsupervised Object-Level Representation Learning from Scene Images", 35th Conference on Neural Information Processing Systems (NeurIPS), 2021.
[12] C. Zhuang, T. She, A. Andonian, M. Mark, D. Yamins, "Unsupervised Learning from Video with Deep

Neural Embeddings", CVPR, Computer Vision Foundation, pp-9563-9572, 2020.
[13] P. Hong, C. Ahn , "unsupervised learning for Stereo Matching Using Single-View Videos", IEEE Access, Vol. 8 , pp. 73804-73815, 2020.
[14] Y. Liu, Y. Li, S. You, F. Lu, "Unsupervised learning for Intrinsic Image Decomposition from a Single Image", CVPR, pp 3248-3257, 2020
[15] M. Minderer, C. Sun, R.Villegas, F. Cole, K. Murphy, H. Lee, "Unsupervised Learning of Object Structure and Dynamics from Video", 33rd Conference on Neural Information Processing System Vancouver, Canada, 2019.
[16] T. Jakab, A. Gupta, H. Bilen, A. Vedaldi, "Unsupervised Learning of Object Landmarks Through Conditional Image Generation", 32ndConference on Neural Information Processing System Vancouver, Canada, 2018
[17] H. Fan, M. KankanHalli, "Motion = Video-Content: Towards Unsupervised Learning of Motion Representation from Videos", MMASia'21, ACM, Dec 1-3, 2021
[18] M. Sridhar, D. Hogg, A. Cohn, "Unsupervised Learning of Event Classes from Video", Proceedings of the 24th AAAI Conference on Artificial Intelligence, pp 1631-1638, 2010.