International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 11 Nov. 2022, pp: 1191-1197 www.ijaem.net ISSN: 2395-5252

Systematic Review of Deep Learning Techniques for Visual Feature Representation and Learning

Rupali Tabakade , Dr. Varsha Jotwani
Research Scholar, Rabindranath Tagore University
Associate Professor Rabindranath Tagore University, Bhopal

Date of Submission: 25-11-2022 Date of Acceptance: 06-12-2022

ABSTRACT :- Visual features representation along with deep learning techniques have a great area of research today as perspective of industries like facebook AI research. These industries aimed to focus on deep features learning with dataset and self-learning model. Most recent efforts in unsupervised feature learning already existed on either small or highly created datasets like ImageNet, whereas using non-curated raw datasets was found to decrease the feature quality when evaluated on a transfer task. From past 5-10 years lot of experimental research has been published and growing path of new ideas to improve accuracy in proposed system is not yet stopped. Paper has a systematic conclusion, the work already done in area related to visual feature representation, similarity computation methods and experimental results comparison. Our goal with paper is to bridge the performance gap with lot many existing techniques of deep leaning.

Index Terms :- Deep features, Deep leaning, Self Learning model.

I. INTRODUCTION:-

Computer vision has been revolutionized by high capacity Convolution Neural Networks (ConvNets) and large-scale labeled data. Recently weakly-supervised training on hundreds of millions of images and thousands of labels has achieved state-of-the-art results on various benchmarks. Interestingly, even at that scale, performance increases only log linearly with the amount of labeled data. Thus, sadly, what has worked for computer vision in the last five years has now become a bottleneck: the size, quality, and availability of supervised data [11].

Unsupervised representation learning is highly successful in natural language processing, e.g., as shown by GPT and BERT [2]. But supervised pre-training is still dominant in computer vision, where unsupervised methods generally lag behind. The reason may stem from differences in their respective signal spaces. Language tasks have discrete signal spaces (words, sub-word units, etc.) for building tokenized dictionaries, on which unsupervised learning can be based. Computer vision, in contrast, further concerns dictionary building, as the raw signal is in a continuous, high-dimensional space and is not structured for human communication (e.g., unlike words). Several recent studies present promising results on unsupervised visual representation learning using approaches related to the contrastive loss. Though driven by various motivations, these methods can be thought of as building dynamic dictionaries. The “keys” (tokens) in the dictionary are sampled from data (e.g., images or patches) and are represented by an encoder network. Unsupervised learning trains encoders to perform dictionary look-up: an encoded “query” should be similar to its matching key and dissimilar to others. Learning is formulated as minimizing a contrastive loss.

Unsupervised learning has been widely studied in the Machine Learning community [9], and algorithms for clustering, dimensionality reduction or density estimation are regularly used in computer vision applications. For example, the \bag of features" model uses clustering on handcrafted local descriptors to produce good image-level features [4]. A key reason for their success is that they can be applied on any specific domain or dataset, like satellite or medical images, or on images captured with a new modality, like depth, where annotations are not always available in quantity. Several works have shown that it was possible to adapt unsupervised methods based on density estimation or dimensionality reduction to deep models.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 11 Nov. 2022, pp: 1191-1197 www.ijaem.net ISSN: 2395-5252

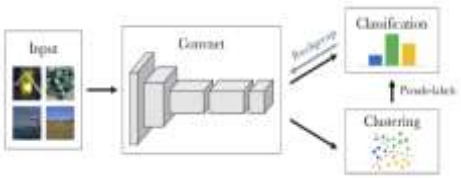


Figure 1: Illustration of the Convent method with clustering and classification.

1.1 Deep Neural Network:

Deep neural networks excel at perceptual tasks when labeled data are abundant, yet their performance degrades substantially when provided with limited supervision (In below fig, red). In contrast, humans and animals can learn about new classes of images from a small number of examples. What accounts for this monumental difference in data-efficiency between biological and machine vision? While highly structured representations may improve data-efficiency, it remains unclear how to program explicit structures that capture the enormous complexity of real world visual scenes, such as those present in the ImageNet dataset. An alternative hypothesis has therefore proposed that intelligent systems need not be structured a priori, but can instead learn about the structure of the world in an unsupervised manner. Choosing an appropriate training objective is an open problem, but a potential guiding principle is that useful representations should make the variability in natural signals more predictable. Indeed, human perceptual representations have been shown to linearize (or 'straighten')the temporal transformations found in natural videos, a property lacking from current supervised image recognition models, and theories of both spatial and temporal predictability have succeeded in describing properties of early visual areas.

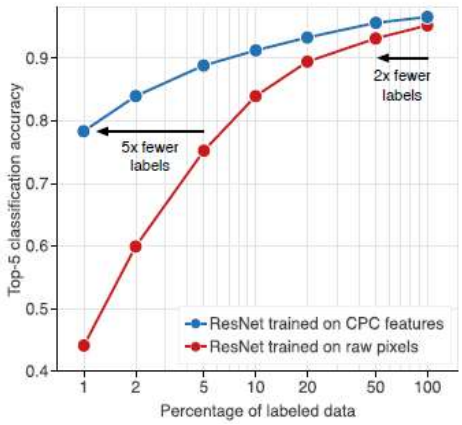


Figure 2: Data-efficient image recognition with Contrastive Predictive Coding. With decreasing amounts of labeled data, supervised networks trained on pixels fail to generalize (red). When trained on unsupervised representations learned with CPC, these networks retain a much higher accuracy in this low-data regime (blue) [3].

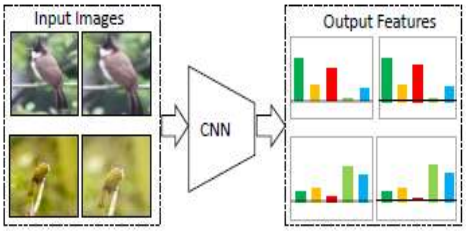


Figure 3: Illustration of our basic idea. The features of the same instance under different data augmentations should be invariant, while features of different image instances should be separated [7].

1.2 Convolutional Neural Network:

Pre-trained convolutional neural networks, or Convnets, have become the building blocks in most computer vision applications. They produce excellent general-purpose features that can be used to improve the generalization of models learned on a limited amount of data. The existence of ImageNet [6], a large fully-supervised dataset, has been fueling advances in pre-training of convents However, As a matter of fact, ImageNet is relatively small by today's standards; it \only" contains a million images that cover the specific domain of object classification. A natural way to move forward is to build a bigger and more diverse dataset, potentially consisting of billions of images. This, in turn, would require a tremendous amount of manual annotations, despite the expert knowledge in crowd sourcing accumulated by the community over the years. Replacing labels by raw metadata leads to biases in the visual representations with unpredictable consequences [4]. Learning a deep neural network together while discovering the data labels can be viewed as simultaneous clustering and representation learning. The latter can be approached by combining cross-entropy minimization with an off-the-shelf clustering algorithm such as K-means. This is precisely the approach adopted by the recent DeepCluster method, which achieves excellent results in unsupervised representation learning. However, combining representation learning, which is a discriminative task, with clustering is not at all trivial. In particular, we show that the combination of cross-entropy minimization and K-means as adopted by DeepCluster cannot be described as the optimization of an overall learning objective; instead, there exist degenerate solutions

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 11 Nov. 2022, pp: 1191-1197 www.ijaem.net ISSN: 2395-5252

that the algorithm avoids via particular implementation choices [9].

II. DISCUSSION:

2.1 Detailed Analysis of different approaches applied:

Human observers can learn to recognize new categories of images from a handful of examples; yet doing so with artificial ones remains an open challenge. The efficient recognition of data is enabled by representations which make the variability in natural signals more predictable. Therefore revisit and improve Contrastive Predictive Coding is a better solution, as unsupervised objective for learning such representations. The given table-1 express an analysis of similarity measured in previous year for features of leafs. Some new implementation produces features which support state-of-the art linear classification accuracy on the ImageNet dataset. When used as input for non-linear classification with deep neural networks, this representation allows us to use 2–5x fewer labels than classifiers trained directly on image pixels.

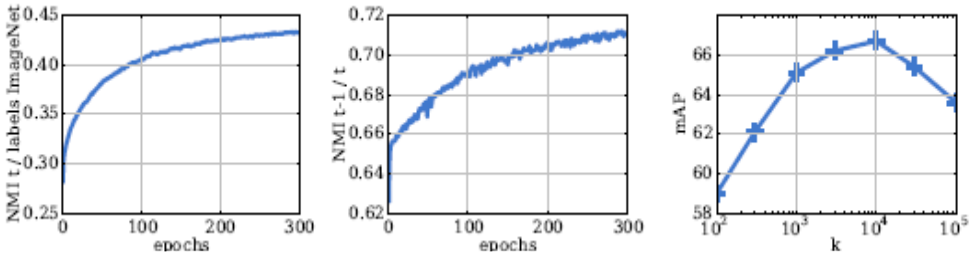
Table 1: Analysis of Feature Extraction Techniques based on similarities (leaf data) [17]

SI. No.	Authors & Year	Methodology/ Approach	Description
1	De Chant S. et.al. 2017, [12]	Deep CNN model have been used for extraction of local and global features of the input image.	They applied the method on maize leaves and used for prediction of disease Northern leaf blight (applied for binary classification only).
2.	Yang Lu et.al. , 2018, [13]	6 layer CNN network proposed for feature extraction with the use of 3 convolution layers , 1 for extraction of low level features other two for extraction of high level features.	16 features are extracted by using 3 convolution and 3 max pooling filters and applied on rice plant diseases of 10 classes. Classification accuracy of 95.48% achieved.
3	Nikos Petrellis, 2019, [14]	Color, area and the number of the lesion spots featured have extracted. Then feature have been put along with additional information like weather metadata, to create disease signature.	Novel method of automated and manual feature extortion has been proposed, where, user's input can also considered as a feature. They do this task using smart phones.
4.	Manso L. el.al. , 2019 [15]	Mathematical equations have been used to extract texture features (like contrast, entropy, homogeneity) and color features (like mean, variance etc.)	15 features extracted, which includes texture and color features only but not considered the affected area of leaf.
5.	Saradhambal. G. et.al. , 2019, [16]	k means clustering for segmentation have been used for segmentation, and shape and texture features are considered as a main features which are calculated by mathematical equations.	Total 10 features extracted. 5 of which are shape features (like area, perimeter, no of component etc.) and 5 are texture features (like contrast, entropy, co relation etc.)

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 11 Nov. 2022, pp: 1191-1197 www.ijaem.net ISSN: 2395-5252

The unsupervised representation substantially improves transfer learning to object detection on the PASCAL VOC dataset, surpassing fully supervised pre-trained ImageNet classifiers. A main purpose of unsupervised learning is to pre-train representations (i.e., features) that can be transferred to downstream tasks by fine-tuning. Authors present Momentum Contrast (MoCo) for unsupervised visual representation learning, a perspective on contrastive learning as dictionary look-up, they build a dynamic dictionary with a queue and a moving-averaged encoder. This enables building a large and consistent dictionary on-the-fly that facilitates contrastive unsupervised learning. MoCo provides competitive results under the common linear protocol on ImageNet classification. More importantly, the representations learned by MoCo transfer well to downstream tasks. MoCo can outperform its supervised pre-training counterpart in 7 detection/segmentation tasks on PASCAL VOC, COCO, and other datasets, sometimes surpassing it by large margins.

Clustering is a class of unsupervised learning methods that has been extensively applied and studied in computer vision. Little work has been done to adapt it to the end-to-end training of visual features on large scale datasets. In this work, author [4] presents DeepCluster, a clustering method that jointly learns the parameters of a neural network and the cluster assignments of the resulting features. DeepCluster iteratively groups the features with a standard clustering algorithm, k-means, and uses the subsequent assignments as supervision to update the weights of the network. They apply DeepCluster to the unsupervised training of convolution neural networks on large datasets like ImageNet and YFCC100M. The primary study and experiment on visual corps image features provides given effect as in figure 4 with respect of clustering goodness [4]. Effect of the experiment is to train DeepCluster on ImageNet [6] unless mentioned otherwise. It contains 1:3M images uniformly distributed into 1; 000 classes.



(a) Clustering Quality (b) Clustering Reassignment (c) Influence of k

Figure 4: (a): evolution of the clustering quality along training epochs, (b): evolution of cluster reassignments at each clustering step;(c): validation mAP classification performance for various choices of k.

2.2Findings and Comparisons Summary for different network & dataset:

As a clustering the unsupervised learning and deep clustering provides a good performance and continuously engaged by many research. It showed the trust of deep clustering algorithm by many researchers. Table 2: shows the work carried by many researches with improves affection on using more convolution layers for ImageNet. DeepCluster outperforms the state of the art from conv3 to conv5 layers by 3 to 5%. The largest improvement is observed in theconv4 layer, while the conv1 layer performs poorly, probably because the Sobel filtering discards color. Linear classification on ImageNet and Places using activations from the convolutional layers of an AlexNet as features. We report classification accuracy on the central crop. Numbers for other methods are from Zhang.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 11 Nov. 2022, pp: 1191-1197 www.ijaem.net ISSN: 2395-5252

Table 2: Analysis by different authors on linear classification on ImageNet and Place

	ImageNet					Place				
Method	Conv1	Conv2	Conv3	Conv4	Conv5	Conv1	Conv2	Conv3	Conv4	Conv5
Place Lables	Na	Na	Na	Na	Na	22.1	35.1	40.2	43.3	44.6
ImageNetLables	19.3	36.3	44.2	48.3	50.5	22.7	34.8	38.4	39.4	38.7
Random	11.6	17.1	16.9	16.3	14.1	15.7	20.3	19.8	19.1	17.5
Pathak[17]	14.1	20.7	21	19.8	15.5	18.2	23.2	23.4	21.9	18.4
Doersch[18]	16.2	23.3	30.2	31.7	29.6	19.7	26.7	31.9	32.7	30.9
Zhang[19]	12.5	24.5	30.4	31.5	30.3	16	25.7	29.6	30.3	29.7
Donahue[20]	17.7	24.5	31	29.9	28	21.4	26.2	27.1	26.1	24
Noroozi and Favaro[21]	18.2	28.8	34	33.9	27.1	23	32.1	37.5	34.8	31.3
Noroozi[22]	18	30.6	34.3	32.5	25.7	23.3	33.9	36.3	34.7	29.6
Zhang[23]	17.7	29.3	35.4	35.2	32.8	21.3	30.7	34	34.1	32.5
Deep Cluster	12.9	29.2	38.2	39.8	36.1	18.6	30.8	37	37.5	33.1

Consistently the conclusive Comparisons on ImageNet linear classification with different techniques have been judged on 100- 400 epoc. All are based on ResNet-50 pre-trained with two 224 x 224 views. The transfer leaning is a new area where we find almost similar kind results statistics. Transfer Learning. All unsupervised methods are based on 200-epoch pre-training in ImageNet. VOC 07 detection: Faster R-CNN fine-tuned in VOC 2007 trainval, evaluated in VOC 2007 test; VOC 07+12 detection: Faster R-CNN fine-tuned in VOC 2007 trainval + 2012 train, evaluated in VOC 2007 test; COCO detection and COCO instance segmentation: Mask R-CNN [18] fine-tuned in COCO 2017 train, evaluated in COCO 2017 value has been shown in table 3:

Table 3: Comparisons on ImageNet linear classification on some latest networks

Method	Batch Size	Negative Pairs	Momentum Encoder	100 ep	200 ep	400 ep	800 ep	VOC 07 detection			VOC 07+12 detection			COCO detection			COCO instance seg		
								AP50	AP	AP75	AP50	AP	AP75	AP50	AP	AP75	AP50	AP	AP75
SimCLR (repro.+)	4096	Yes	No	66.5	68.3	69.8	70.4	75.9	46.8	50.1	81.8	55.5	61.4	57.7	37.9	40.9	54.6	33.3	35.3
MoCo v2 (repro.+)	256	Yes	Yes	67.4	69.9	71	72.2	77.1	48.5	52.5	82.3	57	63.3	58.8	39.2	42.5	55.5	34.3	36.6
BYOL (repro.)	4096	No	Yes	66.5	70.6	73.2	74.3	77.1	47	49.9	81.4	55.3	61.1	57.8	37.9	40.9	54.3	33.2	35
SwAV (repro.+)	4096	No	No	66.5	69.1	70.7	71.8	75.5	46.5	49.6	81.5	55.4	61.4	57.6	37.6	40.3	54.2	33.1	35.1

Unsupervised image representations have significantly reduced the gap with supervised pre-training, notably with the recent achievements of contrastive learning methods. These contrastive methods typically work online and rely on a large number of explicit pair wise feature comparisons, which is computationally challenging. The online algorithm, SwAV, takes advantage of contrastive methods without requiring computing pair wise comparisons. Specifically, our method simultaneously clusters the data while enforcing consistency between cluster assignments produced for different augmentations (or "views") of the same image, instead of comparing features directly as in contrastive learning. Simply put, they use a "swapped" prediction mechanism where they predict the code of a view from the representation of another view. The method can be trained with large and small batches and can scale to unlimited amounts of data. Compared to previous contrastive methods, our method is more memory efficient since it does not require a large memory bank or a special momentum network.

One core objective of deep learning is to discover useful representations, and the simple idea explored here is to train a representation-learning function, i.e. an encoder, to maximize the mutual information (MI) between its inputs and outputs. This work investigates unsupervised learning of representations by maximizing mutual information between an input and the output of a deep neural network encoder. Importantly, [6] they show that structure matters: incorporating knowledge about locality in the input into the objective can significantly improve a representation's suitability for downstream tasks. They further control characteristics of the representation by matching to a prior distribution adversarial. Their method, which they call Deep InfoMax (DIM), outperforms a number of popular unsupervised learning methods and compares favorably with fully-supervised learning on several classification tasks in with some standard architecture. Siamese networks are general models for comparing entities. Their applications include signature and face verification, tracking, one-shot learning, and others. In conventional use cases, the inputs to Siamese networks are from different images, and the comparability is determined by supervision. Siamese networks have become a common structure in various recent models for unsupervised visual representation learning. These models maximize the similarity between two augmentations of one image, subject to certain conditions for avoiding collapsing solutions [1]. In this paper, they report surprising empirical results that simple Siamese networks can learn meaningful representations even using none of the following: (i) negative sample pairs, (ii) large batches, (iii) momentum encoders. Their experiments show that collapsing solutions do exist for the loss and structure, but a stop-gradient operation plays an essential role in preventing collapsing. They also provide a hypothesis on the implication of stop-gradient, and further show proof-of-concept experiments verifying it.

III. CONCLUSION:

Learning visual representations with self-supervised learning has become popular in computer vision. The auxiliary tasks models where labels are free to obtain have most tasks end up providing data to learn specific kinds of invariance useful for recognition. In many articles the exploitation of different self-supervised approaches to learn representations invariant to (i) inter-instance variations (two objects in the same class should have similar features) and (ii) intra-instance variations (viewpoint, pose, deformations, illumination, etc.). Instead of combining two approaches with multi-task learning, they argue to organize and reason the data with multiple variations. Specifically, they propose to generate a graph with millions of objects mined from hundreds of thousands of videos.

Competiveness of minimalist method suggests shape is an important core reason for effectiveness. Representation learning focuses on modeling invariance by different network. Lot many survey and statistics represents that unsupervised learning in variety of computer vision task give and shows better results. MoCo's improvements are considerable and noticeable for small dataset and suggest that it may not be used for large scale data. MoCo can be used with pretext task for constructive learning.

REFERENCES:

[1]. Xinlei Chen, Kaiming He, "Exploring Simple Siamese Representation Learning", IEEE 2020, pp. 1-10.

[2]. Kaiming He, Haoqi Fan, Yuxin Wu, SainingXie,Ross Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning", IEEE 2020, pp. 1-12.

[3]. Olivier J. Henaff, AravindSrinivas, Jeffrey De Fauw, Ali Razavi,Carl Doersch, S. M. Ali Eslami, Aaron van den Oord, "Data-Efficient Image Recognition with Contrastive Predictive Coding", 2020, pp. 1-13.

[4]. Mathilde Caron, PiotrBojanowski, Armand Joulin, MatthijsDouze, "Deep Clustering for Unsupervised Learningof Visual Features", 209, pp. 1-30.

[5]. Mathilde Caron, Ishan Misra, JulienMairal, Priya Goyal, Piotr Bojanowski, Armand Joulin, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments", 34th Conference on Neural Information Processing Systems, 2020, pp. 1-23.

[6]. R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, Yoshua Bengio, "Learning Deep Representations By Mutual Information Estimation And Maximization", Published as a conference paper at ICLR 2019, pp. 1-24.

[7]. Mang Ye, Xu Zhang, Pong C. Yuen, Shih-Fu Chang, "Unsupervised Embedding Learning via Invariant and Spreading Instance Feature", 2019, pp. 1-11.

[8]. Mathilde Caron, PiotrBojanowski, JulienMairal, Armand Joulin, "Unsupervised Pre-Training of Image Features on Non-Curated Data", 2019, pp. 1-14.
[9]. Yuki M. Asano, Christian Rupprecht, Andrea Vedaldi, "Self-LabellingVia Simultaneous Clustering And Representation Learning", Published as a conference paper at ICLR 2020, pp. 1-22.
[10]. Xiaolong Wang, Kaiming He, Abhinav Gupta, "Transitive Invariance for Self-supervised Visual Representation Learning", 2018, pp. 1329-1338.
[11]. PriyaGoyal, DhruvMahajan, Abhinav Gupta, IshanMisra, "Scaling and Benchmarking Self-Supervised Visual Representation Learning", IEEE 2018, pp. 6391-6400.
[12]. DeChant, C., Wiesner-Hanks, T., Chen, S., Stewart, E.L., Yosinski, J., Gore, M.A., Nelson, R.J., Lipson, H., "Automated identification of northern leaf blight-infected maize plants from field imagery using deep learning". Phytopathology (2017).https://doi.org/10.1094/PHYTO-11-16-0417-R.
[13]. Lu, Yang, et al. "Identification of rice diseases using deep convolutional neural networks." Neuro computing 267 (2017): 378-384.
[14]. Petrellis, Nikos. "Plant disease diagnosis for smart phone applications with extensible set of diseases." Applied Sciences 9.9 (2019): 1952.
[15]. Giuliano L. Mansoa, HelderKnidel, Renato A. Krohlinga, José A. Ventura, "A smart phone application to detection and classification of coffee leaf miner and coffee leaf rust", Preprint submitted to Journal of LATEX Templates, arXiv:1904.00742v1 [cs.CV] 19 Mar 2019.
[16]. Saradhambal.G, Dhivya.R, Latha.S, R.Rajesh," Plant Disease Detection And Its Solution Using Image Classification", International Journal of Pure and Applied Mathematics, Volume 119 No. 14 2018, 879-884, pp. 879-884.
[17]. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context en-coders: Feature learning by in painting. In: CVPR. (2016)
[18]. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV. (2015)
[19]. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV. (2016)
[20]. Donahue, J., Kr¨ahenb¨uhl, P., Darrell, T.: Adversarial feature learning. arXivpreprint arXiv:1605.09782 (2016)
[21]. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solvingjigsaw puzzles. In: ECCV. (2016)
[22]. Noroozi, M., Pirsiavash, H., Favaro, P.: Representation learning by learning tocount. In: ICCV. (2017).
[23]. Zhang, R., Isola, P., Efros, A.A.: Split-brain autoencoders: Unsupervised learning by cross-channel prediction. arXiv preprint arXiv:1611.09842 (2016)