ISSN: 2583-5637

International journal of Inventive Research in Science and Technology
Volume 5 Issue 1 January 2026

AN IMAGE SIMILARITY COMPUTATION MODEL USING CONVOLUTIONAL NEURAL NETWORKS: A Review

Aryan Rawat
M.Tech Scholar
Lakshmi Narain College of Technology, Bhopal

Dr Vivek Richhariya
Professor
Lakshmi Narain College of Technology, Bhopal

ABSTRACT: The classification of objects into the right classes has long been one of the most crucial objectives of machine learning or deep learning. Due to the similarities between various things, their textures, colors, and other physical properties, object recognition still presents significant challenges despite the importance of categorizing specific groups of images. In computer vision, object detection has a wide range of uses, such as face and vehicle detection, video surveillance, and plant leaf detection. When conducting studies on flowers, using flowers as medicine, analyzing floral patents, etc., automatic classification of flowers is crucial. Traditionally, low-level characteristics like color, shape, texture, and geometry are used to classify flowers. The feature description has a significant impact on how accurately and robustly flowers are classified. In recent years, deep features have demonstrated good performance on high-resolution photos, but they are unable to extract precise global features from low-resolution images. Deep neural networks have been widely used in computer vision applications because they are effective at identifying picture patterns. An advanced deep-learning model that can precisely calculate the similarity of floral photos is required in the field of flower image analysis. In comparison to the given model mentioned in the exhibit results, the proposed model performs adequately. The proposed network can still be improved in terms of learning parameters, validation accuracy, loss, and training time. Fine-tuned deep learning models for similarity computation on flower datasets have a bright and broad future ahead of them, with lots of room for development and use. In this research work, the future potential of optimized deep learning models for the calculation of similarity on floral datasets is promising for growth and innovation. These models have the power to fundamentally alter the way we see, value, and engage with the world of flowers. The proposed model is supported by augmentation so the distance method is used against two augmentation views of the trained model being calculated.

Keywords: Machine learning, deep learning, object recognition, convolutional neural network, similarity computation.

1.INTRODUCTION

The ability to identify the objects present in an image or scene is one of the most basic requirements when it comes to interacting with one‟s environment. While it seems completely effortless with humans and in fact most animals, trying to teach computers to see and also understand" what they are seeing has proven extremely difficult. The key to understanding visual scenes are three closely related sub-problems. The easiest one will be called classification in the following. For classification, the one dominant object in a given image should be determined and labelled. The next more demanding task is object localization: In addition to labeling the dominant object, it also needs to be localized in the image, usually by determining a bounding box around the image region that is occupied by the object. The difficulty of this task again increases if not only one but all objects in an image need to be labeled, and multiple objects of the same category can appear in one image, with Figure 1.1 showing the process of object detection.

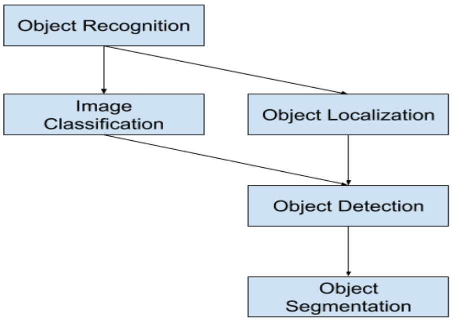

Figure 1.1: Object detection process

Image degradation is inevitable during the transmission and alteration of images. For example, the quality of an image shot by a camera is sometimes low due to the distortion of the camera’s optics scheme, the relative motion of the photographed object and the camera, the ecological change and the arbitrary disturbance [1]. The image enhancement is an important technique that can improve the quality of the degraded image and offer some interesting image features selectively. Image enhancement is basically improving the interpretability or perception of information in images for human viewers and providing better input for other automated image processing techniques. The main objective of image enhancement is to modify attributes of an image to make it more suitable for a given task and a specific spectator. For the duration of this process, one or more characteristic of the image are customized [2]. The alternative of attributes and the way they are customized are specific to a given problem. Moreover, observer-specific factor, such as the person visual system and the observer's experience, will bring in a great deal of subjectivity into the choice of image enhancement methods.

1.2 IMAGE FEATURES Image features refer to the information collected from images that can uniquely identify the image or can be used for further processing. Broadly, image features can be classified into general features and domain-specific features [5]. General features, such as color and texture are applicable to all image data and do not depend on the application being considered. Domain-specific features on the other hand, are specific to the application at hand, such as, minutiae in fingerprints. Figure 1.2 showing the different features of an image. Based on the locality of features, image features can be categorized into [6]:

(i) Local features: Local features are the patterns in images that differ from its immediate neighborhood. These features are extracted from a patch in the image and are useful in applications such as object recognition. Some examples of local features are Shape Invariant Feature Transform (SIFT), Local Binary Pattern (LBP), and Speeded up Robust Features (SURF).

(ii) Global features: Global features represent the whole image. These features are extracted considering the whole image as one patch/object and are useful in applications such as image retrieval and image classification, where a rough segmentation of objects is available. Some examples of global features are Histogram Oriented Gradient (HOG) and Shape Matrices.

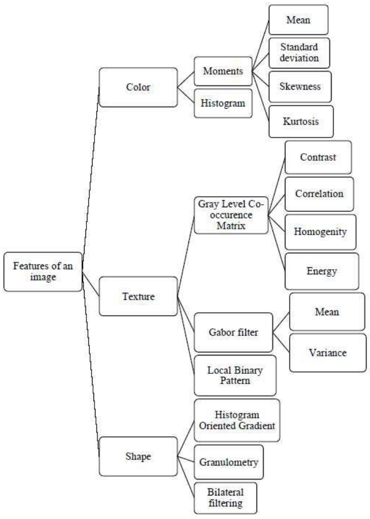

Figure 1.2: Features of an image

1.3 CLASSIFICATION Once a network is trained against a subset of valid objects, one of the tasks that the network can be used for is classification. In the simplest form, a classification problem can be stated as such: given an object and a set, is this object in the set. One of the underlying principles of the deep learning architecture is the reconstruction of valid objects into their original pattern. Of course, if a random image is sent into such a network, it will not resemble itself very well. However, if the object is a close relation to the objects trained, it should be reconstructed with a high fidelity. Figure 1.4 present image classification based on image features. Using this, a simple threshold can be established, and if the reconstruction is in error beyond this threshold, it can be declared not in the set.

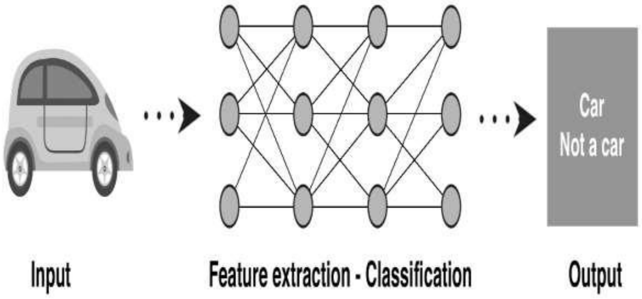

Figure 1.4: Image feature based classification process

1.4 APPLICATION CASES OF CLASSIFICATION Application cases of image classification are classified into scene classification, object detection and object extraction. Scene classification is the process of determining the type of a remote sensing image based on its content. Object detection is the process of determining the locations and types of the targets to be detected in a remote sensing image and labeling their locations and types with bounding boxes. Object extraction is the process of determining the accurate boundaries of the objects to be extracted in a remote sensing image. In this section, we summarize these application cases.

Scene classification Scene classification is a mapping process of learning and discovering the semantic content tags of image scenes. Generally, an image scene is a collection of multiple independent geographic objects. These objects have different structures and contain different texture information, and they form different types of scenes through different combinations and spatial locations. For scene classification studies in the remote sensing field, the UC Merced land use dataset is commonly viewed as the reference dataset.

Object detection Object detection from remote sensing images detects the locations and the types of objects. The object detection application cases from remote sensing images use the candidate region-based object detection method. The method involves three steps: the generation of candidate regions, feature extraction by the image classification techniques and classification of candidate regions. Candidate regions are a series of locations in which the objects may appear in the pre-generated image. All of these locations will be used as the input for the image classification techniques for feature extraction and classification.

Object segmentation To extract objects from a remote sensing image, it is necessary to segment the objects of interest in the image and to produce a pixel-level image classification map. Two types of methods are primarily used in the existing CNN-based studies on object segmentation from remote sensing images, namely patch-based CNN methods and end-to-end CNN methods. A patch-based CNN method generally first obtains a prediction model by training a CNN on a training dataset, and then, based on the prediction model, it generates image patches using a sliding window pixel by pixel and predicts the type of each pixel of the image.

1.5 COMPUTER VISION 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. Unsupervised representation learning is highly successful in natural language processing. 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 [16], 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 [14]. 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.

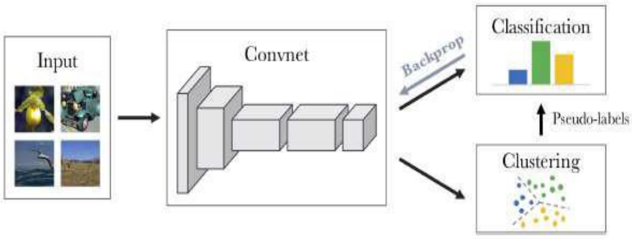

Figure 1.5: Illustration of the convent method with clustering and classification

2. LITERATURE REVIEW

2.1 PREVIOUS WORK DONE

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. A main purpose of unsupervised learning is to pre-train representations (i.e., features) that can be transferred to downstream tasks by fine-tuning

[2]. Here author present Momentum Contrast (MoCo) for unsupervised visual representation learning. From 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. 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.

They [3] hypothesize that data-efficient recognition is enabled by representations which make the variability in natural signals more predictable. They therefore revisit and improve Contrastive Predictive Coding, an unsupervised objective for learning such representations. This 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 less labels than classifiers trained directly on image pixels. 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 research work, author

[4] present Deep Cluster, a clustering method that jointly learns the parameters of a neural network and the cluster assignments of the resulting features. Deep Cluster 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. Unsupervised image representations have significantly reduced the gap with supervised pretraining, 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. In this research work

[5] They propose an online algorithm, SwAV, that takes advantage of contrastive methods without requiring computing pairwise 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. Their 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. The main challenge of unsupervised embedding learning is to discover visual similarity or weak category information from unlabeled samples.

This research works [7] studies the unsupervised embedding learning problem, which requires an effective similarity measurement between samples in low dimensional embedding space. Motivated by the positive concentrated and negative separated properties observed from category-wise supervised learning, they propose to utilize the instance wise supervision to approximate these properties, which aims at learning data augmentation invariant and instance spread out features. To achieve this goal, they propose a novel instance based softmax embedding method, which directly optimizes the „real‟ instance features on top of the softmax function. It achieves significantly faster learning speed and higher accuracy than all existing methods. The proposed method performs well for both seen and unseen testing categories with cosine similarity. It also achieves competitive performance even without pretrained network over samples from fine-grained categories. Pre-training general-purpose visual features with convolution neural networks without relying on annotations is a challenging and important task. Most recent efforts in unsupervised feature learning have focused on either small or highly curated datasets like ImageNet, whereas using non-curated raw datasets was found to decrease the feature quality when evaluated on a transfer task.

Their [8] goal is to bridge the performance gap between unsupervised methods trained on curated data, which are costly to obtain, and massive raw datasets that are easily available. To that effect, they propose a new unsupervised approach which leverages self-supervision and clustering to capture complementary statistics from large-scale data. Combining clustering and representation learning is one of the most promising approaches for unsupervised learning of deep neural networks. However, doing so naively leads to ill posed learning problems with degenerate solutions.

In this research work [9] they propose a novel and principled learning formulation that addresses these issues. The method is obtained by maximizing the information between labels and input data indices. They show that this criterion extends standard cross entropy minimization to an optimal transport problem, which they solve efficiently for millions of input images and thousands of labels using a fast variant of the Sinkhorn-Knopp algorithm. The resulting method is able to self-label visual data so as to train highly competitive image representations without manual labels. Learning visual representations with self-supervised learning has become popular in computer vision. The idea is to design auxiliary tasks where labels are free to obtain. Most of these tasks end up providing data to learn specific kinds of invariance useful for recognition.

In this research work [10] they propose to exploit 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. Self-supervised learning aims to learn representations from the data itself without explicit manual supervision. Existing efforts ignore a crucial aspect of self-supervised learning the ability to scale to large amount of data because self-supervision requires no manual labels.

In this research work [11] revisit this principle and scale two popular self- supervised approaches to 100 million images. They show that by scaling on various axes (including data size and problem „hardness‟), one can largely match or even exceed the performance of supervised pre-training on a variety of tasks such as object detection, surface normal estimation (3D) and visual navigation using reinforcement learning. Virtual population generation is an emerging field in data science with numerous applications in healthcare towards the augmentation of clinical research databases with significant lack of population size. However, the impact of data augmentation on the development of AI (artificial intelligence) models to address clinical unmet needs has not yet been investigated.

In this research work [12] they assess whether the aggregation of real with virtual patient data can improve the performance of the existing risk stratification and disease classification models in two rare clinical domains, namely the primary Sjogren‟s Syndrome (pSS) and the hypertrophic cardiomyopathy (HCM), for the first time in the literature. To do so, multivariate approaches, such as, the multivariate normal distribution (MVND), and straightforward ones, such as, the Bayesian networks, the artificial neural networks (ANNs), and the tree ensembles are compared against their performance towards the generation of high-quality virtual data. 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 research work [13] they present Deep-Cluster, a clustering method that jointly learns the parameters of a neural network and the cluster assignments of the resulting features. Deep-Cluster 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. Deep convolutional neural networks have performed remarkably well on many Computer Vision tasks. However, these networks are heavily reliant on big data to avoid over fitting. Over fitting refers to the phenomenon when a network learns a function with very high variance such as to perfectly model the training data. Unfortunately, many application domains do not have access to big data, such as medical image analysis. This survey focuses on Data Augmentation, a data-space solution to the problem of limited data. Data Augmentation encompasses a suite of techniques that enhance the size and quality of training datasets such that better Deep Learning models can be built using them.

The image augmentation algorithms discussed in this survey [14] include geometric transformations, color space augmentations, kernel filters, mixing images, random erasing, feature space augmentation, adversarial training, generative adversarial networks, neural style transfer, and meta-learning. The application of augmentation methods based on GANs are heavily covered in this survey. In addition to augmentation techniques, this paper will briefly discuss other characteristics of Data Augmentation such as test-time augmentation, resolution impact, final dataset size, and curriculum learning. While there has been remarkable progress in the performance of visual recognition algorithms, the state-of-the-art models tend to be exceptionally data-hungry. Large labeled training datasets, expensive and tedious to produce, are required to optimize millions of parameters in deep network models. Lagging behind the growth in model capacity, the available datasets are quickly becoming outdated in terms of size and density.

To circumvent this bottleneck [15], they propose to amplify human effort through a partially automated labeling scheme, leveraging deep learning with humans in the loop. Starting from a large set of candidate images for each category, they iteratively sample a subset, ask people to label them, classify the others with a trained model, split the set into positives, negatives, and unlabeled based on the classification confidence, and then iterate with the unlabeled set. Learning generic and robust feature representations with data from multiple domains for the same problem is of great value, especially for the problems that have multiple datasets but none of them are large enough to provide abundant data variations.

In this research work they [16] present a pipeline for learning deep feature representations from multiple domains with Convolutional Neural Networks (CNNs). When training a CNN with data from all the domains, some neurons learn representations shared across several domains, while some others are effective only for a specific one. Based on this important observation, they propose a Domain Guided Dropout algorithm to improve the feature learning procedure. Experiments show the effectiveness of their pipeline and the proposed algorithm. Their methods on the person reidentification problem outperform stateof-the-art methods on multiple datasets by large margins.

3. CONCLUSIONS AND FUTURE SCOPE

The classification of objects into the right classes has long been one of the most crucial objectives of machine learning or deep learning. Due to the similarities between various things, their textures, colors, and other physical properties, object recognition still presents significant challenges despite the importance of categorizing specific groups of images. In computer vision, object detection has a wide range of uses, such as face and vehicle detection, video surveillance, and plant leaf detection. When conducting studies on flowers, using flowers as medicine, analyzing floral patents, etc., automatic classification of flowers is crucial. Traditionally, lowlevel characteristics like color, shape, texture, and geometry are used to classify flowers. Between floral classes, there are significant intra-class diversity as well as interclass similarities. Because they are dependent on visual search, search engine-based flower identification and categorization methods are not effective and reliable. The feature description has a significant impact on how accurately and robustly flowers are classified. In recent years, deep features have demonstrated good performance on high-resolution photos, but they are unable to extract precise global features from low-resolution images. Deep neural networks have been widely used in computer vision applications because they are effective at identifying picture patterns. An advanced deep-learning model that can precisely calculate the similarity of floral photos is required in the field of flower image analysis. Computer vision and botanical research have significantly advanced with the development of fine-tuned deep-learning models for similarity computation on flower datasets. These proposed models inception V3 and VGG 16 have been applied to tackle the challenge of manually classifying and comparing flower species due to their diversity and complexity. Similarity computation models have revolutionized the field of botanical research by achieving remarkable accuracy and efficiency in flower image analysis. These models, fine-tuned deep learning models, can handle existing datasets and accommodate new flower species and improved models. This research work states that different similarity computation model; proposed model with augmentation gives better results than existing models.

Future research in this field is expected to focus on enhancing robustness, accuracy, interpretability, and scalability. By improving their capacity to recognize minute visual characteristics, these models will enable them to distinguish precisely between closely related flower species. Additionally, they will be made clearer to promote trust and comprehension of the underlying botanical traits that affect the model's conclusions. A comprehensive understanding of flowers will be given through the integration of multimodal data sources, such as textual descriptions, fragrance profiles, and genetic data, enhancing similarity computations.

REFERENCES

[1] Sarah K. Alhabeeb, Amal A. Al-Shargabi, “Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction”, IEEE Access, 2024, pp. 24412-24427.

[2] Wenbin Yang, Xueluan Gong, “SwiftTheft: A Time-Efficient Model Extraction Attack Framework Against Cloud-Based Deep Neural Networks”, Chinese Journal of Electronics, vol. 33, 2024, pp. 90-100.

[3] Toan Khac Nguyen, Minh Dang, “Utilizing Deep Neural Networks for Chrysanthemum Leaf and Flower Feature Recognition”, AgriEngineering 2024, pp. 1133–1149.

[4] Dipanjali Kundu, Mahbubur Rahman, “Federated Deep Learning for Monkeypox Disease Detection on GAN-Augmented Dataset”, IEEE Access, 2024, pp. 32819-32830.

[5] Giriraj Gautama, Anita Khanna, “Content Based Image Retrieval System Using CNN based Deep Learning Models”, International Conference on Machine Learning and Data Engineering, 2023, pp. 3131-3141.

[6] Q. X. Zhang, W. C. Ma, Y. J. Wang, et al., “Backdoor Attacks on Image Classification Models in Deep Neural Networks,” Chinese Journal of Electronics, 2022, pp. 199-212.

[7] S. Hong and J. Chae, “Active learning with multiple kernels,” IEEE Transactions on Neural Networks and Learning Systems, 2022, pp. 2980–2994.

[8] Vasileios C. Pezoulas, Grigoris I. Grigoriadis, George Gkois, Nikolaos S. Tachos, Tim Smole, “A computational pipeline for data augmentation towards the improvement of disease classification and risk stratification models: A case study in two clinical domains”, Computers in Biology and Medicine, 2021.

[9] Jun Sun and Qiao Sun, “Bearing Prognostics: An Instance-Based Learning Approach with Feature Engineering, Data Augmentation, and Similarity Evaluation”, Signals 2021, pp 662–687.

[10] Thananya Phreeraphattanakarn, Boonserm Kijsirikul, “Text data-augmentation using Text Similarity with Manhattan Siamese long short-term memory for Thai Language”, 2021, pp. 1-7.

[11] Luca Bertinetto, Jack Valmadre, Joao F. Henriques, Andrea Vedaldi, Philip H. S. Tor, “Fully-Convolutional Siamese Networks for Object Tracking”, 2021, pp. 1-16.

[12] Mathilde Caron, Ishan Misra, Julien Mairal, “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments”, 2021, pp. 1-23.

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

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie,Ross Girshick, “Momentum Contrast for Unsupervised Visual Representation Learning”, IEEE 2020, pp. 1-12.

[15] Olivier J. Henaff, Aravind Srinivas, 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.

[16] Mathilde Caron, Piotr Bojanowski, Armand Joulin, Matthijs Douze, “Deep Clustering for Unsupervised Learning of Visual Features”, 2019, pp. 1-30.

[17] Mathilde Caron, Ishan Misra, Julien Mairal, 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.

[18] 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.

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

[20] Mathilde Caron, Piotr Bojanowski, Julien Mairal, Armand Joulin, “Unsupervised Pre- Training of Image Features on Non-Curated Data”, 2019, pp. 1-14.