Brief Report

Momentum Contrast for Unsupervised Visual Representation Learning

Ebou A Sowe *

Posted Date: 9 January 2025

doi: 10.20944/preprints202501.0668.v1

Keywords: Unsupervised learning; Contrastive learning; Visual Representation

Brief Report

Momentum Contrast for Unsupervised Visual Representation Learning

Ebou A Sowe

School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, Hubei, 430070, China; eboua.sowe@whut.edu.cn /ebousowe50@gmail.com

Abstract: This brief report presents a novel unsupervised learning representation learning method called momentum contrast. Momentum contrast uses a contrastive learning technique to learn representations by comparing features of related yet dissimilar images for efficient feature extraction and unsupervised representation learning. Similar images are grouped together, and dissimilar images are placed far apart. The method builds upon previous works in contrastive learning but includes a momentum optimisation step to improve representation learning performance and generate better quality representations. Experiments on various datasets demonstrate that momentum contrast is able to learn high-quality representations, allowing us to directly use them to achieve competitive performance with fewer labelled examples.

Keywords: Unsupervised learning; Contrastive learning; Visual Representation

Table of Contents

Abstract ..... *Error! Bookmark not defined.*

1. Introduction 2

2. Background Study 2

3. Methodology 2

3.1 Contrastive Learning as Dictionary Look-up 2

3.3 Approach 3

3.4 Architecture 4

3.4.1 Dictionary as a Queue 4

3.4.2 Momentum Update 5

4. Performance 5

4.1. Experiment results 5

4.1.1 ImageNet Classification 6

4.1.2 Transfer Learning 6

4.1.3 Few-Shot Learning 6

4.1.4 Robustness to Adversarial Attacks 6

4.1.5 Generalization 7

5. Conclusion 7

5.1 Improved Self-supervised Learning 7

 doi:10.20944/preprints202501.0668.v1

5.2 Used Contrastive Learning 7

5.3 Transferability 8

5.4 Robustness to Label Noise 8

5.5 Promising Applications 8

References 9

1. Introduction

Visual representation learning is a crucial component of many computer vision applications. In recent years, there has been a growing interest in unsupervised methods for learning visual representations. Unsupervised methods do not require labelled data, making them more versatile and applicable to a wider range of tasks. One unsupervised method that has gained popularity in recent years is Momentum Contrast (MoCo) for unsupervised visual representation learning. MoCo is a mechanism for building dynamic dictionaries for contrastive learning and can be used with various pretext tasks[1]. In this report, I will explore the concept of momentum contrast for unsupervised visual representation learning and its performance in comparison to other unsupervised learning methods. I will also explore some experimental results of MoCo, based on these experiments, I will drive some conclusions.

2. Background Study

Traditionally, supervised learning has been the dominant paradigm for training deep neural networks for visual recognition tasks. Contrastive learning (CL) is one of the prominent keystones of self-supervised learning. It fosters discriminability in the representation [8], [9]. [10],[11], [12]. However, there are several limitations to this approach. Firstly, obtaining large labelled datasets can be expensive and time-consuming, especially for specialized tasks. Secondly, even with large labelled datasets, the resulting models may not generalize well to new, unseen images. Finally, supervised learning is not applicable to many domains where labelled data is scarce or unavailable. Unsupervised visual representation learning involves training a neural network to learn a set of features that can be used to represent images in a meaningful way. These features can then be used for a variety of tasks, such as image classification, object detection, and semantic segmentation. Unsupervised learning methods typically rely on data augmentation techniques to generate a large amount of diverse training data. The goal is to train a neural network to learn a set of features that are invariant to these augmentations. The core idea is to pull representations of "similar" images (referred to as positives) close while "dissimilar" images (negatives) are contrasted in feature space. Such methods implemented this idea using an instance discrimination pretext task where only transformed versions of the same images are taken as positives while augmented versions of other images are negatives[3]

3. Methodology

3.1. Contrastive Learning as Dictionary Look-up

Contrastive learning is a machine learning technique that aims to learn useful representations by contrasting pairs of examples. Contrastive learning can drive a variety of pretext tasks[1].

Even though contrastive learning has become prominent in recent years due to the success of large pre-trained models in the fields of natural language processing (NLP) and computer vision (CV), the seminal idea dates back at least to the 1990s [4],[5]. MoCo uses contrastive learning technique by making the dynamic dictionary large and consistence. Among the most successful of

 doi:10.20944/preprints202501.0668.v1

the recent self-supervised approaches to learning visual representations, a subset of these termed ''contrastive'' learning methods have achieved the most success[2].

The negative samples used for contrastive learning are obtained from a dynamic dictionary. Initially, the dictionary is empty. As the model learns, the encoder representations of the input data are stored in the dictionary. The dictionary is maintained using a queue-based mechanism, where new representations replace the oldest ones. This way, the dictionary captures a wide range of negative samples over time, providing diverse negative pairs for contrastive learning.

By continuously updating the dictionary of negative samples and training the encoder using the contrastive loss, MoCo encourages the model to capture useful and semantically meaningful representations. The dynamic nature of the dictionary allows the model to adapt to changing data distributions and learn robust representations.

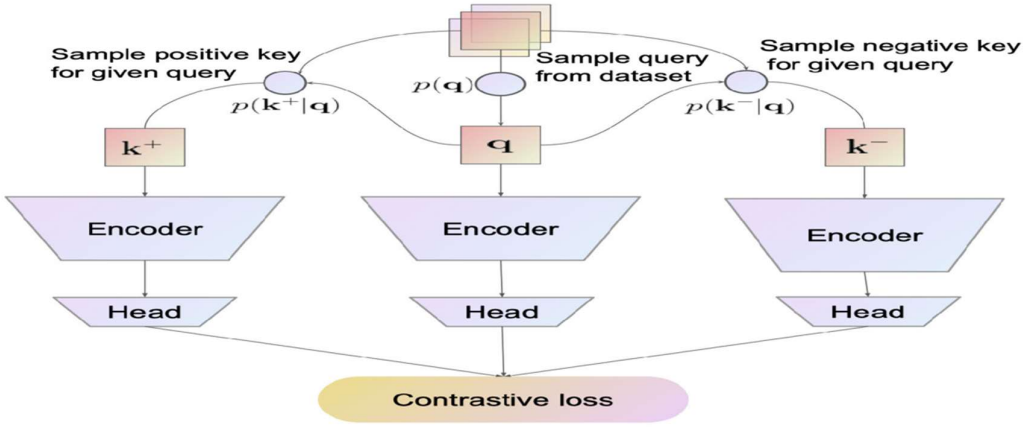

Figure 1. Overview of the Contrastive Representation Learning framework. Its components are: a similarity and dissimilarity distribution to sample positive and negative keys for a query, one or more encoders and transform heads for each data modality and a contrastive loss function evaluate a batch of positive and negative pairs[2].

3.2. Momentum Contrast

Momentum Contrast (MoCo) is an unsupervised learning method that was introduced in a paper by He et al. in 2019. The method is based on the idea of using a momentum encoder to generate a set of target features. During training, the momentum encoder is updated using a moving average of the weights of the online encoder. The online encoder is trained to generate features that are similar to the target features.

The MoCo method has several advantages over other unsupervised learning methods. One advantage is that it is computationally efficient, allowing for larger batch sizes and longer training times. Another advantage is that it is more effective at learning representations that are invariant to data augmentations. This is achieved by using a larger set of augmentations during training. MoCo can outperform its supervised pre-training counterpart in 7 detection/segmentation tasks on PASCAL VOC, COCO, and other datasets, some- times surpassing it by large margins[1].

MoCo v1 has attracted significant attention by demonstrating superior performance over supervised pre-training counterparts in downstream tasks while making use of large negative samples, decoupling the need for batch size by introducing a dynamic dictionary[6].

3.3. Approach

 doi:10.20944/preprints202501.0668.v1

The Momentum Contrast (MoCo) approach is a recent development in unsupervised representation learning that has shown state-of-the-art performance on a variety of visual recognition tasks. The MoCo approach is based on the principle of contrastive learning, which has been shown to be effective for unsupervised representation learning. The basic idea of contrastive learning is to learn representations that are invariant to certain transformations (e.g., rotations, translations, etc.) while maintaining discriminative power for similar images.

The MoCo approach builds on the contrastive learning principle by introducing a momentum-based update rule that improves the stability and convergence of the training process. Specifically, the MoCo approach uses a memory bank to store a large number of negative examples that are used to compute contrastive losses during training. The memory bank is updated using a momentum-based update rule that averages the parameters of the current model with those of a slowly-updated "queue" model. This update rule helps to stabilize the training process by providing a more consistent source of negative examples.

3.4. Architecture

The MoCo architecture consists of two main components: an encoder network and a memory bank. The encoder network is a deep convolution neural network(CNN) that is trained to extract features from raw image data. The memory bank is a large matrix that stores a set of negative examples that are used to compute contrastive losses during training. The memory bank is updated using a momentum-based update rule that averages the parameters of the current model with those of a slowly-updated "queue" model.

The encoder network consists of a series of convolutional layers followed by a global average pooling layer and a fully connected layer. The output of the fully connected layer is a vector of fixed length that represents the image features. The encoder network is trained using a contrastive loss function that encourages similar images to have similar feature representations, while dissimilar images have different feature representations.

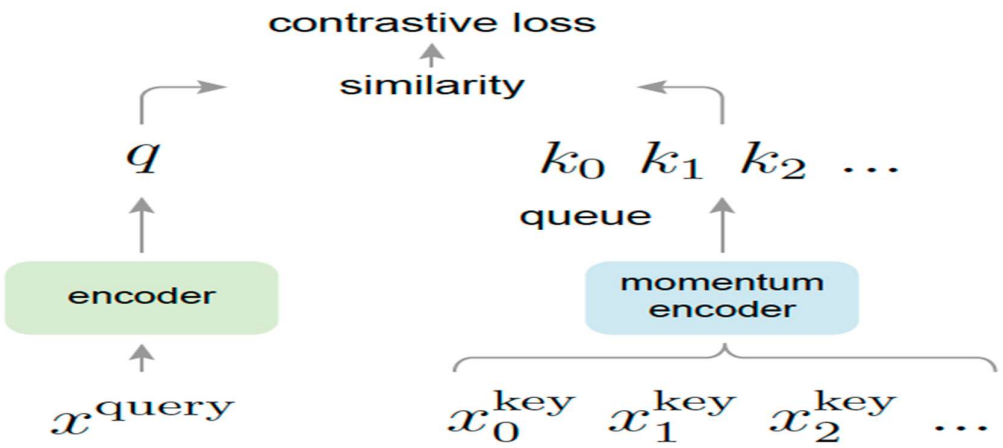

Figure 2. Momentum Contrast (MoCo) trains a visual representation encoder by matching an encoded query q to a dictionary of encoded keys using a contrastive loss[1]. The dictionary keys {k0,k1,k2,...} are defined on-the-fly by a set of data samples[1].

3.4.1. Dictionary as a Queue

Momentum contrast uses a dictionary queue encoded with keys, The queue is updated by adding the representation of the current image to the queue and removing the oldest representation. The dictionary acts as a "memory bank" that stores a history of feature representations. The queue is

 doi:10.20944/preprints202501.0668.v1

updated using a momentum-based update rule, which allows the model to maintain a smooth and stable representation of the feature space.

During training, the images are split into two groups: *a query group* and *a key group*. The query group is used to compute a query feature representation, while the key group is used to compute a set of key feature representations. The query feature representation is then compared to the key feature representations stored in the dictionary using a contrastive loss function.

3.4.2. Momentum Update

Momentum update is a key component in Momentum Contrastive learning, a technique commonly used in self-supervised learning tasks such as image or video representation learning. It helps improve the stability and convergence speed of the learning process by introducing a momentum term during the update of the model's parameters.

In MoCo, the momentum update is used to update the model's parameters based on the current gradient and the momentum term. The momentum update can be visualized as follows:

Initialize the model's parameters and momentum parameters.

At each training iteration:

a). Compute the gradients of the loss function with respect to the parameters using the current mini-batch of data.

b). Update the momentum parameters using the momentum update equation: $\mathbf{v_t} = \boldsymbol{\alpha} * \mathbf{v_{\{t-1\}}} + (\mathbf{1} - \boldsymbol{\alpha}) * \mathbf{g_t}$,where $\mathbf{v_t}$is the velocity term at time step $\mathbf{t}$, $\boldsymbol{\alpha}$ is the momentum coefficient, and g_t is the gradient at time step t.

c). Update the model's parameters using the momentum parameters: $\boldsymbol{\theta}_{\mathbf{\{t+1\}}} = \boldsymbol{\theta}_{\mathbf{t}} + \mathbf{v_t}$, where $\boldsymbol{\theta}_{\mathbf{\{t+1\}}}$ represents the updated parameters at time step **t+1.**

The momentum update equation calculates the velocity term by combining the previous velocity $\mathbf{v_{\{t-1\}}}$ and the current gradient g_t. The momentum coefficient α determines the contribution of the previous velocity compared to the current gradient. A higher α value gives more weight to the previous velocity, resulting in a smoother and more stable update trajectory.

The momentum update allows the model to accumulate information from previous gradients and helps the optimization process by maintaining a consistent direction of updates. This can help the model escape shallow local minima and converge faster to better representations.

Keep in mind that while the momentum update is an essential component of MoCo, the specific implementation details and hyperparameters may vary depending on the exact architecture and training setup.

4. Performance

MoCo has been shown to outperform other unsupervised learning methods on several benchmark datasets, including ImageNet, CIFAR-10, and CIFAR-100. MoCo achieves state-of-the-art performance on these datasets without using any labelled data. MoCo has also been shown to be effective at learning representations for downstream tasks such as object detection and semantic segmentation.

4.1. Experiment results

Momentum Contrast (MoCo) has shown impressive results in various experimental settings and benchmark datasets. MoCo's success can be attributed to its innovative contrastive learning framework, which encourages the model to learn discriminative representations by contrasting positive and negative samples. Positive samples in MoCo are augmented versions of the same image, while negative samples are drawn from a queue. This learning approach enables the model to pull similar samples closer together in the learned representation space while pushing dissimilar samples apart.

 doi:10.20944/preprints202501.0668.v1

Here are some notable experimental results achieved by MoCo:

4.1.1. ImageNet Classification

MoCo achieved state-of-the-art performance on the ImageNet-1K dataset, which consists of 1.28 million labelled images spanning 1,000 object categories. In the MoCo v2 paper, the authors reported top-1 accuracy of 60.6% using ResNet-50, surpassing previous self-supervised methods and approaching the performance of supervised methods.

In the MoCo v2 paper, the authors reported impressive results using the MoCo framework with the ResNet-50 architecture. They achieved a top-1 accuracy of 60.6% on the ImageNet-1K dataset. This performance surpassed previous self-supervised methods and approached the performance of supervised methods, which rely on human-labelled data.

This achievement is significant because it demonstrates the effectiveness of self-supervised learning approaches like MoCo in learning high-quality representations from large amounts of unlabelled data. By leveraging the power of contrastive learning and momentum encoders, MoCo was able to capture meaningful visual features that improved image classification accuracy.

The ability of MoCo to achieve competitive results on the challenging ImageNet-1K dataset indicates that self-supervised learning has the potential to bridge the gap between supervised and unsupervised methods. It opens up possibilities for utilizing vast amounts of unannotated data to learn representations that approach the performance of supervised models, reducing the reliance on human-labelled data.

These advancements in self-supervised learning and the success of MoCo on ImageNet-1K have contributed to the growing interest and exploration of self-supervised methods in various computer vision tasks and domains.

4.1.2. Transfer Learning

MoCo has demonstrated strong transfer learning capabilities. Pretrained models using MoCo representations have been successfully transferred to various downstream tasks such as object detection, semantic segmentation, and instance segmentation. By initializing the models with MoCo pretrained weights, significant performance gains have been observed compared to training from scratch.

4.1.3. Few-Shot Learning

MoCo has also shown promise in few-shot learning scenarios, where the goal is to recognize novel classes with limited labelled examples. By leveraging the learned representations, MoCo has been used as a feature extractor to achieve competitive performance on few-shot learning benchmarks like miniImageNet and tieredImageNet.

4.1.4. Robustness to Adversarial Attacks

MoCo has demonstrated improved robustness to adversarial attacks compared to supervised learning. By training on large-scale unlabelled data, MoCo learns more generalizable representations that are less susceptible to adversarial perturbations.

Adversarial attacks involve making intentional and often imperceptible modifications to input data in order to deceive a machine learning model. These perturbations can lead to incorrect predictions or misclassification. Adversarial attacks are a significant concern in various domains, including computer vision.

By training on large amounts of unlabelled data, MoCo learns to capture underlying patterns and structures in the data that are more resilient to adversarial perturbations. The robustness stems from the model's ability to generalize across a diverse set of samples and learn more invariant representations. This generalizability helps in reducing the vulnerability to adversarial attacks.

 doi:10.20944/preprints202501.0668.v1

Furthermore, the contrastive learning framework of MoCo, where positive samples are augmented versions of the same image and negative samples are drawn from a queue, encourages the model to pull similar samples closer together while pushing dissimilar samples apart. This contrastive objective promotes the learning of discriminative features that are less susceptible to adversarial perturbations.

4.1.5. Generalization

MoCo has been shown to generalize well across different domains and datasets. For example, models pretrained on ImageNet using MoCo have been successfully transferred to domain-specific datasets such as Pascal VOC and COCO, achieving competitive performance.

These results highlight the effectiveness of MoCo in learning meaningful and transferable image representations without relying on explicit labels, thereby enabling broader applications and reducing the need for large amounts of labelled data.

5. Conclusion

Momentum Contrast is an effective unsupervised learning method for visual representation learning. It has several advantages over other unsupervised learning methods, including computational efficiency and better invariance to data augmentations. MoCo has been shown to achieve state-of-the-art performance on several benchmark datasets, making it a promising approach for unsupervised visual representation learning. Based on this, I drive the following conclusions:

5.1. Improved Self-supervised Learning

MoCo has shown significant improvements over traditional self-supervised learning methods. By leveraging a momentum encoder, MoCo creates a dynamic and consistent queue of negative samples, enabling better learning of representations without the need for manual annotations. It addresses the limitations of traditional self-supervised learning approaches by introducing a momentum encoder and a dynamic queue of negative samples.

Overall, MoCo's use of a momentum encoder and a dynamic queue of negative samples has led to significant improvements in self-supervised learning. By leveraging these techniques, MoCo has demonstrated state-of-the-art performance on various benchmark datasets, surpassing previous methods that relied on manual annotations or supervised learning.

5.2. Used Contrastive Learning

MoCo utilizes a contrastive learning framework, where positive samples are augmented versions of the same image and negative samples are drawn from the queue. This encourages the model to pull positive samples together while pushing away negative samples, leading to more discriminative representations.

he main idea behind MoCo is to encourage the model to pull positive samples (augmented versions of the same image) closer together in the embedding space while pushing negative samples (drawn from a queue of other images) further apart. This helps in learning more discriminative and meaningful representations.

The contrastive learning process in MoCo can be summarized in the following steps:

Online Encoder and Target Encoder: MoCo maintains two encoders, the online encoder and the target encoder. The target encoder represents a slowly moving average of the online encoder's weights, which provides a consistent and stable set of representations.

Positive Pair Augmentation: To create positive pairs, an image is randomly augmented multiple times. These augmented versions of the same image serve as positive samples. By applying different augmentations, the model learns to capture different views of the same underlying object or scene.

Negative Sample Selection: Negative samples are drawn from a queue that stores representations of other images in the dataset. The queue acts as a source of negative samples that

 doi:10.20944/preprints202501.0668.v1

the model should be pushed away from. This helps in learning more robust and discriminative representations.

Contrastive Loss: MoCo uses a contrastive loss function to train the model. The contrastive loss encourages the model to maximize the similarity between positive pairs while minimizing the similarity between positive and negative pairs. This loss formulation drives the model to learn representations that are more discriminative and generalize well to downstream tasks.

By training with the contrastive learning framework of MoCo, the model can learn powerful representations from unlabeled data, which can then be fine-tuned or transferred to supervised tasks with limited labeled data, leading to improved performance.

5.3. Transferability

MoCo has demonstrated excellent transferability of learned representations. Pretrained models using MoCo have been shown to achieve state-of-the-art performance on various downstream tasks such as image classification, object detection, and semantic segmentation. This indicates that the learned representations capture general visual concepts that can be transferred across different tasks.

The pretrained models from MoCo provide a strong starting point for fine-tuning on specific supervised tasks with limited labelled data. By leveraging the knowledge acquired during the unsupervised pretraining phase, these models can effectively generalize and adapt to new tasks. This transfer learning approach saves significant computational resources and reduces the need for extensive labelled data.

The success of MoCo and similar self-supervised learning methods highlights the potential of unsupervised learning in capturing meaningful representations that benefit a wide range of downstream applications.

5.4. Robustness to Label Noise

MoCo's self-supervised nature makes it robust to label noise in the training data. Since it does not rely on human annotations, MoCo can learn from large amounts of unlabelled data, which is often easier to obtain compared to accurately labelled data. This is particularly advantageous in scenarios where labelled data is scarce or expensive.

Label noise refers to errors or inconsistencies in the annotations of the training data. In traditional supervised learning, these errors can negatively impact the model's performance as it learns from mislabelled examples. However, self-supervised learning methods like MoCo bypass the need for explicit labels by utilizing pretext tasks that create supervision signals from the data itself.

By leveraging unlabelled data, MoCo can learn powerful representations that are robust to label noise. The model learns to capture inherent patterns and structure in the data, allowing it to generalize well even in the presence of noisy or imperfect labels. This is particularly valuable in real-world scenarios where obtaining accurately labelled data can be challenging, such as in large-scale datasets or domains where expert annotations are scarce.

Furthermore, the ability of MoCo to learn from unlabelled data makes it highly scalable. It can leverage vast amounts of readily available unlabelled data, such as images or text corpora, enabling the model to learn rich and meaningful representations without the need for manual annotations. This scalability makes MoCo an attractive approach in situations where obtaining labelled data is limited or costly.

Overall, MoCo's self-supervised learning paradigm empowers the model to learn robust representations from unlabelled data, making it particularly advantageous in scenarios with label noise, scarce labelled data, or when access to accurately labelled data is challenging.

5.5. Promising Applications

The effectiveness of MoCo in visual representation learning opens up opportunities for various applications. It can be applied to domains such as computer vision, robotics, and autonomous

 doi:10.20944/preprints202501.0668.v1

systems, where understanding visual information is crucial for perception, decision-making, and action.

In conclusion, Momentum Contrast (MoCo) has emerged as a powerful technique for visual representation learning. Its ability to learn from large-scale unlabeled data, robustness to label noise, and excellent transferability make it a valuable tool for advancing computer vision research and applications.

References

1. K. He, H. Fan, Y. Wu, S. Xie and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 9726-9735, doi: 10.1109/CVPR42600.2020.00975.
2. P. H. Le-Khac, G. Healy and A. F. Smeaton, "Contrastive Representation Learning: A Framework and Review," in *IEEE Access*, vol. 8, pp. 193907-193934, 2020, doi: 10.1109/ACCESS.2020.3031549
3. Y. Zhang, X. Hu, N. Sapkota, Y. Shi and D. Z. Chen, "Unsupervised Feature Clustering Improves Contrastive Representation Learning for Medical Image Segmentation," *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Las Vegas, NV, USA, 2022, pp. 1820-1823, doi: 10.1109/BIBM55620.2022.9995129.
4. S. Becker and G. E. Hinton, ''Self-organizing neural network that dis- covers surfaces in random-dot stereograms,'' *Nature*, vol. 355, no. 6356, pp. 161–163, Jan. 1992, doi: 10.1038/355161a0.
5. J. Bromley, I. Guyon, Y. LeCun, E. Sa ̈ckinger, and R. Shah, ''Signature verification using a 'siamese' time delay neural network,'' *in Proc. Adv. Neural Inf. Process. Syst*., Feb. 1993, pp. 737–744.
6. T. Nguyen, T. X. Pham, C. Zhang, T. M. Luu, T. Vu and C. D. Yoo, "DimCL: Dimensional Contrastive Learning for Improving Self-Supervised Learning," in *IEEE Access*, vol. 11, pp. 21534-21545, 2023, doi: 10.1109/ACCESS.2023.3236087.
7. X. Chen, H. Fan, R. Girshick and K. He, "Improved baselines with momentum contrastive learning", *arXiv:2003.04297*, 2020.
8. M. Federici, A. Dutta, P. Forré, N. Kushman and Z. Akata, "Learning robust representations via multi-view information bottleneck", *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.
9. I. Misra, C. L. Zitnick and M. Hebert, "Shuffle and learn: Unsupervised learning using temporal order verification", *Proc. Eur. Conf. Comput. Vis.*, pp. 527-544, 2016.
10. F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 815-823, Jun. 2015.
11. K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective", *Proc. Adv. Neural Inf. Process. Syst.*, pp. 1857-1865, 2016.
12. X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos", *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 2794-2802, Dec. 2015.