Journal of Electronic Research and Application, 2021, Volume 5, Issue 3
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Research on Self-Supervised Comparative Learning for Computer Vision

Yuanyuan Liu[1]*, Qianqian Liu[2]

[1]Heilongjiang University of Science and Technology, Harbin 150000, Heilongjiang Province, China

[2]Shengli Oilfield Central Hospital, Dongying 257000, Shandong Province, China

Corresponding author: Yuanyuan Liu, lyyziyi@163.com

Abstract: In recent years, self-supervised learning which does not require a large number of manual labels generate supervised signals through the data itself to attain the characterization learning of samples. Self-supervised learning solves the problem of learning semantic features from unlabeled data, and realizes pre-training of models in large data sets. Its significant advantages have been extensively studied by scholars in recent years. There are usually three types of self-supervised learning: "Generative, Contrastive, and Generative-Contrastive." The model of the comparative learning method is relatively simple, and the performance of the current downstream task is comparable to that of the supervised learning method. Therefore, we propose a conceptual analysis framework: data augmentation pipeline, architectures, pretext tasks, comparison methods, semi-supervised fine-tuning. Based on this conceptual framework, we qualitatively analyze the existing comparative self-supervised learning methods for computer vision, and then further analyze its performance at different stages, and finally summarize the research status of self-supervised comparative learning methods in other fields.

Keywords: Self-supervised learning; Comparative learning; Conceptual analysis framework; Computer vision field; Performance analysis

Publication date: May 2021; ***Online publication:*** May 31, 2021

1. Introduction

Deep neural networks have the ability to learn rich patterns from a large number of data, and are widely used in most computer vision supervision tasks, such as image classification [1-3], semantic segmentation [4-5], natural language Processing [6-8], graph learning [9-11], etc. However, supervised learning relies on millions of labeled data samples and is vulnerable to generalization errors, false associations, and adversarial attacks. Self-supervised learning has received widespread attention due to its data efficiency and generalization ability, and many advanced models are following this paradigm.

Self-supervised learning does not involve manual labeling. It uses an excuse task to mine the supervised signals of the data from large-scale unlabeled data, and applies the learned representation information to downstream tasks. It belongs to a branch of the field of unsupervised learning. At present, self-supervised learning methods are mainly divided into three types: "Generative, Contrastive, and Generative-Contrastive." The details are shown in **Figure 1.** Their main difference lies in the discriminator, potential distribution z, loss function and so on.

For computer vision, the comparative self-supervised learning method is a distinguishing method, which realizes the learning of unlabeled data by grouping similar samples closer together and different samples grouping further away. Comparative self-supervised learning has a simple structure, and its performance is comparable to supervised learning. Therefore, we collect the self-supervised comparative learning methods of visual representation in recent years, and analyze the current methods in detail based

on the conceptual analysis framework we proposed. In short, the main contribution of this article is：

(1) Propose a conceptual analysis framework for the CSL method, realize the hierarchical analysis of the existing technology, and intuitively understand the difference of the CSL method.

(2) Provide a detailed and up-to-date review of self-supervised comparative learning methods for computer vision. People can easily grasp the cutting-edge ideas in this direction.

(3) On this basis, we analyze the comparative analysis of the quantitative performance of the existing technology in the public data set.

(4) Discuss the current self-supervised comparative learning methods for natural language processing and multi-modal learning related technologies, discuss and analyze future development directions, etc.

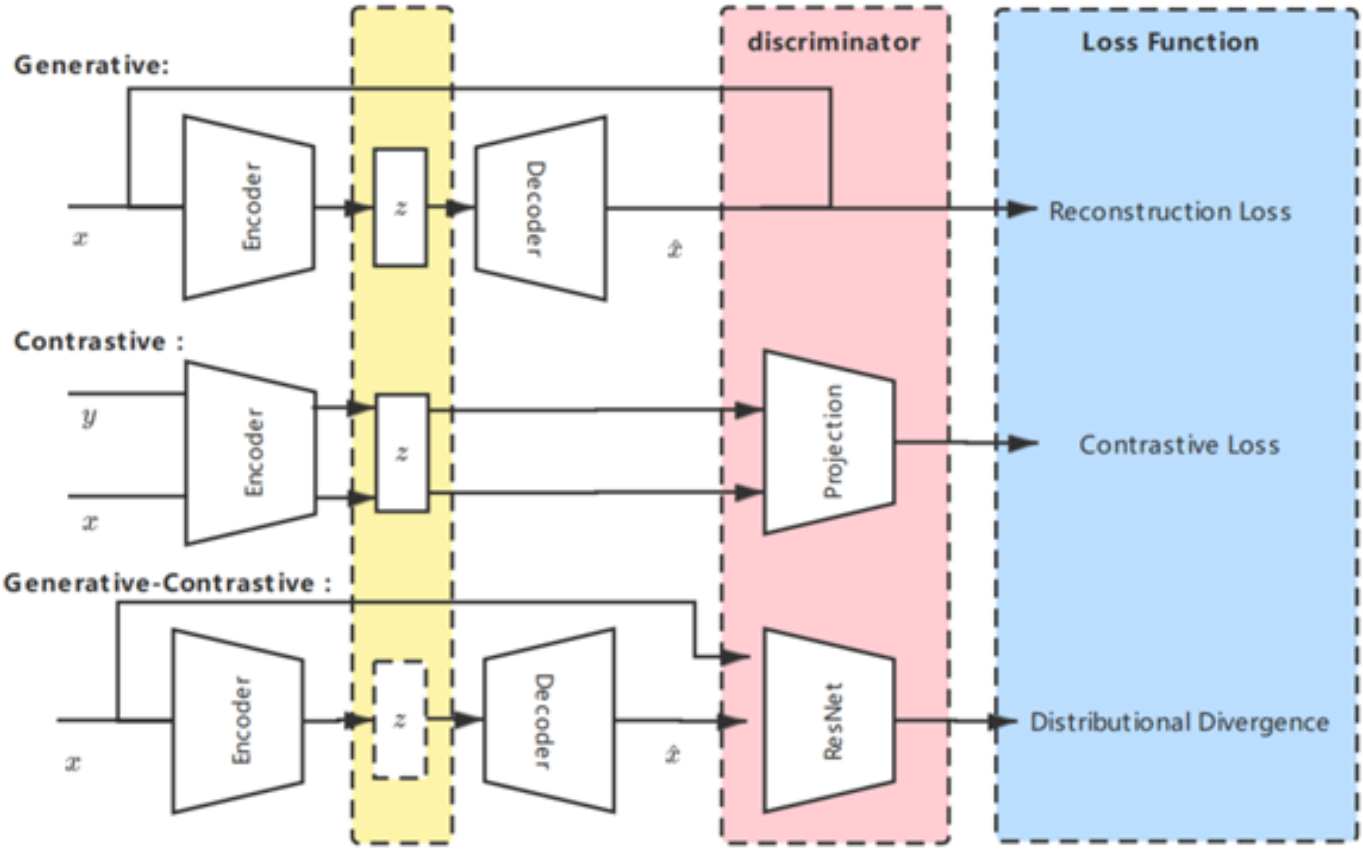

Figure 1. Architecture comparison of Generative, Contrastive, and Generative-Contrastive

2. Contrastive self-supervised learning methods

At present, the comparison of self-supervised learning pipelines is usually shown in **Figure 2.** The augmentation processing of the original sample is a positive sample, while the remaining samples in the batch and data set are considered as negative samples, and the difference learning between positive and negative samples is learned by solving the pretext task. In the process of the pretext task, the model learns the representations in the training set, and then transfers to other downstream tasks after fine-tuning.

This paper is based on the process of comparative learning method and the research of comparative self-supervised learning methods in recent years, as shown in **Table 1.** At the same time, in order to facilitate the qualitative analysis of the differences between different methods, a new conceptual analysis framework is proposed, including five parts: Data augmentation pipeline, architectures, pretext tasks,

comparison methods, semi-supervised fine-tuning.

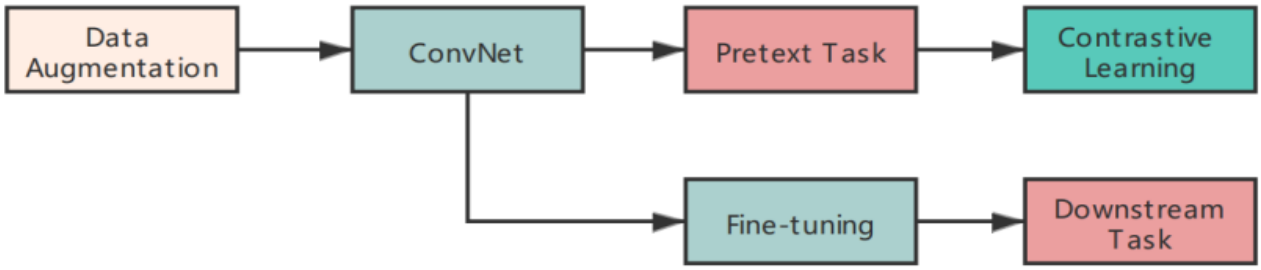

Figure 2. Pipeline based on contrasting self-supervised learning methods

Table 1. Comparison of existing methods.

Method	**Architectures**	**Pretext tasks**	**Comparison**	**NS**	**PS**
RelativePosition [19]	-	Relative postion prediction	Spatial relations (Context-Instance)	-	-
CDJR [20]	End-to-end	Jigsaw + Inpainting + Colorization		×	×
PIRL [21]	Memory bank	Jigsaw		×	√
RotNet [22]	-	Rotation Prediction		-	-
Deep InfoMax [23]	End-to-end	MI Maximization	Belonging (Context-Instance)	×	×
AMDIM [24]	End-to-end			×	√
CPC [25]	End-to-end			×	×
DeepCluster [26]	-	Cluster discrimination	Similarity (Instance-Instance)	-	-
Local Aggregation [27]	-			-	-
ClusterFit [28]	-			-	-
SwAV [29]	Clustering			-	√
SEER [30]	End-to-end			-	√
InstDisc [31]	Memory bank	Instance discrimination	Identity (Instance-Instance)	×	×
CMC [32]	End-to-end			×	√
MoCo [16]	Momentum			×	×
MoCo v2 [33]	Momentum			×	√
SimCLR [17]	End-to-end			×	√
InfoMin [34]	End-to-end			×	√
BYOL [35]	End-to-end			no	√
ReLIC [36]	End-to-end			×	√
SimSiam [37]	End-to-end			no	√

Note: For symbols in “NS” and “PS”: “-” means not applicable, “×” means not adopted, “ √ ” means adopted; “no” particularly means not using negative samples in instance-instance contrast.

2.1. Data augmentation pipeline

The purpose of the data augmentation pipeline is to generate anchor points, positive and negative samples, which maintain the same underlying features as the original samples. In the field of computer vision, there are two common data augmentation methods, one is the data augmentation of image processing technology, and the second is the data augmentation algorithm based on deep learning. SimCLR [17] demonstrated the positive impact of correct data augmentation pipeline on performance.

The augmentation pipeline in AMDIM [24] uses random flipping, image jitter, normalization of mean and standard deviation, etc. The augmentation pipeline is randomly applied twice to generate positive samples, and applied once to negative samples. The author of InfoMin [34] first proposed an unsupervised method of minimizing the mutual information between views to increase the number of positive samples, and combing with a semi-supervised method to find views that only share label information to prevent the loss of predicted label information. This method is about 2% higher than MoCov2. BYOL [35] only uses positive examples, using a random image augmentation channel similar to SimCRL.

2.2. Architecture

The contrastive learning method relies on the calculation of the similarity of negative samples, which can be regarded as a dictionary lookup process. The size of the dictionary is different for different architecture, and it is usually divided into four structures: end-to-end, memory bank, momentum encoder, and clustering. As shown in **Figure 3.**

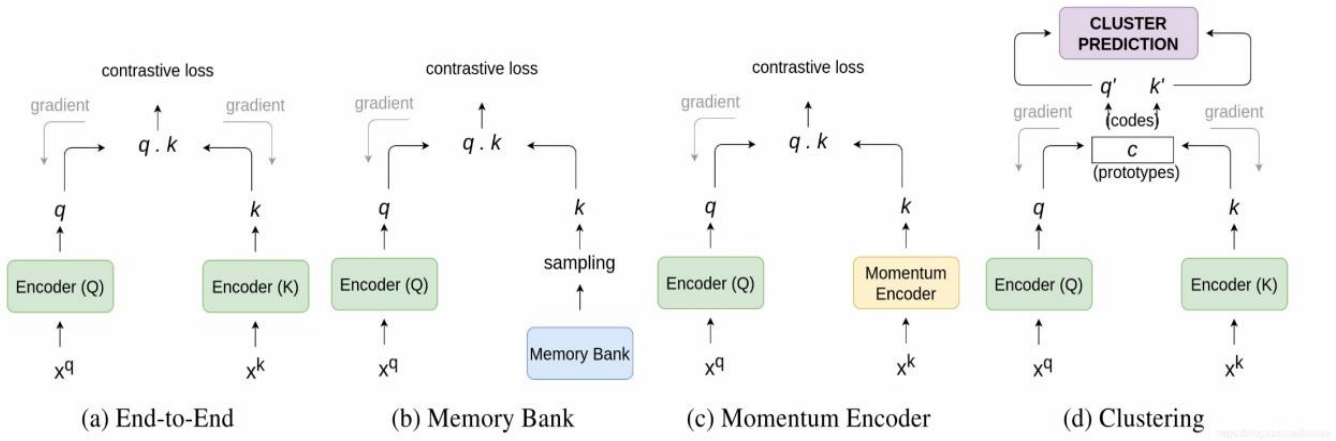

Figure 3. Different architecture in contrastive learning method

The end-to-end is a complex learning system that uses gradient learning, and all modules are differentiable. The original image and its augmentation image are positive samples, and other samples in the same batch are negative samples. Its structure is simple, suitable for large-scale, large-epoch use. The number of negative samples is related to the batch size. However, the batch size is limited by GPU memory, [15] pointed out this structure needs to be optimized in small batches.

SimCLR [17] uses 4096 batches to process 100 epochs. The structure is shown in **Figure 4.**, including image augmentation pipeline, encoder, projection, similarity calculation, and InfoNCE loss function. It adopts the two-mapping structure of encoder and projection, the upper and lower branches are symmetrical, and the two can share parameters. Oord et al. [25] proposed another popular end-to-end architecture. They used a powerful autoregressive model and contrast loss to predict the future of the latent space and learned the feature representation of high-dimensional time series data [20,23,24,30,32,34-37].

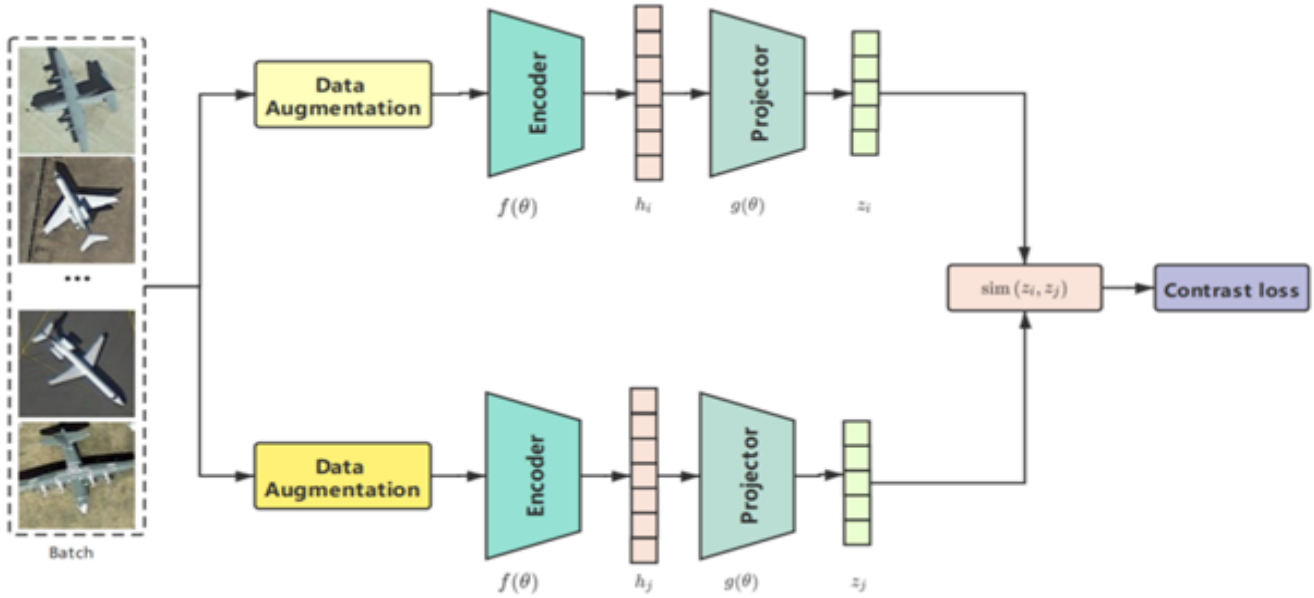

Figure 4. End-to-end in SimCLR

It was proposed to use the repository to save all the features of the image and provide negative samples during training [21]. This structure solves the problem of excessive batch processing, but it consumes a lot of memory, its updates are slow and the complexity of update calculations is high.

The structure of PIRL [21] is shown in **Figure 5.** The storage library M contains the feature representation of each sample, which is convenient to provide sufficient negative samples and provide an intermediate benchmark between the original image and the changed image, and it is updated in a moving average manner [31].

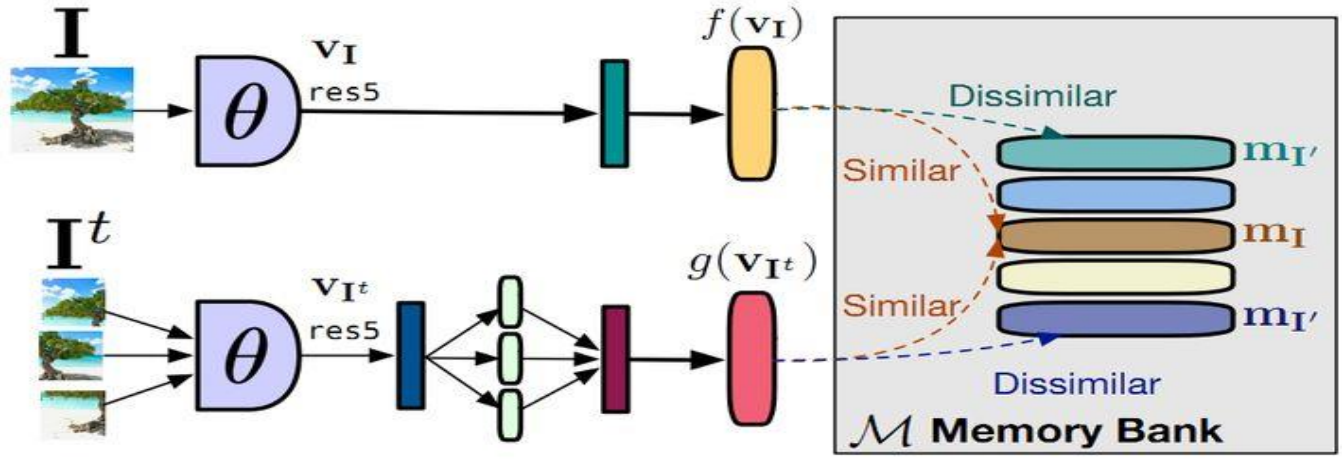

Figure 5. PIRL's memory bank

The momentum encoder generates a dictionary as a queue of encoded keys, and adopts the momentum update method to realize that the dictionary keys are dynamically defined by a batch of data samples during training. This structure further solves the problem of the memory bank.

The structure of MoCo V2 [33] is shown in **Figure 6.**, similar to SimCLR. The parameters of the upper branch model of MoCo V2 are updated by backpropagation gradient, and the parameters of the lower branch

model is updated using **Formula 1.**, and the upper and lower branches do not share parameters.

Formula 1. $$\xi = m\ \xi + (1 - m)\vartheta$$

Among them, ϑ are the model parameters of the upper branch structure, ξ are the parameters of the lower branch structure model, m and are the adjustment parameters of the weight. Usually m will take a larger value. Compared with the upper branch parameter, the lower branch parameter changes slowly and steadily, and iterates from the random value to the optimal value bit by bit [16].

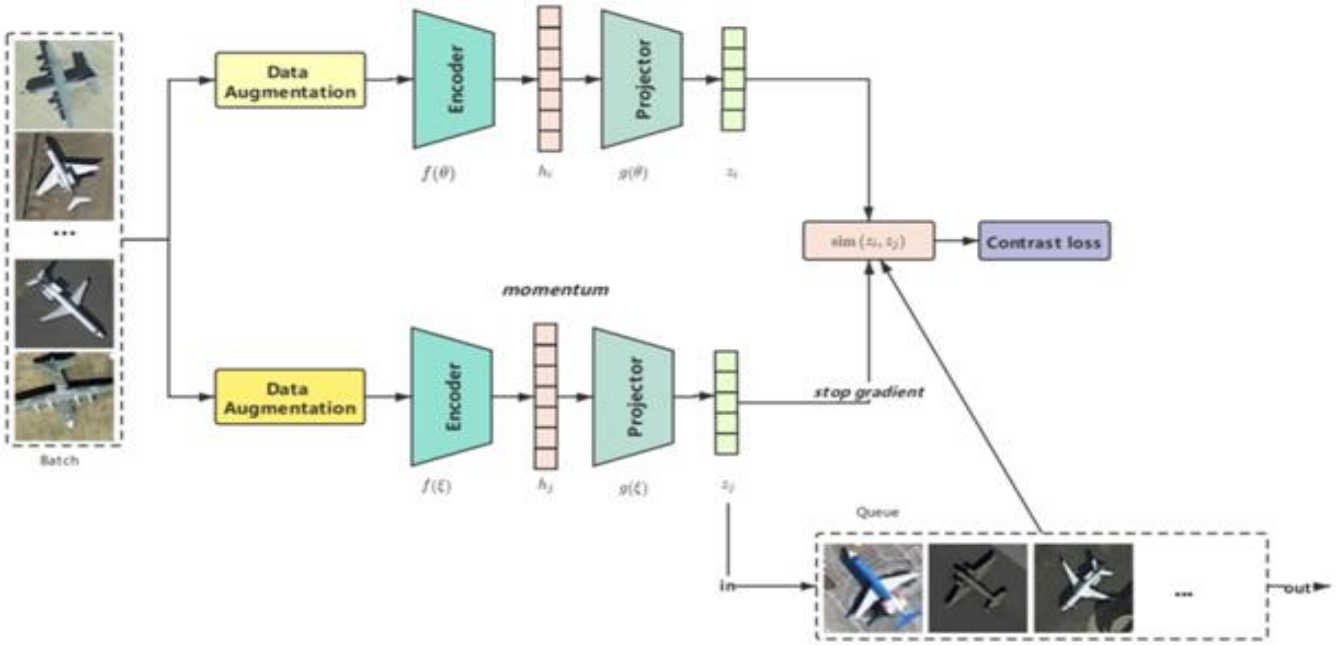

Figure 6. Momentum encoder used by MoCo

The clustering follows an end-to-end approach, where two encoders share parameters and use a clustering algorithm to group similar features together. This structure implicitly solves the problem that in the Instance-based learning method, the comparison of different samples of the same class within the same batch cannot be achieved. The typical method SWAV of this structure, its structure is shown as in **Figure 7.**

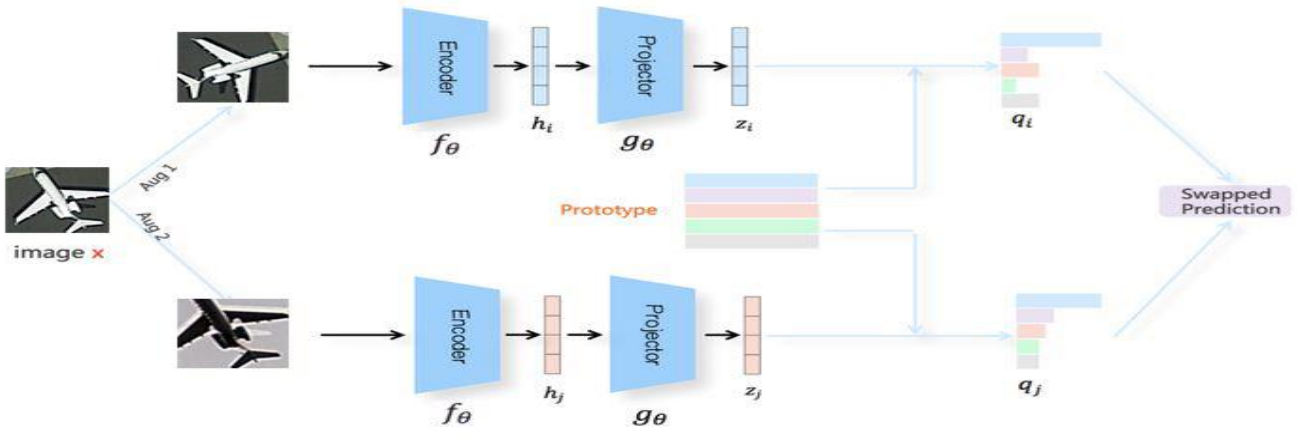

Figure 7. Clustering used by SWAV

2.3. Pretext tasks

The pretext tasks a self-supervised task, which automatically generates pseudo-labels based on the attributes found in the data to achieve characterization learning. Commonly used pretext tasks are mainly divided into four categories: color transformation, geometric transformation, context-based and view prediction (cross modal-based). Currently, cross modal-based commonly used the learning of video representations. Research has shown that it is necessary to choose the appropriate pretext tasks according to the problem to be solved. At the same time, different pretext tasks also affect the way of subsequent characterization extraction and comparison.

Color transformation involves the basic adjustment of image color levels, such as blur, color distortion, grayscale conversion and so on. Geometric transformation is a spatial transformation that modifies the geometric shape of an image without changing its actual pixel information, such as zooming, random cropping, flipping etc. The purpose of those transformation used as pretext tasks is usually to restore, such as relative position prediction, rotation prediction and so on. RotNet [22] is the prediction of the rotation angle based on the transformation of rotation.

For computer vision, context-based often include puzzles, future predictions and so on. PIRL [21] uses puzzle as a pretext task, the original image is the anchor point, the image after the position transformation is the positive sample, and other images in the batch are considered as negative samples [21]. CPC [25] research shows that high-dimensional data is compressed into a compact low-dimensional potential embedding space. A powerful autoregressive model can summarize the information of the potential space and generate a potential context representation. Based on the distributed vector of its composition, the maximum Ground retains the mutual information of the original signal.

2.4. Comparison method

2.4.1. Characterization extraction method

At present, the comparative learning framework is divided into context-Instance comparison and Instance-Instance comparison. Context-instance comparison focuses on the attribution relationship between the local features of the modeling sample and the global context. The Instance-Instance comparison directly studies the relationship between instance-level local representations of different samples.

Context-instance comparison methods can usually be divided into two types: comparison based on spatial relations and comparison based on belonging. Three typical spatial relationship comparison methods: predict relative position [19], rotate [22] and solve puzzles [20,21,48]. The comparison method based on the attribution relationship focuses on the relationship between the local view and the global view of the sample, and solves it with the help of mutual information. Deep InfoMax [23] is the first method to explicitly model mutual information through comparative learning tasks, which promotes the development of self-supervised learning. AMDIM [24] enhances the positive correlation between a local feature and its context. The comparison Instance-Instance method mainly includes two categories: comparison method by clustering and comparison method by identity [26,49-51].

DeepCluster [26] first uses clustering to generate pseudo-labels. The training process is mainly two steps. The first is to cluster through clustering algorithms. Class, each sample generates a pseudo-label, the second is that the discriminator predicts whether two samples are from the same cluster and backpropagates to the encoder. SwAV [29] introduced online clustering ideas and multi-view data augmentation strategies into clustering discrimination methods. The prototype is InstDisc [31] with the comparison method by identity. Later, the research of this method focuses on the selection of positive and negative samples. Based on InstDisc, CMC [32] proposed to take multiple different views of one image as positive samples, and the other image as negative samples. MoCo [16] proposed a way through momentum encoder. SimCLR [17] proposes to add a nonlinear Projection transformation between representation and comparison to achieve the

extraction of lower-level information. InfoMin [34] chooses views with less mutual information to achieve better enhanced views and achieve an increase in positive samples. BYOL [35] abandons negative samples and proposes a new architecture that uses an exponential moving average strategy to update the target encoder.

2.4.2. Similarity measure and loss function

The measure of similarity is to measure the closeness of the embedding between two samples. The commonly used similarity measure is the cosine similarity (as shown in **Formula 2.**), which serves as the basis for different contrast loss functions.

Formula 2.

$$\cos\_ sim(A, B) = \frac{AB}{\|A\| \|B\|}$$

The loss function uses contrast positive-negative samples to express learning ability, and is defined as a combination of positive and negative scores that reflect learning progress. Commonly used loss functions are NCE [18], InfoNCE [52], triplet loss [32].

2.5. Semi-supervised fine-tuning processing

Improve the self-supervised learning model and improve the ability to extract data representations, but in order to transfer to downstream tasks, we need more or less tags. In order to narrow the gap between self-supervised upstream tasks and downstream tasks, semi-supervised learning is usually used.

Analysis based on the contrast self-supervised learning method is beneficial in many downstream vision tasks, but it cannot improve the target detection task in COCO [12]. Studies have found that the improvement of pre-training and self-training contributes to performance from different perspectives [13]. It is found that ResNet-50 uses 10% of the ImageNet label, which can surpass the supervised label of joint pre-training and self-training [14]. Therefore, for self-supervised comparative learning, a three-step framework is proposed:

(1) Do self-supervised pre-training as SimCLR v1, with some minor architecture modification and a deeper ResNet.

(2) Fine-tune the last few layers with only 1% or 10% of original ImageNet labels.

(3) Use the fine-tuned network as teacher to yield labels on unlabeled data to train a smaller student ResNet-50.

The success in combining self-supervised contrastive pre-training and semi-supervised self-training opens up our eyes for a future data-efficient deep learning paradigm. More work is expected for investigating their latent mechanisms.

3. Image representation learning performance analysis

In order to evaluate and compare the performance of self-supervised learning, it is usually analyzed from two aspects: the effectiveness of the pretext task and the specific performance of the downstream task.

Assessing the effectiveness of the pretext task is usually analyzed by kernel visualization, feature map, and nearest neighbor-based methods. For example, attention maps generated from different layers of the encoder can be used to evaluate whether the excuse task is effective, as shown in **Figure 8.**

In the downstream task, image classification, most methods use two commonly data sets of ImageNet and Places to evaluation. In terms of target detection, the Pascal VOC data set is often used to evaluation. The performance of these methods is better than the best supervised models.

As shown in **Table 2.**, without considering the proportion of batch size, epoch and the number of semi-supervised fine-tuning labels, the current method's classification performance on ImageNet and the

performance of target detection on Pascal VOC are roughly counted. At present, the top-1 classification accuracy on the ImageNet dataset based on the comparative self-supervised learning method is comparable to the supervised classification accuracy, and the current top-1 accuracy is stable at 65% on the basic ResNet structure. At the same time, there is a small gap between performance and supervised learning in target detection tasks, and the average detection accuracy is above 80%. However, whether it is in classification tasks or target detection tasks, the gap between comparative learning methods and supervised learning methods is based on a deeper and wider network structure and training with a large number of samples.

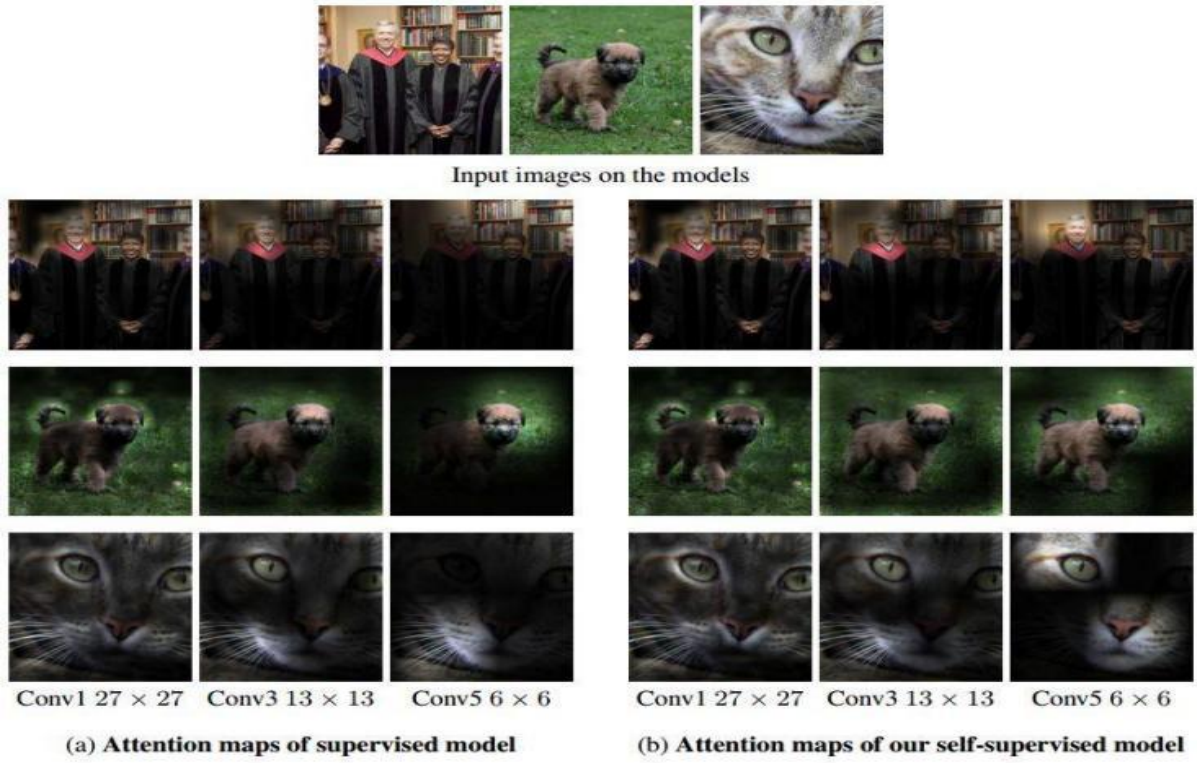

Figure 8. Attention map applied on Conv1 27x27, Conv3 13x13 and Conv5 6x6 features

Table 2. (1) The linear classification accuracy based on frozen features is the highest (top-1); (2) VOC7+12 features based on Faster-CNN fine-tuned target detection

Method	Encoder	Parameter	Top-1	Detection
Supervised	ResNet50	25.6M	53.2	81.3
DeepCluster	AlexNet	61M	37.5	55.4
PIRL	ResNet50	25.6M	49.8	80.7
MoCo	ResNet50	25.6M	60.6	81.4
MoCo V2	ResNet50	25.6M	67.5	-
SwAv	ResNet50	25.6M	56.7	82.6
CPC v2	ResNet50	25.6M	63.8	
CPC v2	ResNet161	305M	71.5	-
SimCLR	ResNet50	25.6M	65.6	84.1
SimCLR	ResNet50(2*)	94M	74.2	-
SimCLR	ResNet50(4*)	375M	76.5	-
BYOL	ResNet50	25.6M	68.8	85.4
BYOL	ResNet200(2*)	250M	77.7	-
SimSiam	ResNet50	25.6M	74.3	82.4

4. Contrastive learning in other areas of research

At present, comparative learning is widely studied in the fields of natural language processing, multi-modality, graph neural network and other fields. And different fields influence each other.

Comparative learning was first introduced into natural language processing, using co-occurring words as semantic similarities, and using negative sampling to learn word embedding [38]. Learn useful feature representations from unlabeled data, and introduce latent classes to formalize semantically similar concepts. This method is comparable to state-of-the-art supervision methods on the Wiki-3029 dataset [39]. Researchers discuss the coherence and encodes the fine-grained sentence ordering in the text [40]. Although it has the same number of parameters as BERT-Base, it is better than the BERT-Large model. At the same time, it shows significant improvement on multiple downstream tasks [41-43].

Contrastive learning in the multi-modal field relies on the alignment of different modal information, and is basically supervised comparative learning. CM-ACC [44] learns the joint representation of audio and vision in video data, similar to MOCO, with an encoder and momentum encoder for each of the audio and visual modalities.CM-ACC adopts a method of dynamically constructing a queue to ensure the amount of information and diversity of negative cases. CLIP [45] and ALIGN [46] focus on text and visual modalities, and adopt a comparative learning mode in which there is only one encoder for each modal, and negative examples are selected in the batch. There is also WenLan [47].

5. Conclusion

This article reviews the development status of comparative learning methods in self-supervised learning for computer vision. The comparative model is lightweight and uses unlabeled data to self-learn to generate supervised signals. At present, the performance on the basic data set is comparable to supervised learning. However, there are still many problems to be solved. For example, based on the problem of sampling efficiency, the theory of the role of negative samples in the method is not clear. SimCLR proves that data augmentation can improve the performance of the method, but the reasons and theories are not yet demonstrated. Evaluation of migration capability based on other data sets, etc.

Disclosure statement

The author declares no conflict of interest.

References

[1] Deng J, Dong W, Socher R, et al., 2009, Imagenet: A Large-Scale Hierarchical Image Database. In CVPR:248–255.IEEE.

[2] He K, Zhang X, et al., 2016, Deep Residual Learning for Image Recognition. In CVPR: 770–778.

[3] Huang G, et al., 2017, Weinberger. Densely Connected Convolutional Networks. 2017 IEEE CVPR: 2261–2269.

[4] Girshick R, et al., 2014, Rich Feature Hier-Archies for Accurate Object Detection and Semantic Segmentation. In CVPR: 580–587.

[5] Long J, et al., 2015, Fully Convolutional Networks for Semantic Segmentation. In CVPR: 3431–3440.

[6] Devlin J, et al., 2015, Pre-Training of Deep Bidirectional Transformers for Language Under-Standing. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (1):4171–4186.

[7] Lan Z, et al., 2019, Albert: A Lite Bert for Self-Supervised Learning of Language Representations. arXiv preprint arXiv:1909.11942.

[8] Asai A, et al., 2019, Learning to Retrieve Reasoning Paths Over Wikipedia Graph for Question Answering. arXiv preprint arXiv:1911.10470.

[9] Hu Z, et al., 2020, Heterogeneous Graph Transformer. arXiv preprint arXiv:2003.01332.

[10] Zhang M, et al., 2018, An End-To-End Deep Learning Architecture for Graph Classification. In AAAI.

[11] Hoffmann J, et al., 2019, Infograph: Unsupervised and Semi-Supervised Graph-Level Representation Learning Via Mutual Information Maximization. arXiv preprint arXiv:1908.01000.

[12] Deng J, 2020, How Useful is Self-Supervised Pretraining for Visual Tasks? In CVPR: 7345–7354.

[13] Zoph B, et al., 2020, Rethinking Pre-Training and Self-Training. arXiv:2006.06882.

[14] Chen T, et al., 2020, Big Self-Supervised Models are Strong Semi-Supervised Learners. arXiv preprint arXiv:2006.10029.

[15] Goyal P, et al., 2017, Accurate, Large Minibatch Sgd: Training Imagenet in 1 Hour.

[16] He K, et al., 2019, Momentum Contrast for Unsupervised Visual Representation Learning.

[17] T. Chen, et al., 2020, A Simple Framework for Contrastive Learning of Visual Representations. arXiv preprint arXiv:2002.05709.

[18] Gutmann M, Hyvärinen A, 2010, Noise-Contrastive Estimation: A New Estimation Principle for Unnormalized Statistical Models. In AISTATS.

[19] Doersch C, Gupta A, et al., 2015, Unsupervised Visual Representation Learning by Context Prediction. In Proceedings of the IEEE ICCV: 1422–1430.

[20] Kim D, et al., 2018, Learning Image Representations by Completing Damaged Jigsaw Puzzles. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV): 793–802.

[21] Misra I, et al., 2019, Self-Supervised Learning of Pretext-Invariant Representations. arXiv preprint arXiv:1912.01991.

[22] Gidaris S, Singh P, et al., 2018, Unsupervised Repre-Sentation Learning by Predicting Image Rotations. arXiv preprint arXiv:1803.07728.

[23] Hjelm R, et al., 2018, Learning Deep Representations by Mutual Information Estimation and Maximization. arXiv preprint arXiv:1808.06670.

[24] Bachman P, Hjelm R, 2019, Learning Represen-Ations by Maximizing Mutual Information Across Views. In NIPS: 15509–15519.

[25] Li Y, Vinyals O, 2018, Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

[26] Caron M, et al., 2018, Deep Clustering for Unsupervised Learning of Visual Features. In Proceedings of the ECCV (ECCV): 132–149.

[27] Zhuang C, et al., 2019, Local Aggregation for Unsupervised Learning of Visual Embeddings. In Proceedings of the IEEE ICCV: 6002–6012.

[28] Yan X, Misra I, et al., 2019, Clusterfit: Improving Generalization of Visual Representations. arXiv:1912.03330.

[29] Caron M, Misra I, et al., 2020, Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. arXiv preprint arXiv:2006.09882, 2020.

[30] Goyal JP, Caron M, et al., 2021, Self-Supervised Pretraining of Visual Features in the Wild. arXiv preprint arXiv:2103.01988.

[31] Wu Z, Xiong Y, et al., 2018, Unsupervised Feature Learning Via Non-Parametric Instance Discrimination. In CVPR: 3733–3742, 2018.

[32] Tian Y, Krishnan D, et al., 2019, Contrastive Multiview Coding. arXiv preprint arXiv:1906.05849.

[33] X. Chen, H. Fan, et al., 2020, Improved Baselines with Momentum Contrastive Learning. arXiv preprint arXiv:2003.04297.

[34] Tian Y, Sun C, et al., 2005, What Makes for Good Views for Contrastive Learning. arXiv preprint arXiv:2005.10243.

[35] Grill JB, Strub F, et al., 2006, Bootstrap your Own Latent: A New Approach to Self-Supervised Learning. arXiv preprint arXiv:2006.07733.

[36] J. Mitrovic, B. McWilliams, et al., 2010, Representation Learning Via Invariant Causal Mechanisms. arXiv preprint arXiv:2010.07922.

[37] Chen X, et al., 2011, Exploring Simple Siamese Representation Learning. arXiv:2011.10566.

[38] Mikolov T, Sutskever I, et al., 2013, Distributed Representations of Words and Phrases and Their Compositionality.

[39] Arora S, Khandeparkar H, et al., 2019, A Theoretical Analysis of Contrastive Unsupervised Representation Learning.

[40] Iter D, Guu K, et al., 2020, Pretraining with Contrastive Sentence Objectives Improves Discourse Performance of Language Models.

[41] Chi Z, Dong L, Wei F, et al., 2020, Infoxlm: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training.

[42] Fang H, Wang S, Zhou M, et al., 2020, Cert: Contrastive Self-Supervised Learning for Language Understanding.

[43] Giorgi J, Nitski O, Bader G, et al., 2020, Declutr: Deep Contrastive Learning for Unsupervised Textual Representations.

[44] Ma, Shuang, et al., 2021, “Active Contrastive Learning of Audio-Visual Video Representations.” ICLR 2021: The Ninth International Conference on Learning Representations.

[45] Radford, Alec, et al., 2021, “Learning Transferable Visual Models from Natural Language Supervision.” ArXiv Preprint ArXiv:2103.00020.

[46] Jia, Chao, et al., 2021, “Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision.” ArXiv Preprint ArXiv:2102.05918.

[47] Huo Y, et al., 2021, “WenLan: Bridging Vision and Language by Large-Scale Multi-Modal Pre-Training.” ArXiv Preprint ArXiv:2103.06561.

[48] Wei C, Xie L, Ren X, et al., 2019, Iterative Reorganization with Weak Spatial Constraints: Solving Arbitrary Jigsaw Puzzles for Unsupervised Representation Learning. In CVPR: 1910–1919.

[49] Li, W. Hung J, Huang S, et al., 2016, Unsupervised Visual Representation Learning by Graph-Based Consistent Constraints. In ECCV: 678–694. Springer.

[50] Noroozi M, Vinjimoor A, Favaro P, et al., 2018, Boosting Self-Supervised Learning Via Knowledge Transfer. In CVPR: 9359–9367.

[51] Yang J, Parikh D, Batra D, 2016, Joint Unsupervised Learning of Deep Representations and Image Clusters. In CVPR: 5147–5156.

[52] Olivier J, Hénaff, Razavi A, Doersch C, et al., 2019, Data-Efficient Image Recognition with Contrastive Predictive Coding. arXiv preprint arXiv:1905.09272.