International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 07 July 2024, pp: 427-432 www.ijaem.net ISSN: 2395-5252

A comprehensive analysis of Self-supervised methods for image classification

Vu Thi Khanh Trinh
Tan Trao University, Tuyen Quang, Vietnam

Date of Submission: 10-07-2024 Date of Acceptance: 20-07-2024

ABSTRACT: Self-supervised learning is a type of machine learning where the model learns to predict part of its input data based on other parts of the data. Unlike supervised learning, which relies on labeled datasets, self-supervised learning generates its own labels from the input data itself, thus reducing the need for manual labeling. In self-supervised learning, the model typically learns to solve a pretext task, an auxiliary task that doesn't directly relate to the main goal but helps the model learn useful features from the data. Once the model is trained on this pretext task, it can be fine-tuned or used directly for the main task, often resulting in improved performance. There are many benefits from this approach such as reducing the need for largelabeled datasets, which can be expensive and time-consuming to produce, leveraging of vast amounts of unlabeled data, which is often more readily available than labeled data, etc. In this paper, we present a comprehensive analysis of popular self-supervised methods proposed in recent years. Specifically, we categorize it into two main approaches including Contrastive Learning and Non-Contrastive Learning. For each group, we present the details of several proposed methods and evaluate these method's effectiveness.

KEYWORDS: Self-supervised learning, image classification, deep learning.

I. INTRODUCTION

Over the past decade, advancements in computer vision [1, 2, 3], speech processing [4, 5], and natural language processing [6] have drawn significant attention, largely due to the availability of large-scale datasets[7,8]. Nonetheless, annotating new datasets remains essential for addressing issues in emerging domains. This process is both time-consuming and labor-intensive, making the ability to leverage unlabeled data highly advantageous. For instance, ImageNet [9], a widely used dataset for pre-training deep 2D Convolutional Neural Networks (CNNs), contains approximately 1.3 million labeled images across 1,000 classes.

To circumvent the need for time-intensive and costly data annotations, numerous self-supervised methods have been developed [8, 10, 11, 12, 13]. These methods aim to learn visual features from large-scale datasets for video recognition tasks without relying on human annotations. Self-supervised learning involves creating a pre-training or “pretext” task to extract knowledge from unlabeled data. Once a model is trained on the pretext task, it can be adapted to the target task through transfer learning.

In self-supervised learning, the input data is often transformed to challenge the model to predict missing parts, recognize transformations applied to the data, or handle information bottlenecks. Pseudo-labels are automatically generated for the pretext task to exploit the data structure. A CNN model is then trained to solve tasks derived from these pseudo-labels without human intervention. Examples of such tasks include solving image patch jigsaw puzzles [11], predicting frame order [12, 14], analyzing motion and appearance statistics [10], and identifying image color channels [15]. The trained CNN can subsequently be used as a feature extractor for other video tasks or as a weight initializer for downstream tasks.

II. BACKGROUND

Supervised Learning are algorithms that use labeled data to model the relationship between input data and their labels. Supervised learning requires very high labeling data, often the data sets will have to be labeled by humans, some data sets require high accuracy and importance, which often takes a lot of resources, moreover, the labeled data sets are not necessarily comprehensive and appropriate.

Unsupervised Learning are algorithms that use unlabeled data. These algorithms often aim to model the structure or hidden information in the

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 07 July 2024, pp: 427-432 www.ijaem.net ISSN: 2395-5252

data, thereby describing the characteristics and properties of the data set. These methods are often used in the process of analyzing and visualizing data

Obviously generating a good dataset with full labels is very expensive but unlabeled data is always generated. To take advantage of this much larger amount of unlabeled data, an approach can learn key features from within the data. The fact that the model can learn key features from within the data can be used to improve the modeling of some task on that data set, and such approaches are commonly known for their Sefl-Supervised learning mathematics.

Self-supervised learning is a machine learning process in which the model trains itself to learn the internal characteristics of the data. In this process, unsupervised machine learning problems are transformed into supervised machine learning problems by automatically generating labels. To take advantage of large amounts of unlabeled data, it is important to set appropriate learning goals so that the model can learn features from the data itself.

After being trained, self-supervised models will be used to finetune (using the parameters of the model trained for the task of the self-supervised problem for the model of the main task) for downstream tasks. (tasks that the model wants to learn) makes the models more robust, thereby increasing the model's performance. For example, we will use parameters of a backbone learned from unlabeled data (pictures of cars, houses, cats...) through the process of training a self-supervised model to train a classification model.

Unlike supervised learning, most self-supervised learning methods require a data pair $\{x_i, z_i\}$ where z_i is automatically generated for a predefined pretext task without human annotation. Figure 1 provides an overview of self-supervised learning-based methods. In these methods, a deep network acts as a feature extractor to learn spatio-temporal features from the input video through pretext tasks. Once self-supervised training is complete, the learned visual features can be transferred to downstream tasks, i.e., the target tasks.

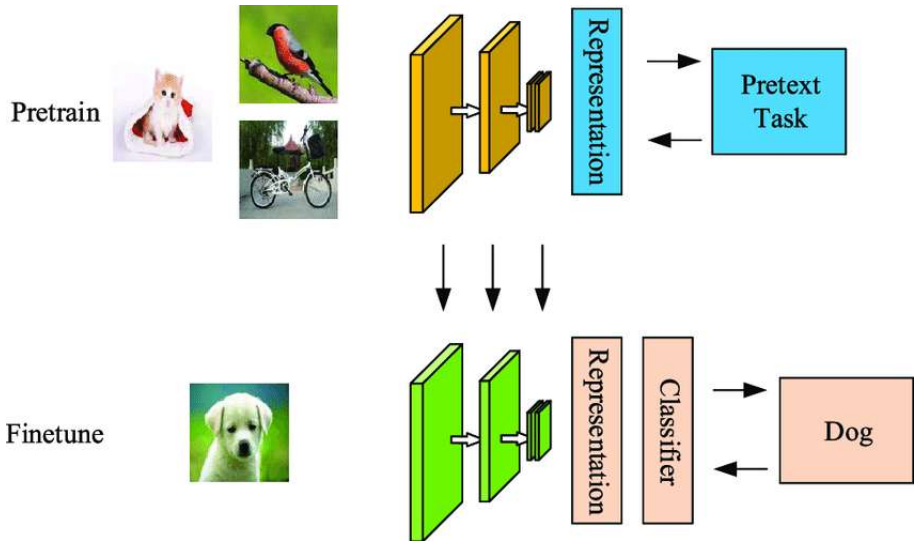

Figure 1. An overview of self-supervised learning approaches

III. THE STATE-OF-THE-ART SELF-SUPERVISED APPROACHES

3.1. Contrastive Learning

The goal of constrastive learning is to learn an embedded space in which pairs of similar samples are close together and pairs of different samples are far apart. The tasks of self-supervised learning are called pretext tasks and aim to automatically create pseudo-labels. There are many different ways to automatically create hypothetical tasks for example in the image there will be some methods like:

- Change color
- Rotate and crop photos
- Other geometric transformations

There are many new methods and techniques for conducting training that combine self-supervised and contrastive learning. The three most important components of these techniques are

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 07 July 2024, pp: 427-432 www.ijaem.net ISSN: 2395-5252

how we define and construct the pretext tasks, backbone, and contrastive loss. I have researched many studies around this topic and found that optimizing this combination is also one of the popular research trends today. Some popular studies in this topic include:

3.1.1. SimCLR

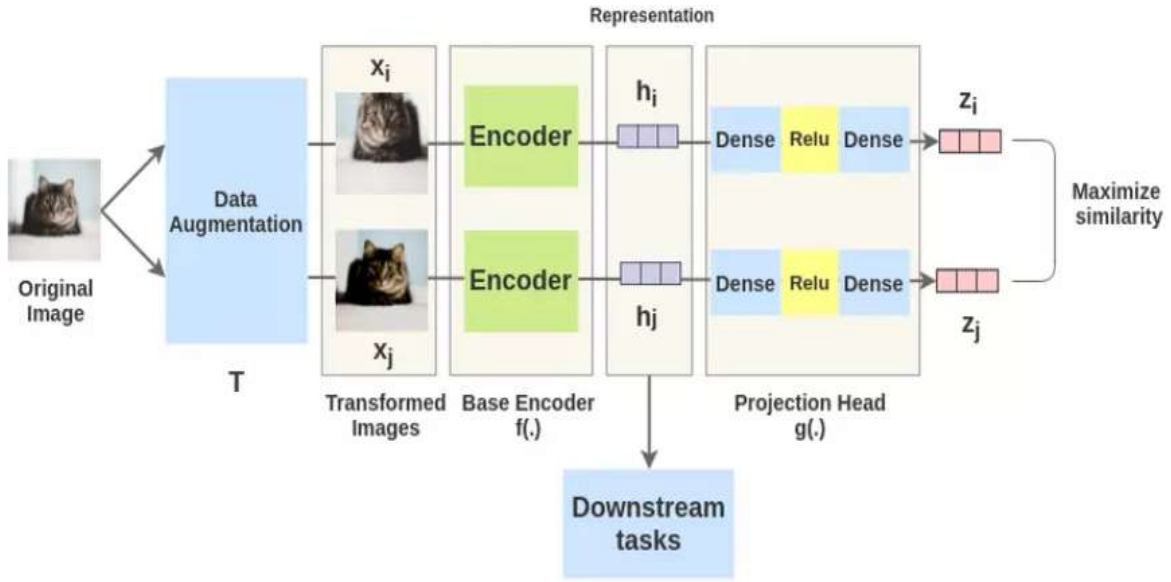

Figure 2. An overview of SimCLR approach

Based on SimCLR framework [16], it can be seen that pretext tasks are defined based on data augmentation methods, the model will create corresponding features for each sample. A loss function is designed based on constrastive learning to optimize the distance between features (similar samples will be closer together and different samples will be farther apart). The parameters learned by the model will be fine-tuned for downstream tasks (see Figure 2).

3.1.2. SwAV

In 2020, a paper proposed the SwAV (Swapping A ssignments between many Views) [17] model, which is a method for comparing cluster assignments to contrast different image views and does not rely on Compare features between images. The goal of this method is to learn image features in an online, unsupervised manner. Therefore, the authors propose a self-supervised learning method based on online clustering or we can call it Contrasting Cluster Assignments (see Figure 3).

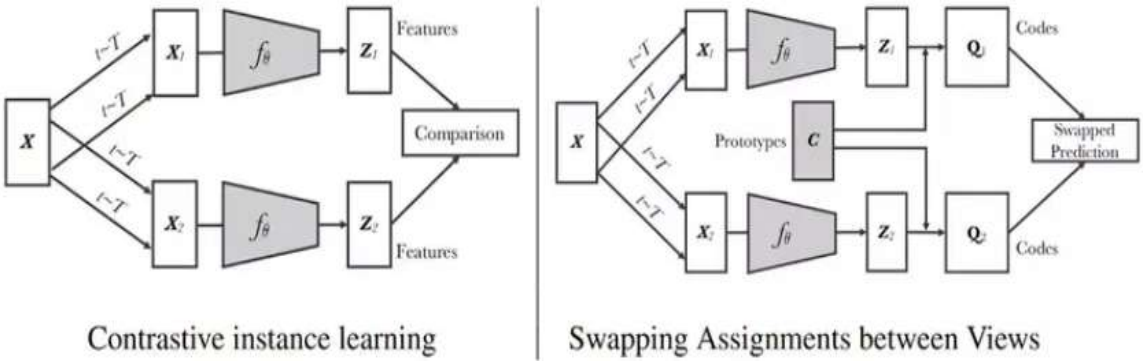

Figure 3. An overview of SwAV approach

3.2. Non-Contrastive Learning

Unlike contrastive learning, Non-Contrastive Self Supervised Learning (NC-SSL) trains a machine learning model in which only positive sample pairs are used to train a model. This idea seems unreasonable, because the model

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 07 July 2024, pp: 427-432 www.ijaem.net ISSN: 2395-5252

only focuses on minimizing the distance between pairs of positive data. However, NC-SSL has shown that it can learn a good representation using only positive samples combined with some extra prediction methods and stop-gradient operations. Furthermore, the learned representation shows comparable (or even better) performance for downstrem tasks. You can refer to this research here: Non-Contrastive Learning

3.2.1. Contrastive Predictive Coding (CPC)

The central idea of Contrastive Predictive Coding [18] is to first divide the entire image into an image grid and given information about the upper rows of the image, the task is to predict the lower rows of the same image. To perform this task, the model must learn the structure of the object in the image (for example, seeing the face of a dog, the model must predict that it will have 4 legs). Having the model trained like this will have a huge impact on downstream tasks (see Figure 4).

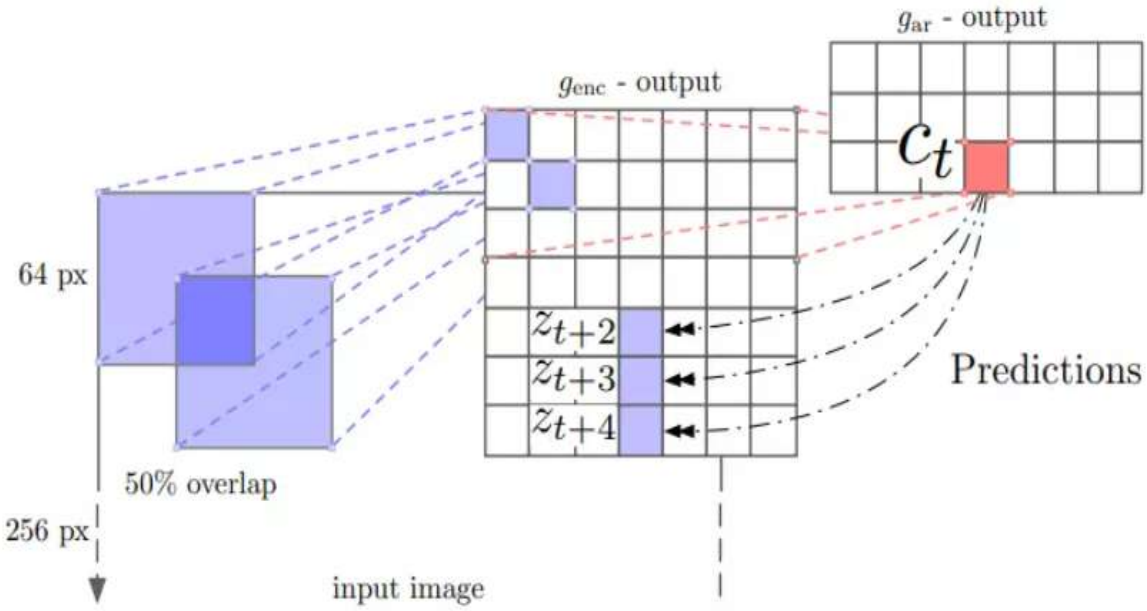

Figure 4. The detail of CPC approach

3.2.2. Self-Supervised Learning via Multi-Transformation Classification

In [19], the authors present a self-supervised video representation learning method based on multi-transformation classification for efficient human action classification. This approach leverages various transformations in self-supervised learning to provide richer contextual information and enhance the robustness of visual representations to these transformations. The spatio-temporal representation of the video is learned in a self-supervised manner by classifying seven different transformations: rotation, clip inversion, permutation, split, join transformation, color switch, frame replacement, and noise addition. Initially, these seven video transformations are applied to video clips. Then, 3D convolutional neural networks are employed to extract features from the clips, and these features are processed to classify the pseudo-labels. The learned models are used as pre-trained models in pretext tasks and fine-tuned for recognizing human actions in downstream tasks. Experiments conducted on the UCF101 and HMDB51 datasets, using C3D and 3D Resnet-18 as backbone networks, demonstrate that the proposed framework outperforms other state-of-the-art self-supervised action recognition approaches.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 07 July 2024, pp: 427-432 www.ijaem.net ISSN: 2395-5252

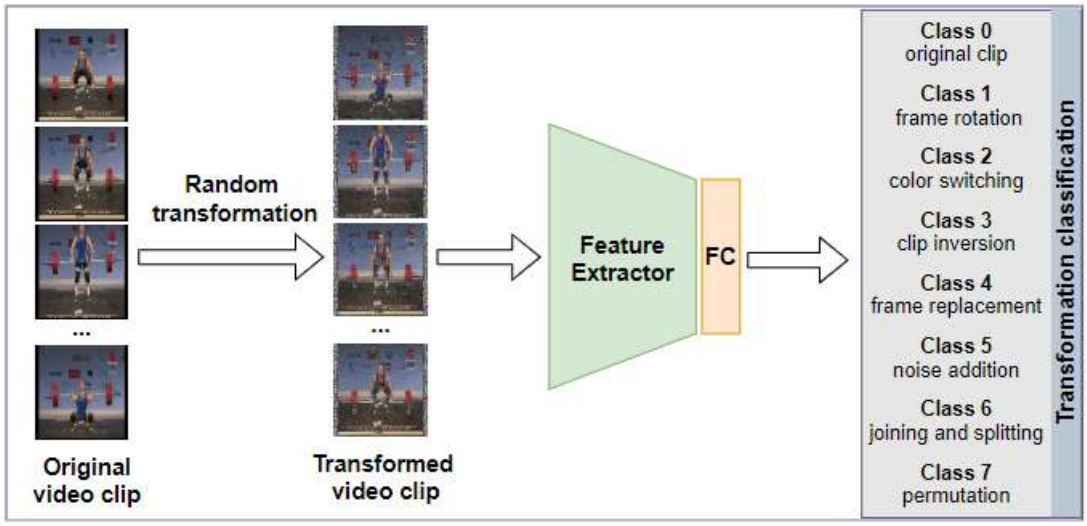

Figure 5. The detail of the Self-Supervised Learning method via Multi-Transformation Classification

As shown in Figure 5, this paper introduces a novel approach to self-supervised spatio-temporal video representation learning by predicting a set of video transformations. This task is particularly suitable for 3D CNNs, which are capable of modeling spatio-temporal information. Experimental results show that the proposed method achieves state-of-the-art performance in self-supervised video action recognition on the UCF101 and HMDB51 datasets. Additionally, it outperforms some methods that utilize the much larger-scale Kinetics dataset. These findings demonstrate the efficacy of the proposed method in predicting video transformations. The authors also suggest that the proposed model can serve as a powerful feature extractor for other tasks.

IV. CONCLUSION

In recent years, the approach of exploiting unlabeled data information has been a research trend of great interest with many methods such as self-supervised learning, semi-supervised learning, active learning, etc... has been bringing efficiency in the field of AI. Self-supervised learning is a popular method to deploy and apply in many problems and is surprisingly effective, especially in specific data problems (biomedical data problems, data problems, human data...).

REFERENCES

[1]. Vu, D. Q., Phung, T. T., Wang, J. C., & Mai, S. T. (2024). LCSL: Long-tailed Classification via Self-labeling. IEEE Transactions on Circuits and Systems for Video Technology.

[2]. Cheng, W. C., Mai, T. H., & Lin, H. T. (2023, December). From SMOTE to Mixup for Deep Imbalanced Classification. In International Conference on Technologies and Applications of Artificial Intelligence (pp. 75-96). Singapore: Springer Nature Singapore.

[3]. Vu, D. Q., & Thu, T. P. T. (2023). Simultaneous context and motion learning in video prediction. Signal, Image and Video Processing, 17(8), 3933-3942.

[4]. Tan, H. M., Vu, D. Q., & Wang, J. C. (2023, June). Selinet: a lightweight model for single channel speech separation. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[5]. Tan, H. M., Vu, D. Q., Thi, D. N., & Thu, T. P. T. (2023, December). Voice Separation Using Multi Learning on Squash-Norm Embedding Matrix and Mask. In International Conference on Advances in Information and Communication Technology (pp. 327-333). Cham: Springer Nature Switzerland.

[6]. Nga, C. H., Vu, D. Q., Luong, H. H., Huang, C. L., & Wang, J. C. (2023). Cyclic Transfer Learning for Mandarin-English Code-Switching Speech Recognition. IEEE Signal Processing Letters.

[7]. Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299-6308).

[8]. Fernando, B., Bilen, H., Gavves, E., & Gould, S. (2017). Self-supervised video

representation learning with odd-one-out networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3636-3645).

[9]. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[10]. Wang, J., Jiao, J., Bao, L., He, S., Liu, Y., & Liu, W. (2019). Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4006-4015).

[11]. Ahsan, U., Madhok, R., & Essa, I. (2019, January). Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 179-189). IEEE.

[12]. Misra, I., Zitnick, C. L., & Hebert, M. (2016). Shuffle and learn: unsupervised learning using temporal order verification. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14 (pp. 527-544). Springer International Publishing.

[13]. Vu, D. Q., Le, N. T., & Wang, J. C. (2021). Self-supervised learning via multi-transformation classification for action recognition. arXiv preprint arXiv:2102.10378.

[14]. Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., & Zhuang, Y. (2019). Self-supervised spatiotemporal learning via video clip order prediction. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10334-10343).

[15]. Zhang, Richard, Phillip Isola, and Alexei A. Efros. "Colorful image colorization." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. Springer International Publishing, 2016.

[16]. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

[17]. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems, 33, 9912-9924.

[18]. Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

[19]. Vu, D. Q., Le, N. T., & Wang, J. C. (2021). Self-supervised learning via multi-transformation classification for action recognition. arXiv preprint arXiv:2102.10378.