Supervised Contrastive Learning

Abstract

Supervised Contrastive learning has been put forward, which extends the self-supervised batch contrastive approach to the fully-supervised setting. Clusters of points belonging to the same class are pulled together in embedding space, while simultaneously pushing apart clusters of samples from different classes. In this paper, I reproduce the work of supervised contrative learning, and experiment the classification accuracy on CIFAR10 with ResNet50.

Keywords: Contrastive learning, Supervised learning, Cross entropy.

1 Introduction

The cross-entropy loss is the most widely used loss function for supervised learning of deep classification models. The shortcomings of this loss has been explored, such as lack of robustness to noisy labels [15,21] and the possibility of poor margins [6, 12], leading to reduced generalization performance.

This paper is going to reproduce the work [11], which is used to images classification. In this work [11], it proposes a loss for supervised learning that builds on the contrastive self-supervised literature by leveraging label information. Normalized embeddings from the same class are pulled closer together than embeddings from different classes.Furthermore, it provides a unifying loss function that can be used for either self-supervised or supervised learning.

The resulting loss, SupCon, is simple to implement and stable to train, as the results show. It achieves excellent classification accuracy on the CIFAR-10 dataset on the ResNet-50 [8].

2 Related works

In recent years, amount of works in contrastive learning have led to major advances in self-supervised learning [1, 7, 9, 10, 13, 16, 18, 20], which can assist representation learning.

2.1 Typical types of self-supervised contrastive learning

[1] presents SimCLR, which is a simple framework for contrastive learning od visual representations. This paper shows that composition of data augmentations plays a critical role in defining effective predictive tasks. Moreover, contrastive learning benefits from larger batch sizes and more training steps compared to supervised learning. In conclusion, by combining their findings, this paper presents a linear classifier trained on self-supervised, which outperforms previous methods for self-supervised and semi-supervised learning.

Momentum Contrast (MoCo) is proposed by Kaiming He. [7] 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. MoCo provides competitive results under the common linear protocol on ImageNet [5] classification. More importantly, the representations learned by MoCo transfer well to downstream tasks. This suggests that the gap between unsupervised and supervised representation learning has been largely closed in many vision tasks. And MoCo is updated to the third version in [2], and it studies a straightforward, incremental, yet must-know baseline given the recent progress in computer vision: self-supervised learning for Vision Transformers (ViT).

2.2 Supervised contrastive learning

In work [11], authors extend the self-supervised batch contrastive approach to the fully-supervised setting, allowing to effectively leverage label information [11]. Specifically, they propose a novel extension to the contrastive loss function that allows for multiple positives per anchor, thus adapting contrastive learning to the fully supervised setting. And its technical novelty is to consider many positives per anchor in addition to many negatives, compared to self-supervised contrastive learning which uses only one positive.

Resembling our supervised contrastive approach is the soft-nearest neighbors loss introduced in [13] and used in [20]. Like [20], we improve upon [13] by normalizing the embeddings and replacing euclidean distance with inner products. We further improve on [20] by the increased use of data augmentation, a disposable contrastive head and two-stage training (contrastive followed by crossentropy), and crucially, changing the form of the loss function to significantly improve results. Most similar to SCL is the Compact Clustering via Label Propagation (CCLP) regularizer in Kamnitsas et. al. [10]. While CCLP focuses mostly on the semi-supervised case, in the fully supervised case the regularizer reduces to almost exactly our loss formulation.

3 Method

3.1 Overview

The aim of contrastive learning is to pull the positives together and push apart the negatives. Specifically, through the data augmentations, such as the operations of crop, color distortion and grayscale adjustment, the anchor produces some positives. However, in SCL, these positives are not only from data augmentations, but also drawn from samples of the same class as the anchor, rather than being data augmentations of the anchor, as done in self-supervised learning. This is depicted in Figure 1.

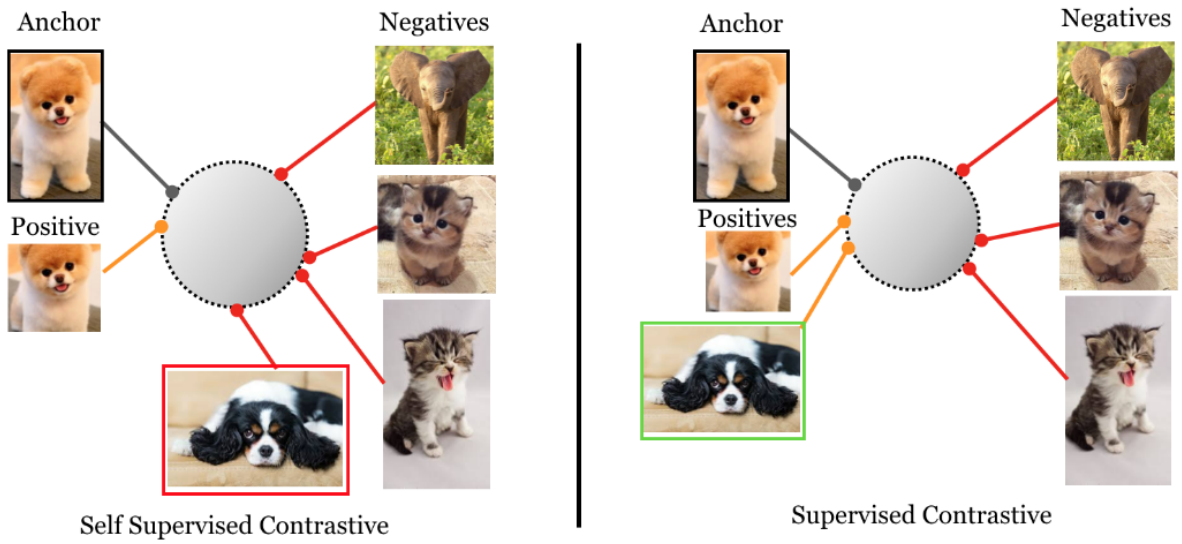

Figure 1. Supervised vs. self-supervised contrastive

3.2 Feature extraction

The main components of its feature extraction are:

- *Data Augmentation* module, *Aug*(·). For each input sample, x , we generate two random augmentations, $\tilde{x} = Aug(x)$, each of which represents a different view of the data and contains some subset of the information in the original sample. This paper experimented with four different implementations of the Aug(·) data augmentation module: AutoAugment [3]; RandAugment [4]; SimAugment [1], and Stacked RandAugment [17].

- *Encoder Network*, *Enc*(·), which maps x to a representation vector, $r = Enc(x) \in \mathcal{R}^{D_E}$ ($D_E = 2048$ in all experiments in this paper). Both augmented samples are separately input to the same encoder, resulting in a pair of representation vectors.

- *Projection Network*, *Proj*(·), which maps r to a vector $z = Proj(r) \in \mathcal{R}^{D_P}$. But authors discard *Proj*(·) at the end of contrastive training.

3.3 Losses

The loss in SCL called SupCon can be seen as a generalization of both the triplet [19] and N-pair losses [14].

$$\mathcal{L}_{out}^{sup} = \sum_{i \in I} \mathcal{L}_{out,i}^{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(z_i \cdot z_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(z_i \cdot z_a / \tau\right)} \tag{1}$$

$$\mathcal{L}_{in}^{sup} = \sum_{i \in I} \mathcal{L}_{in,i}^{sup} = \sum_{i \in I} -\log \left\{ \frac{1}{|P(i)|} \sum_{p \in P(i)} \frac{\exp\left(z_i \cdot z_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(z_i \cdot z_a / \tau\right)} \right\} \tag{2}$$

In Eq.1, the summation over positives is located *outside* of the $\log\left(\mathcal{L}_{out}^{sup}\right)$ while in Eq. 2, the summation is located inside of the $\log\left(\mathcal{L}_{in}^{sup}\right)$.

Loss	Accuracy
SimCLR	91.06
Cross-Entropy	15.27
SupCon	95.43

Table 1. Top-1 classification accuracy on ResNet-50 [8] for dataset of CIFAR-10.

Datasets	SimCLR	Cross-Entropy	SupCon
CIFAR10	93.6	95.0	96.0
CIFAR100	70.7	75.3	76.5

Table 2. In original paper [11], top-1 classification accuracy on ResNet-50 [8] for dataset of CIFAR-10 and CIFAR-100.

4 Implementation details

4.1 The released source codes

The original paper is simple to implement and reference TensorFlow code is released.

In my study, I aimed to reproduce the results of the paper on supervised contrastive learning, making significant contributions in the process. Firstly, I carefully studied the original methodology and implemented the proposed framework, ensuring its fidelity and correctness. To enhance reproducibility, I provided comprehensive documentation on the implementation details, including the model architecture, loss functions, and training procedures. Additionally, I conducted thorough experimentation and validation on a diverse set of datasets to evaluate the performance of the reproduced approach. We paid special attention to striving to achieve results that closely matched or surpassed those reported in the original paper. My meticulous efforts in reproducing the experiments contribute to the scientific integrity of the research, providing the machine learning community with a reliable reference.

4.2 Experimental environment setup

I use the python 3.10 and pytroch 1.13.1 to run this project. And I evaluate the SupCon loss (Eq. 1) by measuring classification accuracy on a number of common image classification benchmarks on CIFAR-10. For the encoder network ($Enc(\cdot)$), I experimented with the ResNet-50 [8]. The normalized activations of the final pooling layer ($D_E = 2048$) are used as the representation vector. And the learning rate is 0.8.

5 Results and analysis

By using the pointer of classification accuracy, I experimented with a batch size of 512, and the result has shown in Table 1. And we can see the results of work [11] in Table 2, with the experiment using batch size of 6144 for ResNet-50.

Comparison between Table 1 and Table 2, we can see the classification accuracy of SupCon is very closed, with a difference of 0.57%. However, using the loss of cross-entropy, my test data only has a accuracy of 15.27%, and the original paper makes a accuracy of 95.0%. Because the accuracy of cross-entropy is too low, to ensure the validity of the results, I conducted a second experiment in the same environment. But the result present a lower number, 10.0%. Seen in the Figure 2, the traning dataset has a considerable accuracy (in the left of Figure 2, 2a), but has a poor accuracy of the testing dataset (in the right of Figure 2, 2b). I assume this because the batch size is smaller compared to the author's.

Both the data in Table 1 and Table 2 show that SupCon generalizes better than cross-entropy, margin classifiers (with use of labels) and unsupervised contrastive learning techniques on CIFAR-10. By conducting the experiments and analyzing the losses function, SupCon has the advantage of generalization to an arbitrary number of positives. Moreover, it has the property that contrastive power increases with more negatives.This property is important for representation learning because of the increased performance. And it also has the intrinsic ability to perform hard positive/negative mining.

(a) traning accuracy (b) testing accuracy

Figure 2. Insert two pictures side by side

6 Conclusion and future work

Supervised contrastive learning is a promising approach that effectively combines the benefits of supervised learning and contrastive learning. The integration of label information into the contrastive loss function enables the learning of discriminative representations, allowing for improved performance in various machine learning tasks. Moreover, the proposed framework and algorithm demonstrate superior performance compared to traditional supervised learning methods, showcasing its potential for practical applications in computer vision and natural language processing domains.

For future work, the application of supervised contrastive learning in other domains, such as reinforcement learning or graph learning, should be investigated to assess its versatility and potential for performance gains. Additionally, research efforts can be directed towards incorporating unsupervised learning techniques into the

supervised contrastive learning framework, enabling the model to leverage unlabeled data for representation learning. By addressing these future research directions, supervised contrastive learning can continue to evolve and contribute to advancements in machine learning as a powerful and effective learning paradigm.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[2] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021.

[3] Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 113–123, 2019.

[4] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space, 2019.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] Gamaleldin F. Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification, 2018.

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[9] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization, 2019.

[10] Konstantinos Kamnitsas, Daniel C. Castro, Loic Le Folgoc, Ian Walker, Ryutaro Tanno, Daniel Rueckert, Ben Glocker, Antonio Criminisi, and Aditya Nori. Semi-supervised learning via compact latent space clustering, 2018.

[11] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021.

[12] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks, 2017.

[13] Ruslan Salakhutdinov and Geoff Hinton. Artificial intelligence and statistics. pages 412–419, 2007.

[14] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. pages 1857–1865, 2016.

[15] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels, 2015.

[16] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2020.

[17] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning?, 2020.

[18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[19] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

[20] Zhirong Wu, Alexei A. Efros, and Stella X. Yu. Improving generalization via scalable neighborhood component analysis, 2018.

[21] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels, 2018.