CSCE 689 Final Project Report

Haotian Xu, Chi Zhang
Texas A&M University
hx105@tamu.edu,czhang241@tamu.edu

1 Introduction

Self-supervised Learning (SSL) for pre-training deep neural networks has emerged to be a popular paradigm in computer vision for learning visual representations. A simple but effective framework of self-supervised learning is contrastive learning. The goal of contrastive learning is to learn an embedding space in which similar sample pairs are close to each other while dissimilar sample pairs are separated from each other. SimCLR, a popular contrastive learning method, has demonstrated great success on benchmarks in computer vision tasks (Chen et al. [2020]). However, it suffers from the problem of large batch size. SogCLR is proposed to mitigate the problem (Yuan et al. [2022]). It uses an idea of running average to track the global contrastive loss and applies stochastic compositional optimization techniques to achieve a better gradient estimator.

In this project [1], we aim to develop better algorithms to optimize the global contrastive objectives based on SogCLR. Specifically, we propose several improvements that are summarized below:

- We use additional data augmentation techniques such as Gaussian noise, salt and pepper noise, and Gaussian blur to generate noisy versions of the samples and to facilitate the algorithms to learn better representations.
- We combine a memory bank approach (He et al. [2019]) with SogCLR. The memory bank approach decouples the batch size from the number of negatives. Therefore, it can provide a large sample of negatives for contrastive learning and help to learn better representations.
- We use sharpness aware minimization technique (Keskar et al. [2016], Foret et al. [2020]) to improve the generalization ability of the model by performing noise attacks on model parameters.
- We use retrieval augmentation technique (Guu et al. [2020], Xu et al. [2021]) to select hard negative samples for contrastive learning to boost the model performance.

We test our proposed methods on CIFAR 10 and CIFAR 100 datasets. The experiment results show that our proposed methods can achieve or surpass the SogCLR baseline.

2 Related Work

2.1 Contrastive Learning

Contrastive learning (Wu et al. [2018], Oord et al. [2018], Ye et al. [2019], Tian et al. [2019]) aims to learn effective representations by making the representations agree with one another under proper transformations while pushing different ones away. It requires a set of paired examples $\{(x_n, x_n^+)\}$, in which x_n and x_n^+ are semantically equivalent, i.e., x_n and x_n^+ should be closed in terms of 'distance' in the representation metric space. For any $i \neq j$ x_i and x_j are regarded as negative pair. The

[1] Github link: https://github.com/hxu105/TAMU_CSCE689_OPT

objective function of contrastive learning is

$$\mathbb{E}_{x_i \sim \mathcal{D}} \left[\log - \frac{e^{f(x_i) \cdot f(x_i^+)/\tau}}{\sum\limits_{j \neq i} e^{f(x_i) \cdot f(x_j^+)/\tau}} \right], \tag{1}$$

where $\mathcal{D}$ is the data distribution, $f(\cdot)$ is the function mapping from input space to the representation space, and τ is a scalar temperature hyperparameter.

Contrastive representation learning has been outstandingly successful in practice. There are many works using contrastive learning to generate a better visual representation in computer vision tasks (He et al. [2019], Chen et al. [2020], Caron et al. [2020], Zbontar et al. [2021], Chen et al. [2021b], Caron et al. [2021]). Contrastive learning has also been applied to natural language processing (Gao et al. [2021]), graph neural network (You et al. [2020]) and vision-language tasks (Zhang et al. [2021], Radford et al. [2021], Li et al. [2022b]).

Besides the success on the engineering side, there are also concrete theoretical explanations for the advantages of contrastive learning. Oord et al. [2018] shows that minimizing the contrastive objective loss can maximize the mutual information between samples and their positive pairs. In that sense, we can view contrastive learning as a means to learn a robust representation that is invariant towards different transformations in the input space. Wang and Isola [2020] offers another analysis by demonstrating that contrastive learning is capable of making positive pairs more aligned and making the representation distribution on unit hypersphere more uniform. They also conclude that optimizing for alignment and uniformity leads to effective representations.

However, there are still some issues remaining unsolved in contrastive learning. Previously, SimCLR (Chen et al. [2020]) uses contrastive learning to achieve stunning results in image classification while suffering from large batch sizes. Chen et al. [2021a] provide an explanation of why contrastive learning with InfoNCE loss fails in the small-batch-size regime. They show that learning efficiency plunges in the small batch size due to limited numerical precision resulting from using InfoNCE loss. To address the limitation of InfoNCE loss, they propose FlatNCE objective, a self-normalized contrastive objective, which puts larger weights on harder negative samples to facilitate the models learning better representation. Experimental evidence shows that FlatNCE is far less sensitive to the choice of mini-batch size. Tsai et al. [2021] proposes an objective function based on Relative Predictive Coding (RPC), which aims to maintain a good balance between training stability and minibatch size sensitivity. Yeh et al. [2021] aims to tackle the problems of large batch size and extensive training epochs. They identify a negative-positive-coupling (NPC) effect in InfoNCE-based loss, which results in low learning efficiency. To address the NCE effect, they propose a decoupled contrastive learning loss (DCL) by removing the positive term from the denominator and improving the learning efficiency. Experimental evidence shows that DCL requires neither large batches nor large epochs to achieve competitive performance. SogCLR, Yuan et al. [2022], applies the running average idea to keep tracking the global contrastive loss in order to mitigate the problem of requiring large batch size. Experiment results indicate that SogCLR can achieve comparable or even better performance while using a small batch size.

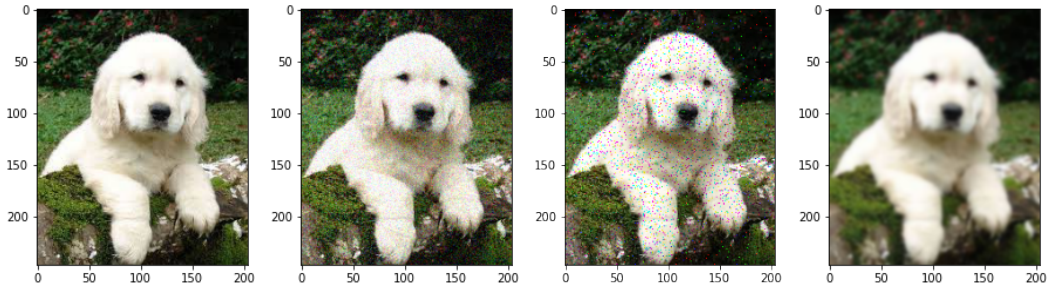
In addition, another stream of literature studied efficient sampling strategies for positive pairs (Chen et al. [2020], Tian et al. [2020]) and negative pairs (Kalantidis et al. [2020], Robinson et al. [2020]) in contrastive learning. For image data, positive sampling strategies usually rely on transformations that preserve semantic content (Chen et al. [2020]). Tian et al. [2020] suggest an "InfoMin principle", which should reduce the mutual information between views while preserving task-relevant information intact, to generate a good set of views. Kalantidis et al. [2020] proposes a hard negative mixing strategy to generate hard negatives in latent space. Robinson et al. [2020] designs an importance sampling technique for efficiently selecting hard negative samples where the hardness can be controlled.

3 Proposed Improvement

3.1 Data Augmentation

Strong data augmentation has been found effective for learning good visual representations. (Chen et al. [2020]) Two types of augmentation were used in the literature. One type of augmentation

employs spatial/geometric transformation of data, including cropping and resizing, rotation, and cutout. The other type of augmentation involves color jitter and color dropping. Following the second type of data augmentation approach in the literature, we introduce additional data augmentation operations: adding noise and applying Gaussian blur to the image data. Specifically, we explore two types of noise. One is Gaussian noise, which is a type of statistical noise added to each pixel of the image. The statistical noise follows the normal distribution. The other type is salt and pepper noise, which is a type of impulse noise. It adds random dark and random bright noises all over the images. And Gaussian blur is performing convolution operation on the image with a Gaussian function. Both adding noise and applying Gaussian blur can generate a noisy version of the data. Figure 1 illustrates the examples from applying additional data augmentation operators.



(a) Original (b) Gaussian Noise (c) Salt & Pepper Noise (d) Gaussian Blur

Figure 1: Illustrations of the additional data augmentation operators

3.2 Dynamic MoCo

He et al. [2019] points out that unsupervised learning in computer vision tasks requires building dictionaries since the data signals are continuous, distributed in high spatial dimensions and not structured like NLP. After classifying existing unsupervised methods as dictionary learning, the authors propose that building a dictionary depends on two necessary conditions: 1. the size of the dictionary needs to be large enough to represent the high-dimensional, continuous space well; 2. the keys of the dictionary need to be encoded from the same or similar encoders so that the distance measure between the query and the key can be consistent and meaningful.

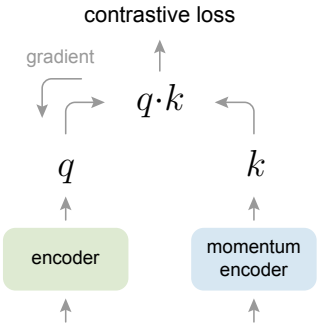


Figure 2: MoCo has the advantage of decoupling the batch size from the number of negatives. It trains a visual representation encoder by matching an encoded query q to a dictionary of encoded keys using a contrastive loss. The dictionary is built as a queue, with the current mini-batch enqueued and the oldest mini-batch dequeued. The keys are encoded by a slowly progressing encoder, which is updated by the query encoder through a momentum step. This updating rule guarantees the differences among these encoders are small.

Combining MoCo with SogCLR is a simple but effective approach. This approach can benefit from the large batch size provided by the memory bank in MoCo and tracking the global contrast loss gradient provided by SogCLR. Algorithm 1 gives a pesudocode for implementing MoCo+SogCLR.

Algorithm 1 Dynamic MOCO

for $img, \ldots$ in loader **do**
  $x_1, x_2 = aug1(img),\ aug2(img)$
  $q, k = f_q(x_1), f_k(x_2)$
  k.detach()
  l_{pos} = bmm(q.view(N,1,C), k.view(N,C,1))
  l_{neg} = mm(q.view(N,1,C), queue.view(N,C,1))
  loss = SogCLR($l_{pos},\ l_{neg}$)
  loss.backward()
  update(f)
end for

3.3 Sharpness Aware Minimization

There have been evidences showing that flat minima can generalize better than sharp minima in terms of loss landscape. Contrastive learning usually requires large batch size, and Keskar et al. [2016] argues that 1) the large batch size is more likely to converge to sharp minima while the small batch size is more likely to converge to flat minima, and the large batch size is difficult to jump out of the sharp minima pit; 2) the weakened generalization caused by large batch size is not caused by overfitting, because the early stopping method does not help in the experiment; 3) there is a threshold value for batch size, once exceeded, the performance of the model degrades and the accuracy drops dramatically. Thus, most of contrastive learning which requires a large batch size during training may also encounter the same sharp minima issue. Foret et al. [2020] proposes Sharpness-Aware Minimization (SAM) to efficiently learn a model with strong generalization ability. They applied adversarial training to attack the model parameter, and then perform gradient descent on the attacked parameters, i.e., seeking parameters that lie in neighborhoods having uniformly low loss.

$$\min_w L(w+\epsilon) + \lambda \|w\|_2^2, \tag{2}$$

$$\epsilon = \arg\max_{\|\epsilon\|_p \leq p} L(w+\epsilon), \tag{3}$$

From the adversarial training approach, ϵ is the attack that changes the parameter to make a maximum loss. It is well-known that parameters with low loss can form a manifold in the weight space. Since $L(w+\epsilon) \geq L(w)$ for any ϵ that we generated, performing gradient descent on $L(w+\epsilon)$ will lead to a lower loss of $L(w)$. SAM can be treated as a way of doing weight perturbation, and Zhu et al. [2019], Aghajanyan et al. [2020], Yu et al. [2022] show that adding perturbation in the weight space can not only improve model generalization ability but also enhance the robustness of the model against various attacks.

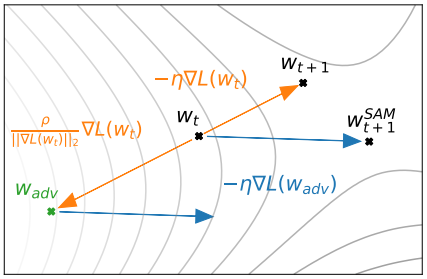


Figure 3: Schematic of the SAM parameter update

3.4 Retrieval Augmentation

One of the main concerns in contrastive learning is how to generate positive and negative samples. Using various data augmentation techniques can help generate different positive samples while keeping their semantic meaning intact. However, random sampling within the batch to form the negative samples will not always provide the true negatives. Retrieval augmentation Guu et al.

[2020] serves as a latent knowledge retriever, which allows the model to retrieve and acquire outside information to capture knowledge in a more modular and interpretable way. In contrastive learning, one can apply the same idea to generate negative samples. Xu et al. [2021] has used the retrieval augmentation method to generate harder negative samples in their video-text pretraining. They also claim that contrastive learning and retrieval-augmented training can have good synergy because the former aims to discriminate against examples and the latter aims to find harder examples for discrimination. In this project, we try to implement the similar idea to construct harder negative samples to boost the performance of contrastive learning.

4 Experiments and Results

This section introduces our experiment results and their comparison with SogCLR.

4.1 Experiments

Experiment setup. We use the standard CIFAR 10 dataset for training and testing the model. The backbone structure of the network is ResNet-50 with a 2-layer non-linear projection head with the hidden size of 128. The batch size is fixed as 64. The number of epochs for training is fixed as 400. We use the provided train/test splits of the dataset to perform the training and testing procedure. In the first stage, the model is pre-trained for 400 epochs using the unlabeled samples. In the second stage, 40k labeled samples are used to train the model for 90 epochs and 10k samples are used to perform the validation. For the first stage of training, we use a linear learning rate scaling with an initial leaning rate of 0.25. The LARS optimizer is used to optimize the model with a weight decay of 1e-4. The temperature parameter τ is set to be a fixed value of 0.3. The gamma parameter γ for maintaining the moving average estimator in global contrastive learning is set to be 0.9. For the second stage of training, we use momentum-SGD without weight decay and a batch size of 1024 of linear classification on frozen features/weights.

4.2 Results

Method	**Cifar 10**	**Cifar 100**
Baseline	89.36%	59.22%
GN $\sigma^2 = 9/10000$	89.71%	-
GN $\sigma^2 = 2/255$	72.89%	-
GN $\sigma^2 = 8/255$	70.85%	-
S&P Noise (50-50)	89.14%	-
MOCO+SogCLR	**89.87%**	**60.21%**
SAM	89.79%	59.34%
REAL	87.47%	55.36%

Table 1: Experiment results for comparison between data augmentation and training methods. Overall, our implemented methods can achieve or surpass the baseline performance.

We test our pretrained model generalization ability on Cifar10 and Cifar100 datasets. As shown in Table 1, our proposed methods can achieve or surpass the SogCLR baseline, and the methods are orthogonal to each other. Thus, one can combine them to see if there is more improvement.

Data Augmentation The motivation for using Gaussian noise or Gaussian blur is to generate corrupted images and approximate the effect of using an adversarial attack. There is a clear pattern that the performance on the clean dataset can be improved by different scales if we decrease the variance in noisy sampler or attack magnitude. Due to the time limit, we only try a few hyperparameter settings, but one can test different variance values to check the enhancement.

MOCO+SogCLR Essentially, MOCO+SogCLR is just having a large "batch" instead of randomly sampled, we memorize the past data instance within the memory bank. A direct comparison should be made between MOCO+SogCLR and SogCLR with large batch sizes. However, the default memory

bank size is 65336 which will be way larger than the common batch sizes we use on the machine. Thus, MOCO+SogCLR can be argued as an efficient way to asymptote the SogCLR with very large batch size. And the test results show the improvement of combining MOCO with SogCLR.

SAM There is a hypothesis about model generalization ability: flat minima will have better generalization ability than sharp ones. In the original SAM paper, they use adversarial attacks to perturb model parameters. However, generating adversarial attacks is time-consuming. We choose a uniform noise sampler for creating noise attacks. The noise-induced SAM method can prompt the model to achieve better performance on downstream tasks, and such implementation is also easy and the add-on training time is marginal as well. SAM can be a good trick to improve model generalization ability while not suffering from training time increments if using noise attacks.

REAL Retrieval Augmentation learning has shown a great performance in many NLP fields, and VideoClip also implements the idea of retrieval augmentation as a way of sampling hard negative examples. Moreover, contrastive learning and retrieval augmentation can be considered complementary to each other. Since contrastive learning is to make positive samples near the anchor points while pushing them away from the negative samples, retrieval augmentation, on the other hand, desires to provide hard negative samples. We make one hypothesis here: selecting only hard negative samples through retrieval augmentation should help contrastive learning since we mitigate the dominance of easy negative samples and shape the distribution of hidden features more uniformly on the hypersphere. However, the experiment results do not support our hypothesis. But we still think retrieval augmentation is a promising direction in contrastive learning. The time was limited to fully exploring the power of retrieval augmentation. One can try to implement retrieval augmentation differently or use other distance metrics.

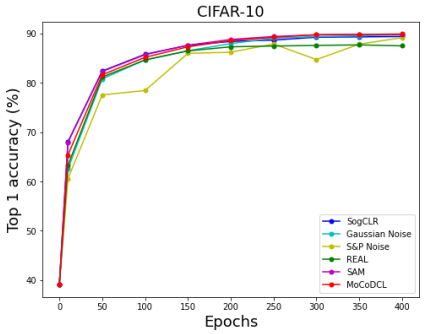


Figure 4: Testing trajectory for different methods

5 Conclusion

We have tried several interesting approaches in this project such as sharpness-aware minimization and retrieval augmentation. Even though the results of the experiments can only provide a marginal improvement, we have shown the possible direction for future development in contrastive learning. As professor Yang mentioned, one of the challenges in contrastive learning is how to obtain the hard negatives without passing through all data. The memory bank approach can help to set a criterion for choosing negative samples. But it would increase the training time cost and difficulty to set up a reasonable hyperparameter or distance metric. We think meta-learning can be another approach to be combined with retrieval augmentation to generate hard negatives. And there is a recently released paper emphasizing this approach. (Li et al. [2022a])

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse, 2020. URL `https://arxiv.org/abs/2008.03156`.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2020. URL https://arxiv.org/abs/2006.09882.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.14294.

Junya Chen, Zhe Gan, Xuan Li, Qing Guo, Liqun Chen, Shuyang Gao, Tagyoung Chung, Yi Xu, Belinda Zeng, Wenlian Lu, Fan Li, Lawrence Carin, and Chenyang Tao. Simpler, faster, stronger: Breaking the log-k curse on contrastive learners with flatnce, 2021a. URL https://arxiv.org/abs/2107.01152.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021b. URL https://arxiv.org/abs/2104.02057.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization, 2020. URL https://arxiv.org/abs/2010.01412.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings, 2021. URL https://arxiv.org/abs/2104.08821.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training, 2020. URL https://arxiv.org/abs/2002.08909.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2019. URL https://arxiv.org/abs/1911.05722.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21798–21809. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/f7cade80b7cc92b991cf4d2806d6bd78-Paper.pdf.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima, 2016. URL https://arxiv.org/abs/1609.04836.

Jiangmeng Li, Wenwen Qiang, Changwen Zheng, Bing Su, and Hui Xiong. Metaug: Contrastive learning via meta feature augmentation, 2022a. URL https://arxiv.org/abs/2203.05119.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022b. URL https://arxiv.org/abs/2201.12086.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples, 2020. URL https://arxiv.org/abs/2010.04592.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2019. URL https://arxiv.org/abs/1906.05849.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6827–6839. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c2e5eaae9152079b9e95845750bb9ab-Paper.pdf.

Yao-Hung Hubert Tsai, Martin Q. Ma, Muqiao Yang, Han Zhao, Louis-Philippe Morency, and Ruslan Salakhutdinov. Self-supervised representation learning with relative predictive coding, 2021. URL https://arxiv.org/abs/2103.11275.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere, 2020. URL https://arxiv.org/abs/2005.10242.

Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination, 2018. URL https://arxiv.org/abs/1805.01978.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding, 2021. URL https://arxiv.org/abs/2109.14084.

Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature, 2019. URL https://arxiv.org/abs/1904.03436.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning, 2021. URL https://arxiv.org/abs/2110.06848.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations, 2020. URL https://arxiv.org/abs/2010.13902.

Chaojian Yu, Bo Han, Mingming Gong, Li Shen, Shiming Ge, Bo Du, and Tongliang Liu. Robust weight perturbation for adversarial training, 2022. URL https://arxiv.org/abs/2205.14826.

Zhuoning Yuan, Yuexin Wu, Zi-Hao Qiu, Xianzhi Du, Lijun Zhang, Denny Zhou, and Tianbao Yang. Provable stochastic optimization for global contrastive learning: Small batch does not harm performance, 2022. URL https://arxiv.org/abs/2202.12387.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021. URL https://arxiv.org/abs/2103.03230.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis Langlotz. Contrastive learning of medical visual representations from paired images and text, 2021. URL https://openreview.net/forum?id=T4gXBOXoIUr.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding, 2019. URL https://arxiv.org/abs/1909.11764.

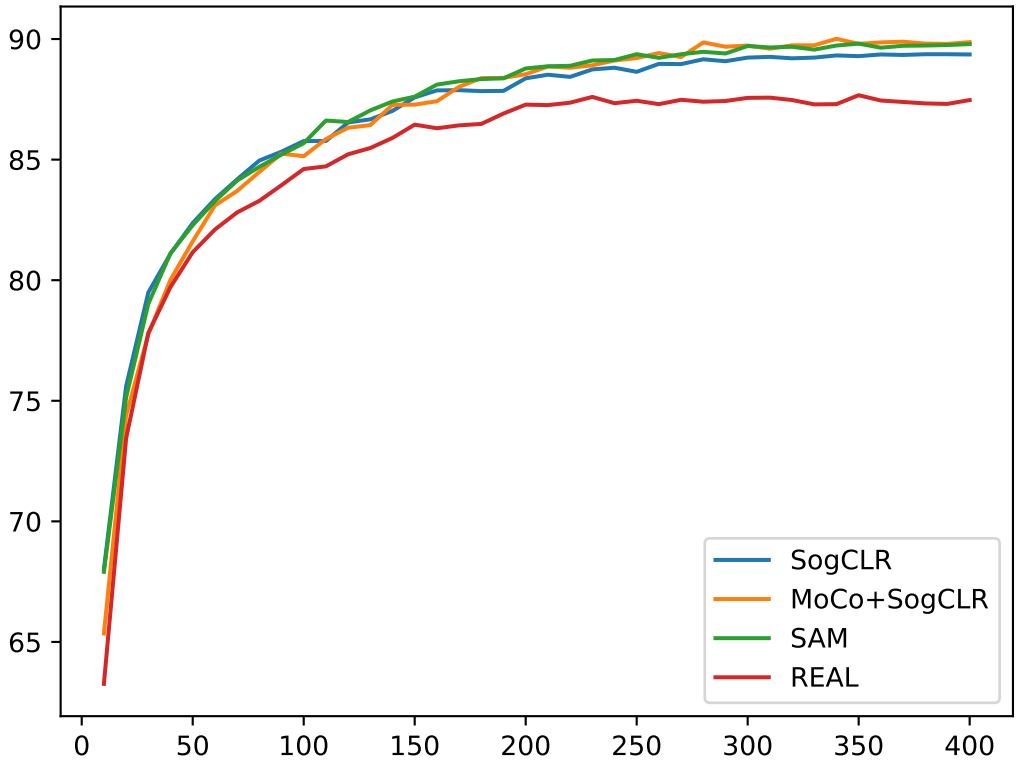


Figure 5: Cifar10 Testing trajectory for every 10 epochs. The x-axis is epochs, and y-axis is top 1 accuracy