TOWARDS GOOD PRACTICES IN SELF-SUPERVISED REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised representation learning has seen remarkable progress in the last few years. More recently, contrastive instance learning has shown impressive results compared to its supervised learning counterparts, in particular on downstream tasks like image classification and object detection. However, even with the ever increased interest in contrastive instance learning, it is still largely unclear why these methods work so well. In this paper, we aim to unravel some of the mysteries behind their success, which are the *good practices*. In particular, we investigate why the nonlinear projection head is essential, why instance discrimination does not suffer from strong data augmentation, and if large amounts of negative samples are required during contrastive loss computation. Through an extensive empirical analysis, we hope to not only provide insights but also lay out a set of best practices that led to the success of recent work in self-supervised representation learning.

1 INTRODUCTION

Self-supervised representation learning (SSL) has been a hot area of research in the last several years. The allure of SSL is the promise of annotation-free ground-truth to ultimately learn a superior data representation (when compared to supervised learning). Initial attempts to improve the quality of learned representations were via designing various pretext tasks, such as predicting the rotation (Gidaris et al., 2018), learning to count (Noroozi et al., 2017), context completion (Doersch et al., 2015; Pathak et al., 2016; Noroozi & Favaro, 2016), image colorization (Zhang et al., 2016), deep clustering (Caron et al., 2018), generative modeling (Donahue et al., 2016), motion prediction (Agrawal et al., 2015), etc. More recently, a type of contrastive learning method based on instance discrimination as pretext task (Dosovitskiy et al., 2014; Wu et al., 2018) has taken-off as it has been consistently demonstrated to outperform its supervised counterparts on downstream tasks like image classification and object detection (Oord et al., 2018; Tian et al., 2019; Misra & Maaten, 2020; He et al., 2020; Khosla et al., 2020; Chen et al., 2020a; Caron et al., 2020).

Putting the progress in contrastive instance learning aside for a moment, many recent observations seem to contradict what we have known from supervised learning. For example, Bachman et al. (2019); Chen et al. (2020a) have shown that simply adding a nonlinear projection head after the last pooled convolutional feature can significantly improve the quality of learned representations. Quantitatively, the nonlinear projection head can help to improve top-1 classification accuracy on ImageNet by over 10% in Chen et al. (2020a) and 5.6% in Chen et al. (2020c). However, adding such a shallow multilayer perceptron (MLP) projection head (i.e., one fully-connected layer and one activation layer) is often not effective in supervised learning. Take another example, recent methods (He et al., 2020; Chen et al., 2020a) adopt aggressive and strong data augmentation during contrastive pre-training. Although data augmentation has been proven to be useful, overly aggressive augmentations often lead to worse results in supervised or other self-supervised learning methods. Then why contrastive instance learning does not suffer from strong data augmentation? At this moment and to the best of our knowledge, there is no concrete evidence to answer these questions.

After closely observing recent contrastive instance learning work, it becomes apparent that what seem to be design choices are in-fact good practices which are largely responsible to their disruptive success. Furthermore, some of these good practices can effectively be transferred to other non-contrastive, unsupervised learning methods (Grill et al., 2020). Hence in this work, we focus on

the importance of these design choices using extensive experimental evidence and evangelize certain good practices which lead to their success. We hope to provide insights to the self-supervised learning community, with the potential impact and application even beyond it. Specifically, our contributions include:

- We empirically show why the nonlinear projection head helps contrastive instance learning and visualize it using a feature inversion approach (Section 3).
- We present the semantic label shift problem caused by strong data augmentation in supervised learning and study the difference between supervised and contrastive instance learning (Section 4).
- We investigate both quantity and quality of negative samples used in computing contrastive loss. We find that good practices can help to eliminate the need of using large number of negative samples, thereby could simplify the framework design (Section 5).

2 BACKGROUND

The goal of contrastive instance learning is to learn representations by distinguishing between similar and dissimilar instances. Images are typically perturbed using various augmentation techniques (e.g. random cropping, or color jittering) and a model is trained to map the original images and the perturbed ones as the same latent representation. It is important to note that in instance discrimination, each image together with its augmented versions is treated as its own class requiring no *a priori* class annotations. A generic visual depiction of recent contrastive instance learning methods is shown in Figure 1.

More formally, given a set of images $\mathcal{X}$, an image x is uniformly sampled from $\mathcal{X}$ and is augmented using various augmentation techniques $t \in \mathcal{T}$ to generate the positive pairs x_1^+ and x_2^+. f_θ and f_ϕ are encoders which map the images to visual representations h_1 and h_2, i.e., 2048-dim features from the last average pooling layer of a ResNet (He et al., 2016). These visual representations are then projected via g_θ and g_ϕ, which are often MLP heads, to lower dimensional features z_1 and z_2 for similarity comparison. Networks are usually trained by InfoNCE (Oord et al., 2018) loss,

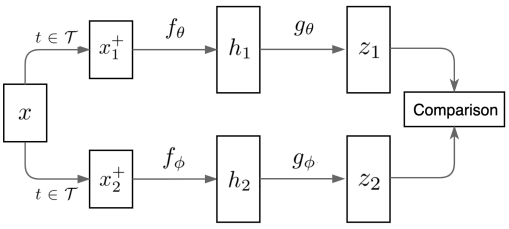


Figure 1: A generic visual depiction of recent contrastive instance learning methods.

$$L_{\text{NCE}} = -\log \frac{\exp[\text{sim}(z_1, z_2)]}{\exp[\text{sim}(z_1, z_2)] + \sum_{i=1}^{K} \exp[\text{sim}(z_1, g_\phi(f_\phi(x_i^-)))]}. \tag{1}$$

Here, $\text{sim}(.)$ is the distance metric used to measure the similarity between feature embeddings. K denotes the number of negative examples x^- for each x used in computing contrastive loss. It has been posited that negative examples are crucial to avoid learning collapsed representations (Grill et al., 2020), and can be selected from the rest of the dataset $\mathcal{X}$ via different mechanisms, e.g., memory bank (Wu et al., 2018), queue (He et al., 2020) or large mini-batch (Chen et al., 2020a). By optimizing the objective, the model learns to map similar instances closer and dissimilar instances farther apart in the embedding space. For simplicity, we use f and g to denote encoder network and MLP head, h and z to represent the feature before and after MLP head for the rest of the paper.

In this paper, we mainly use MoCov2 (Chen et al., 2020c) as an illustrating example, but also include experiments on other methods when needed. In terms of unsupervised pre-training, we train a ResNet50 backbone for 100 epochs, unless otherwise stated. In terms of evaluation, we follow the standard protocol of (Wu et al., 2018; He et al., 2020) to perform linear probe on ImageNet and report top-1 classification accuracy. To be specific, we freeze the encoder f and train a supervised linear classifier using feature h. MLP head g is discarded in the evaluation.

Table 1: Investigation of nonlinear projection head in MoCov2 (Chen et al., 2020c).

Method	Top-1 Acc (%)
MoCov2	64.4
(a) without projection head	59.2
(b) with fixed projection head	62.9
(c) with more hidden layers in projection head	63.9
(d) with reduced feature dimension in projection head	62.2

3 NONLINEAR PROJECTION HEAD

A main goal of unsupervised representation learning is to learn features that are transferable to downstream tasks. Typically the outputs of the penultimate layer (i.e. the pooling of the last convolution layer) are considered for transferring to other tasks. Recently, Bachman et al. (2019); Chen et al. (2020a;c) have shown that simply adding a nonlinear projection head as shown in Figure 1 can significantly improve the quality of learned feature representation h. For example, the nonlinear projection head helps to improve the top-1 classification accuracy on ImageNet by over 10% in Chen et al. (2020a) and 5.6% in Chen et al. (2020c). However, the nonlinear projection head typically consists of only one fully-connected (fc) layer and one ReLU activation layer. Adding such a shallow MLP projection head is often not so effective in supervised learning regime. So the question arises, *why adding a nonlinear projection head is so important for contrastive instance learning*?

In this paper we attempt to answer this question by designing experiments to explore different aspects of the nonlinear projection head in MoCov2 during unsupervised pre-training. First, as shown in Table 1 (a), removing the projection head g significantly degrades the classification accuracy compared to baseline ($64.4 \rightarrow 59.2$). This is consistent with observations made in (Chen et al., 2020a;c). Next, we initialize the nonlinear projection head g with a uniform distribution and freeze its parameters during the unsupervised pre-training. Interestingly, as we can see in Table 1 (b), we obtain better representations compared to removing the projection head (62.9 vs 59.2). This indicates that the nonlinear projection head is useful beyond its learning capability offered by the extra two layers. In fact, it is the nonlinear transformation itself that somehow benefits the learning process even if the parameters of this transformation is randomly initialized. To strengthen our observation, we investigate two other model variations. We first deepen the nonlinear projection head with more hidden layers, i.e., fc $\rightarrow$ ReLU $\rightarrow$ fc $\rightarrow$ ReLU which in theory adds more learning capacity. As shown in Table 1 (c), this seems not to bring extra benefits (63.9 vs 64.4). We then narrow the nonlinear projection head by reducing the embedding dimensionality, e.g, $2048 \rightarrow 128$ for a ResNet50. As shown in Table 1 (d), such a drastic dimension reduction indeed lowers the performance compared to baseline, but still outperforms setting (a) by a large margin (62.2 vs 59.2). With these insights we are more confident to conclude: it is the transformation of the projection head, which separates the pooled convolutional features from the final classification layer, that helps the representation learning. But why is such separation beneficial?

We argue that the nonlinear projection head acts as a filter separating the information-rich features useful for downstream tasks (i.e. color, rotation, or shape of objects) from the shallow, more discriminative features that are more useful for the contrastive loss. This conjecture was previously introduced in (Chen et al., 2020a) and verified by using features to predict transformations applied during the pre-training. In this work, we provide further visual evidence to support this hypothesis.

Inspired by deep image prior (DIP) Ulyanov et al. (2018), we perform feature inversion to obtain natural pre-images. By looking at the natural pre-image, we can diagnose which information is lost and which invariances are gained by the network. Given a fixed random noise $\mathcal{Z}$ and a reconstruction network $\mathcal{R}$, a realistic natural image can be generated by

$$x' = \mathcal{R}(\mathcal{Z}).$$

Here, $\mathcal{Z}$ is sampled from a uniform noise between 0 and 0.1. $\mathcal{R}$ is an U-Net architecture with skip connections. In order to optimize $\mathcal{R}$, we formulate the inverse problem as an image reconstruction problem by forcing x' to be close to a given input image x in the embedding space of the encoder network f. The objective can be minimized by

$$L = \min E(f(x'), f(x)),$$

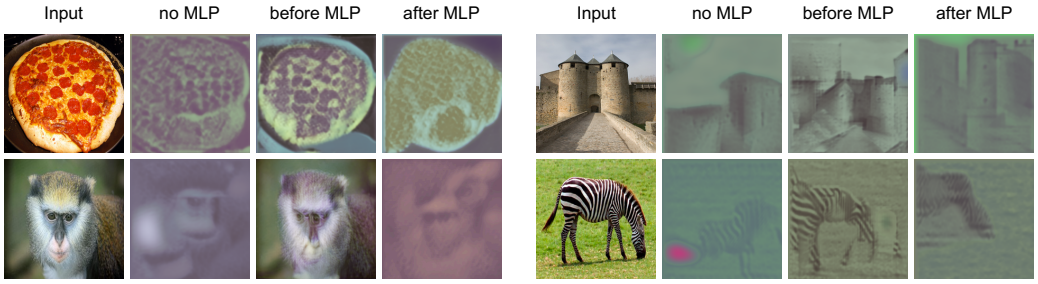
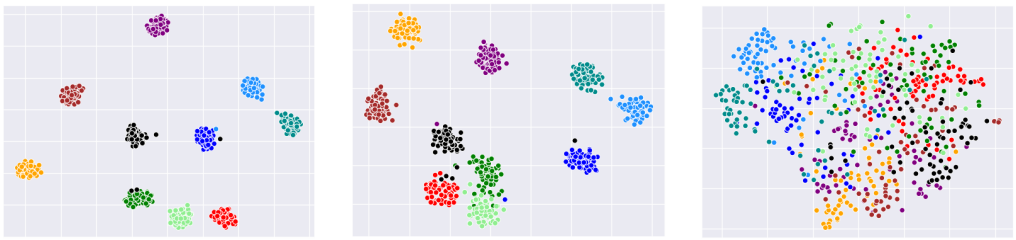


Figure 2: Visualization of feature inversion results by DIP (Ulyanov et al., 2018). We show four examples, each example contains: (a) original image, (b) reconstructed image using h from a model trained without MLP projection head, (c) reconstructed image using h from a model trained with MLP projection head and (d) reconstructed image using z from a model trained with MLP projection head. Best viewed in color and digitally.



(a) (b) (c)

Figure 3: t-SNE plots from 10 randomly chosen ImageNet classes where each class is represented by 64 images. We use the 1000-dim features from a ResNet50 model trained on ImageNet in supervised manner. The three sub-figures use the same images but under different augmentations: (a) no augmentation; (b) weak augmentation; and (c) strong augmentation. Best viewed in color.

where L2 distance is adopted. Specifically, we invert the features before and after the nonlinear projection head, h and z respectively (both features has a dimension of 1×2048). By checking on the quality of feature inversion, we can find out the information contained within the activations of a pre-trained network. We refer the readers to (Ulyanov et al., 2018) for more details.

As we can see in Figure 2, using features before MLP projection head gives the best reconstruction result. Even though we use globally pooled features without spatial dimension, they are able to generate decent image reconstructions, maintaining most color, shape, location and orientation information. However, features learned without projection or features after projection head only preserve the most discriminative information to make classification. This observation supports our claim that layers close to loss computation will lose information due to invariances to data transformations induced by the contrastive loss.

4 Strong Data Augmentation

Data augmentation is an important regularization technique in training most deep learning models. In the supervised setting, people have applied a range of augmentation techniques going back to the seminal work of AlexNet (Krizhevsky et al., 2012) which used random flipping, crop-and-resize to more recent and sophisticated techniques such as cutout (DeVries & Taylor, 2017) and autoAugment (Cubuk et al., 2019).

Despite the evolving augmentation techniques and mounting complexity of them, empirical experience shows that augmentation should not be too strong in the supervised setting as it will have an adverse effect on the network training. For example, the augmentation stack in MoCov2 unsupervised pre-training stage is far more aggressive than the augmentation stack used in its linear probe

Table 2: Investigation on different data augmentation settings' effect on MoCov2 versus supervised training. The setting of color jittering strength follows SimCLR (Chen et al., 2020a). Numbers in second line of each row indicate the performance difference compared to baseline.

(a) Random Cropping

Method	Cropping strength (min/max size)		
	baseline 20%/100%	medium 20%/50%	extreme 2%/10%
MoCo v2	64.4	63.7 −0.7	47.1 −17.3
Super-vised	75.5	74.8 −0.7	52.0 −23.5

(b) Color Jittering

Method	Color jittering strength			
	from weak to strong			
	1/8	1/4	1/2	1
MoCo v2	63.1	63.9 +0.8	64.2 +1.1	64.3 +1.2
Super-vised	75.8	75.7 −0.1	75.6 −0.2	74.6 −1.2

stage (Chen et al., 2020c). In this section, we try to explain *why unsupervised contrastive instance learning can benefit from strong augmentation more effectively than supervised setting*.

We hypothesize that the exact instance discrimination learning technique plays an important role. Without class label information, each image becomes its own class (or as in Tian et al. (2019), several views of the same object/scene become its own class) and as a result, there are no clear class boundaries like they exist in supervised training. Let us take FixMatch (Sohn et al., 2020) as an example. In Figure 1 of their paper, a strongly augmented horse image can look like a cow and easily confuse the model. To address this concern, the authors in Sohn et al. (2020) propose to use label-propagation from the model's prediction on a series of weakly augmented images (but not the original images) to supervise its predictions on the strongly augmented counterparts. Because strong data augmentation would probably break the class boundary and hinder the model from learning class labels effectively.

In order to see clearly how this hurts supervised learning, we provide a t-SNE visualization (van der Maaten & Hinton, 2008) in Figure 3 to show how the class boundaries in ImageNet break down when strong augmentation is applied, i.e., images under no/weak augmentations are separated nicely but collapse into one cluster under strong data augmentation. We term this phenomenon as *semantic label shift problem*. However, this phenomenon does not apply to instance discrimination since there are no clear semantic class boundaries. We could use strong data augmentation without worrying about the semantic label shift problem.

We also want to justify the hypothesis quantitatively. In Table 2a and Table 2b, we show how different random cropping and color jittering settings can affect MoCov2 training and supervised training. From Table 2a, we can see that as the cropping augmentation becomes more "extreme", performance on both supervised learning and MoCov2 degrade. We think this is due to the fact that MoCov2 is learning occlusion invariant features and cropping augmentation is essential in this process. This has been demonstrated recently in (Purushwalkam & Gupta, 2020) that contrastive instance learning is good at occlusion invariance. Despite this drawback for instance discrimination, MoCov2 still suffers less than its supervised counterparts. From Table 2b, we can see a different trend where stronger color jittering benefits MoCov2 but hurts supervised learning. This is consistent with the findings from SimCLR (Chen et al., 2020a).

5 Negative Samples

Recall from Equation. (1), computing contrastive loss requires sampling negative pairs, which is essential to avoid learning collapsed representations (Grill et al., 2020). Recently, Wu et al. (2018); He et al. (2020); Chen et al. (2020a) have empirically shown that using a large number of negative samples is beneficial to learn good features in contrastive instance learning. However, one needs to design sophisticated mechanisms to store the negative examples due to hardware limitation, as well as ways to update them. So we ask, *is it possible to use less negative examples during contrastive loss computation without performance degradation*?

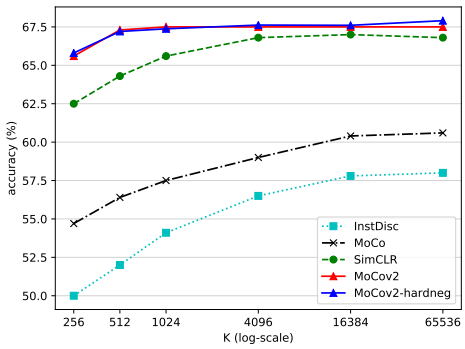


Figure 4: Comparison of different algorithms with varying number of negative samples. K denotes the number of negative samples. We show that MoCov2 performs the same ranging from $K = 512$ to $K = 65536$.

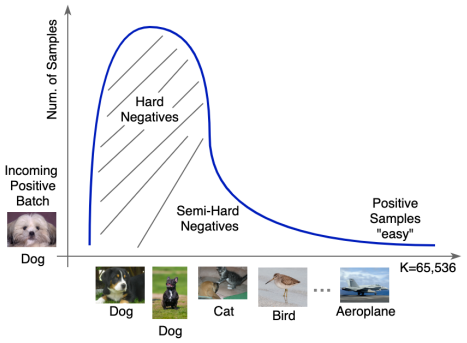


Figure 5: Illustration of how we perform dynamic hard-negative sampling on MoCov2 during training time for each mini-batch. We sort the entire dynamic queue based on similarity with incoming positive samples. We use positive skew-normal distribution to select more of hard and semi-hard samples and less of "easy" positive samples.

Quantity of Negative Samples: In order to answer the question, we first run experiments on recent contrastive instance learning approaches using different number of negative examples to see the effects. We choose InstDisc (Wu et al., 2018), MoCo (He et al., 2020), SimCLR (Chen et al., 2020a) and MoCov2 (Chen et al., 2020c) as illustrating methods. Although they use negative examples in different manners, e.g., memory bank in InstDisc, queue in MoCo and MoCov2, and large mini-batch in SimCLR, what they have in common is using a large number of negative examples. We adopt their official released code and follow the same hyper-parameter setting. All the models here are trained 200 epochs in terms of fair comparison.

We report the numbers in Figure 4 and have several interesting findings. First, the performance of InstDisc and MoCo drops when number of negatives decreases, which is consistent with the conclusion in previous literature (Wu et al., 2018; He et al., 2020; Chen et al., 2020a). However, for SimCLR and MoCov2, the performance drop is not obvious when using fewer negative samples. In fact, MoCov2 performs consistently with respect to the number of negatives. Its accuracy on ImageNet stays the same ranging from $K = 512$ to $K = 65536$, K denotes the number of negative samples. Figure 4 does indicate that the *quantity* of negative samples during contrastive loss computation has little impact on final linear probe performance.

Despite the accuracy on ImageNet stay the same for using less negative samples, it is possible that the learned features may not transfer well to downstream tasks. Here, we also perform a sanity check on the learned representations' transferability on object detection and follow the setting of (He et al., 2020). Evaluation is performed on the PASCAL VOC test2007 set. We find that AP stay the same no matter the features are learned using 512 or 65536 negative samples, which indicates that negative samples has little impact on transfer learning as well.

Quality of Negative Samples: We ask ourselves another question - What if we alter the *quality* of negative samples? Here, we perform dynamic hard-negative mining in MoCov2 where for an incoming batch of positive images we sample hard-negatives from the dynamic queue (instead of taking all samples from the queue as negatives). Specifically, during unsupervised training for a positive batch of images B and an existing dynamic queue Q, we re-arrange Q for each batch based on cosine-similarity of query-branch feature q and the latent features in Q. After re-arrange, we have similar embeddings at top of the queue, least at the bottom. We mine hard and semi-hard negative exemplars from dynamic Q using a skew-norm distribution. Note, we over-sample in order to keep the overall Q size same as MoCov2. The intuition for this experiment: each image is treated as a distinct class of its own in instance discrimination learning. Particular to ImageNet, since we are not using labels, for a specific dog image all other dog images (possibly cats looking like dogs) are considered hard-negative samples as shown in Figure 5. Contrasting among these hard-negative samples intuitively should improve the learned representation.

Table 3: Investigation of good practices' impact on the number of negative examples (K) being used in computing contrastive loss. We show that MLP projection head and momentum encoder are essential to keep the performance when using fewer negative samples.

Method	K=512	K=65536
MoCov2	67.3	67.5
(a) without MLP projection head	61.7	63.6
(b) without GaussianBlur	65.5	66.4
(c) without cosine learning schedule	67.2	67.3
(d) with smaller momentum (0.5) in momentum encoder	59.8	64.5

However, as shown in Figure 4 (blue line), in spite of over-sampling hard-negative samples, performance of MoCov2 does not change for various sizes of Q. Note that by only including hard and semi-hard negative samples, the training could collapse as this leads to an intractable optimization problem (Schroff et al., 2015). To prevent this, we also sampled positives (as shown in Figure 5). The use of skew-normal distribution easily allows us to do this. We also tried various other sampling strategies like random sampling, adaptive sampling, but the final linear probe performance does not out-perform MoCov2.

To summarize, in spite of *quantity* and *quality* of negative samples in queue Q, MoCov2 performance is stable. Naturally, we ask why? We know that MoCov2 is an improved version of MoCo with good practices, like adding MLP projection head, adding GaussianBlur data augmentation and using cosine learning schedule. Hence, the question now becomes which good practice(s) eliminate the need of using large number of negatives and why?

Good Practices Here, we perform ablation studies to see the contribution from each good practice. From Table 3, we can see that the MLP projection head and momentum encoder have the biggest impact (row a and d). Without these two techniques, the performance quickly drops as the number of negative samples decreases. In terms of MLP projection head, it can preserve more information, like color, rotation and shape, as we show in Section 3. Such information contains generic low/mid-level features that can help discriminate among instances during loss computation, which effectively reduce the need of using large number of negative samples. This also explains why SimCLR is robust to number of negatives given its performance using $K = 4096$ and $K = 65536$ is similar. In terms of momentum encoder, it is essential because using a slowly changing network can ensure the comparisons between negative and query are consistent, which avoids degenerated solution. We suspect the large performance drop of SimCLR from $K = 4096$ to $K = 256$ is due to the lack of mechanisms like momentum encoder[1]. Using stronger data augmentation and cosine learning schedule are helpful to keep the performance when using fewer negative samples, but not the key.

Note that, we are not claiming contrastive loss computation does not need negative samples, because it does in order to avoid learning collapsed representations. Our finding is that by using appropriate good practices, we do not need a huge number of negative samples during loss computation. This can simplify the framework design and potentially get more of the community interested in contrastive instance learning. Furthermore, we want to emphasize the importance of good practices because they can generalize beyond contrastive instance learning. For example, recent state-of-the-art SSL approaches BYOL (Grill et al., 2020) and SwAV (Caron et al., 2020) are consistency based and clustering based respectively, but they adopt good practices like MLP projection head, strong data augmentation and momentum encoder to achieve good performance.

6 Related Work

Contrastive instance learning Instance discrimination is a classification task where each image in a dataset is considered to be in its own class (Dosovitskiy et al., 2014). This task can either learn from both positive and negative pairs (Wu et al., 2018; He et al., 2020) or simply learn from positive pairs through consistency (Grill et al., 2020). Recently, most work on instance discrimination use contrastive learning to train the network due to its simplicity and effectiveness. He et al. (2020); Chen et al. (2020a); Caron et al. (2020) have been particularly impressive as these self-supervised models do better than supervised variants (He et al., 2016) on downstream tasks.

[1]SimCLRv2 (Chen et al., 2020b) adopts momentum encoder, which would be a good example to support our claim. However, their official released code does not include momentum encoder at this moment.

Interestingly, we find that part of the success of recent self-supervised learning methods actually come from using good practices. Inspired by (Bachman et al., 2019; Cubuk et al., 2019), SimCLR (Chen et al., 2020a) adopts nonlinear projection head and stronger data augmentation to learn better feature representations. Then, MoCov2 (Chen et al., 2020c) incorporates these good practices with longer training, and significantly outperform MoCo (He et al., 2020). Followed by this, SimCLRv2 (Chen et al., 2020b) uses more MLP heads and momentum encoder in He et al. (2020) to again refresh the state-of-the-art. Recently, Caron et al. (2020) show that clustering based representation learning can benefit from aforementioned good practices as well. Furthermore, consistency based methods like BYOL (Grill et al., 2020) also used MLP head and momentum encoder to achieve promising results, which suggests that these good practices can be generalized beyond contrastive learning paradigm.

Hence in this work, we deviate from proposing new pretext tasks or new constraints. Instead we analyze the effect of good practices in self-supervised learning and unravel the mystery behind the recent success of contrastive instance learning methods.

Insights and analysis With this ever increased interest in contrastive instance learning, there are only few work that shed light on *why* they work so well. Wang & Isola (2020) identifies two key properties, alignment and uniformity, related to the contrastive loss and empirically confirms the strong agreement between both metrics and downstream task performance. Tian et al. (2020) is a follow-up work of Contrastive Multiview Coding (Tian et al., 2019) which studies the influence of different view choices in contrastive learning. Their analysis suggests an "InfoMin principle" which inspires a task-dependent representation learning strategy. Zhao et al. (2020) investigates the problem of what makes instance discrimination pre-training good for transfer learning. Interestingly, they find that the low-/mid-level representations matter the most for object detection transfer, not high-level representations. Purushwalkam & Gupta (2020) demystifies the gains from contrastive learning by investigating the invariances encoded in the learned visual representations. Their results indicate that a large portion of recent gains come from occlusion invariances. Recently, Falcon & Cho (2020) analyze leading SSL approaches and find that despite different motivations, they are special cases under one unified framework. All aforementioned work provide great insights from different perspectives, while we go with a new one: good practices.

Good practices in deep learning Deep models are known to be sensitive to training techniques. A small change in network architecture or learning schedule can significantly impact the final performance. Numerous good practices have been proposed over the last decade and lead to the current deep learning era. The great success of AlexNet (Krizhevsky et al., 2012) benefits from using color augmentation, Dropout layer and ReLU activation. Gulrajani et al. (2017) stabilizes the training of generative adversarial networks (GAN) by introducing gradient penalty, and makes GAN accessible to more people. He et al. (2018) ensembles multiple training procedure refinements in image classification, like learning rate warmup, no bias decay, label smoothing, etc., so that stronger backbones can be trained and transferred to downstream tasks. Similarly, Bochkovskiy et al. (2020) combines good practices in object detection over the years and builds a more accurate and faster object detector YOLOv4. Hence, good practices are important, sometimes even essential, for deep network learning. Furthermore, it is also beneficial to understand how they work, which could be inspiration for further improvement.

7 Conclusion

In this paper, we dive into the details of recent contrastive instance learning methods and unravel the mysteries behind their success. We emphasize that good practices are an integral part of the algorithms. For example, MoCo cannot converge without momentum encoder. In particular, we investigate on MLP projection head and find that it acts like a filter, separating the information-rich features useful for downstream tasks from the more discriminative feature that are suitable for minimizing the contrastive loss. We also find that instance discrimination task does not suffer from the semantic label shift problem presented in supervised learning. Finally, by analyzing both quantity and quality of negative samples, we make an interesting observation that good practices during training eliminate the need of using large number of negatives. We hope our empirical evidence can provide insights and better understanding of recent progress, as well as advance future development of self-supervised representation learning.

REFERENCES

Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *Proceedings of the IEEE international conference on computer vision*, pp. 37–45, 2015.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning Representations by Maximizing Mutual Information Across Views. In *NeurIPS*. 2019.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv preprint arXiv:2004.10934*, 2020.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132–149, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *arXiv preprint arXiv:2006.09882*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*. 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. *arXiv preprint arXiv:2006.10029*, 2020b.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved Baselines with Momentum Contrastive Learning. *arXiv preprint arXiv:2003.04297*, 2020c.

E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation strategies from data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 113–123, 2019.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative Unsupervised Feature Learning with Convolutional Neural Networks. In *NeurIPS*. 2014.

William Falcon and Kyunghyun Cho. A Framework For Contrastive Self-Supervised Learning And Designing A New Approach. *arXiv preprint arXiv:2009.00104*, 2020.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *arXiv preprint arXiv:2006.07733*, 2020.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs. In *NeurIPS*. 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*. 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*. 2020.

Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of Tricks for Image Classification with Convolutional Neural Networks. *arXiv preprint arXiv:1812.01187*, 2018.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning. *arXiv preprint arXiv:2004.11362*, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NeurIPS*. 2012.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pp. 69–84. Springer, 2016.

Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5898–5906, 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Senthil Purushwalkam and Abhinav Gupta. Demystifying Contrastive Self-Supervised Learning: Invariances, Augmentations and Dataset Biases? *arXiv preprint arXiv:2007.13916*, 2020.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What Makes for Good Views for Contrastive Learning? *arXiv preprint arXiv:2005.10243*, 2020.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep Image Prior. In *CVPR*. 2018.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *IMLR*, 2008.

Tongzhou Wang and Phillip Isola. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. *arXiv preprint arXiv:2005.10242*, 2020.

Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination. In *CVPR*. 2018.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.

Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606*, 2020.