Learning from Long-Tailed Data with Noisy Labels

Shyamgopal Karthik
IIIT Hyderabad
shyamgopalkart@gmail.com

Jeromve Revaud
Naver Labs Europe
jerome.revaud@naverlabs.com

Boris Chidlovskii
Naver Labs Europe
boris.chidlovskii@naverlabs.com

Abstract

Class imbalance and noisy labels are the norm rather than the exception in many large-scale classification datasets. Nevertheless, most works in machine learning typically assume balanced and clean data. There have been some recent attempts to tackle, on one side, the problem of learning from noisy labels and, on the other side, learning from long-tailed data. Due to this separation, the proposed solutions often underperform when both assumptions are violated. In this work, we present a simple two-stage approach based on recent advances in self-supervised learning to treat both challenges simultaneously. It consists of, first, task-agnostic self-supervised pre-training, followed by task-specific fine-tuning using an appropriate loss. Most significantly, we find that self-supervised learning approaches are effectively able to cope with severe class imbalance. In addition, the resulting learned representations are also remarkably robust to label noise, when fine-tuned with an imbalance- and noise-resistant loss function. We validate our claims with experiments on CIFAR-10 and CIFAR-100 augmented with synthetic imbalance and noise, as well as the large-scale inherently noisy Clothing-1M dataset.

1. Introduction

Deep Neural Networks (DNNs) have been remarkably successful when trained under supervision of large-scale labeled data. However, this success has hinged upon two strong yet implicit assumptions: (i) data is balanced, *i.e*. there are equal number of samples for all categories; and (ii) all annotated labels are clean and reliable. In practice, unfortunately, these assumptions are incredibly difficult and expensive to meet. In fact, the price to collect and annotate by human annotators a large-scale dataset such as ImageNet is immense [27]. Conversely, it is now clear that collecting large-scale datasets can be cheap and fast when affording to violate these two assumptions [3, 21, 26, 30, 36]. It is therefore desirable to conceive learning algorithms that can handle imbalance and noise *simultaneously*.

Class imbalance and noisy labels both pose significant challenges. A vast amount of research has looked into mitigating the impact of these aspects separately. Recent methods coping with noisy labels apply different techniques ranging from sample selection [16, 19] to label correction [1, 33] as well as noise-aware losses [7, 15, 29, 37]. To learn effectively from long-tailed distributions, some works have proposed to modify the sampling algorithm to ensure all classes are represented equally [8, 23], modify the loss function [31], or perform a post-hoc correction [20]. However, existing methods designed to learn from noisy labels assume a balanced class distribution and, conversely, methods tailored to learn from long-tailed class distributions assume labels to be clean.

We argue that such a separation is artificial since label noise and long-tailed class distributions occur simultaneously in real-world datasets. For instance, the Clothing1M dataset [40], collected automatically from shopping websites, contains an estimated amount of 38.5% incorrect labels and its most populated class contains almost 5 times more instances than the smallest one (this ratio is denoted as the *class imbalance ratio*).

To address both class imbalance and label noise in deep learning, we propose to split the training procedure in two stages: *representation learning* and *classifier training*. We first pre-train the model in a self-supervised manner by discarding the training labels. This is followed by fine-tuning the model (*i.e*. learning a robust classifier) using the noisy and long-tailed class labels. This procedure is inspired by recent findings in semi-supervised learning where it was discovered that self-supervised pre-training leads to state-of-the-art performance [10]. We experiment with diverse formula-

tions of recently proposed self-supervised learning methods, namely SimCLR [9], Barlow Twins [42], BYOL [14] and SimSiam [11]. We find that all these methods are able to learn high-quality representations even when the samples are drawn from strongly imbalanced distributions.

In the second stage, the self-supervised model is fine-tuned using the noisy labels. To cope with class imbalance and noise, we adopt a simple solution based on two strong baselines. Namely, we show that a combination of the Logit Adjustment loss [31], a classification loss adapted to long-tailed data, and the SuperLoss [7], a generic loss robust to label noise, can be used to fine-tune a classifier that is robust to both class imbalance and label noise.

2. Method

Our approach is inspired by recent advances in semi-supervised learning [10]. It leverages available data in, first, task-agnostic and, second, task-specific ways. Given an image dataset with possible class imbalance and noisy labels, we first use an augmentation invariance criterion to pre-train a model in a self-supervised manner. Second, we fine-tune this representation using a loss function tailored for long-tailed data and noisy labels.

Self-Supervised Pre-training Stage. The first stage of our approach consists of pre-training the model in a self-supervised manner, thereby discarding the instance labels. Since one of our goals is to study the impact of class-imbalanced distributions on different self-supervised techniques, we select a diverse subset of those methods. Namely, we experiment with SimCLR [9] (uses positive and negative pairs), BYOL [14] (only positive pairs and momentum encoder), SimSiam [11] (same as BYOL but no momentum encoder) and Barlow Twins [42] (none of the previous). Perhaps surprisingly, as recent attempts to explain the success of these self-supervised approaches assume batch-level balanced data [34], we find that all of them work well even under severe class imbalance (see Section 3).

All these methods use Siamese networks where each image x is augmented twice. The two augmented views are fed to an *encoder* network (a ResNet [18]) and then transformed with a MLP *projection head* composed of 2 or 3 fully-connected (FC) layers. During training, obtained representations are fed to the contrastive loss in the case of SimCLR or to a redundancy reduction loss for Barlow Twins, and to a prediction head followed by similarity losses in the cases of BYOL and SimSiam (see Supplementary material for more details.)

Fine-tuning stage. Fine-tuning is a common way to adapt a task-agnostic pre-trained network for a specific task, which consists of learning with noisy labels. We follow recent strategies for semi-supervised learning [10, 11, 14]. We freeze the entire encoder network during finetuning and only train the MLP projection head. The head can be trained entirely or partially. In the latter case, we fine-tune the model from a middle layer of the projection head.

Loss functions. We leverage two recently proposed loss functions during finetuning to ensure robustness against both label noise and class imbalance. These are chosen for their simplicity and effectiveness, as well as for the fact that they can be easily combined. The first one, the logit adjustment loss [31], is a modified version of the Cross-Entropy loss that can handle class imbalance. Given a model f_θ and observed class distribution π_y which predicts logits $f(x)$ for a sample x, the logit adjustment corrects the logits as follows:

$$f^*(x) = f(x) + \log(\pi_y). \tag{1}$$

Taking softmax over the adjusted logits, the cross-entropy loss can be applied for classification of x:

$$L_{\text{LA}} = -\log\left(\frac{\exp(f^*_y(x))}{\sum_{y'}\exp(f^*_{y'}(x))}\right). \tag{2}$$

The second loss is the SuperLoss [7], a generic loss for curriculum learning. Even though this loss is not primarily meant to handle noise, recent works suggest that curriculum learning has strong noise-resistant abilities [7, 39]. Its effect is to downweight the contribution of hard samples (*i.e.* those having a higher loss value), effectively preventing the memorization of noisy labels. Given the loss L_{LA} from Eq. (2), the SuperLoss computes a new loss as follows:

$$L_{\text{LA+SL}}(L_{\text{LA}}, \sigma^*) = (L_{\text{LA}} - \tau)\sigma^* + \lambda(\log\sigma^*)^2, \tag{3}$$

where λ is a regularization trade-off and σ^* corresponds to a per-sample confidence whose optimal value can be computed in closed-form as:

$$\sigma^*_\lambda(L_{\text{LA}}) = \exp\left[-W\left(\frac{1}{2}\max\left(\frac{L_{\text{LA}}-\tau}{\lambda}, \frac{2}{e}\right)\right)\right], \tag{4}$$

where W stands for the Lambert W function, τ is the expected loss for the "average" sample and is used to separate the easy samples from the hard samples.

3. Experiments/Results

Datasets. We evaluate the proposed methodology on two standard benchmarks with simulated label noise and class imbalance, CIFAR-10 and CIFAR-100, and one large-scale, real-world dataset, Clothing1M. CIFAR datasets consist of 32×32 color images composed of 10 and 100 classes, respectively. Each dataset contains 50,000 train and 10,000 test images. For both CIFAR datasets, we simulate label noise by replacing the labels for a certain fraction of the training samples with labels chosen from a uniform distribution.

Clothing1M contains 14 classes with 1 million 256×256 train images collected from online shopping websites with labels generated using surrounding text. The proportion of annotation errors is estimated at 38.5% [40].

Class imbalance. Following prior work on class imbalance, we create imbalanced versions of CIFAR-10 and CIFAR-100 and down-sample the number of samples per class by following the Exponential profile [4, 12] with imbalance ratio $\gamma = max_y p(y)/min_y p(y)$. We test our method on CIFAR-10 with $\gamma = 50$ and $\gamma = 100$. For $\gamma = 100$, the smallest class then comprises $5000/\gamma = 50$ training instances. When we further apply 90% of label noise, only 5 of them keep their original labels while the other 45 get their labels switched to random ones (possibly keeping the same labels). Note that the proportion of incorrect labels for CIFAR-10 needs to be strictly less than 90%, as under 90% actual noise the true labels cannot possibly be recovered [25]. In the case of CIFAR-100, we test our method with $\gamma = 5$ and $\gamma = 10$. In the latter case, the smallest class has $500/\gamma = 50$ instances, which matches the case of CIFAR-10 with $\gamma = 100$. For Clothing-1M, which has 14 classes, we also experiment with long-tailed versions with $\gamma = 50$ and $\gamma = 100$. The implementation details are mentioned in the appendix.

Baselines. We compare our approach to several baselines and state-of-the-art approaches. First, we compare against strong baselines consisting of training a network from scratch in a single stage using one of the aforementioned specialized losses: the Logit-Adjusted loss [31], the SuperLoss [7], and our proposed combination of both losses in order to deal jointly with class imbalance and label noise. For reference, we also compare against a standard Cross-Entropy baseline.

Finally, we compare against DivideMix [25] and ELR [29], two state-of-the-art methods which have both shown excellent robustness to label noise. Note that these methods require significant modifications compared to the standard learning procedure used by baselines and our approach, such as network ensembling, weight averaging and mix-up data augmentation [43]. We use their default implementations available, and we adapt these to the long-tailed settings.

3.1. CIFAR experiments

Fine-tuning losses. We first study the impact of the imbalance- and noise-tailored losses considered in Section 2 during finetuning of the two-stage learning process. Namely, we consider the 4 following configurations: CE, CE+SL, LA, LA+SL where CE and LA respectively refers to the Cross-Entropy and Logit-Adjusted losses, and "+SL" denotes applying SuperLoss on top of another loss [7]. For the SuperLoss hyper-parameters, we always use a fixed threshold $\tau = \log(C)$ and we set the regularisation parameter to $\lambda = 4$. Results are presented in Figure 1 in terms of absolute improvement compared to the CE loss for models pre-trained using SimSiam and Barlow Twins (BYOL and SimCLR pre-training yield similar outcomes). While we observe some variability depending on which self-supervised method is used, we find that combining LA+SL almost always achieves the best performance overall. The accuracy gain can reach over 10%, which validates the effectiveness of these two losses. Here, it is important to note that the gains from both LA and SL are essentially for free as it does not require any additional information nor additional training time. In the remainder of this section, we therefore use LA+SL as default fine-tuning losses unless stated otherwise.

Comparison with single-stage training. To measure how much two-stage self-supervised training is beneficial, we compare with networks trained from scratch in a single stage using the same losses. This can be thought of an ablation study where we remove the self-supervised pre-training while keeping other things equal (*e.g.* losses, number of epochs, etc). We compare single-stage learning results in Figure 2 with two-stage training using BYOL. In the absence of noise, we observe that single-stage training is on par or slightly better than two-stage training. However, as soon as labels get noisy, two-stage training outperforms single-stage training by a large margin (*e.g.* $+24\%$ for CIFAR-100 at 60% noise and $\gamma = 10$), even at very low noise levels. Interestingly, two-stage training requires only little more effort that single-stage training, as the second stage (*i.e.* fine-tuning) is very short (10 epochs) and only updates a few layers. This advocates for the use of SSL pre-training in imbalanced and possibly noisy situations, as it can provide state-of-the-art performance in difficult conditions.

Discussion. The poor performance achieved by ELR and DivideMix in the presence of class imbalance and noise is due to assumptions implicitly made in their design. Specifically, the regularization applied by ELR assumes that the clean samples are learnt first, followed by the noisy samples, and the regularization prevents the noisy samples from being memorized. However, in the imbalanced setting, the dominant classes are learnt first, followed by the classes on the long-tail. The regularization applied by ELR prevents the model from being able to learn the rare classes. In the case of DivideMix, there is a regularization term which encourages the model to predict a uniform class distribution, which is again violated in the long-tailed setting. In contrast, SimCLR, Barlow Twins, BYOL and particularly SimSiam are fairly robust to both the long-tailed distribution as well as the increasing label noise.

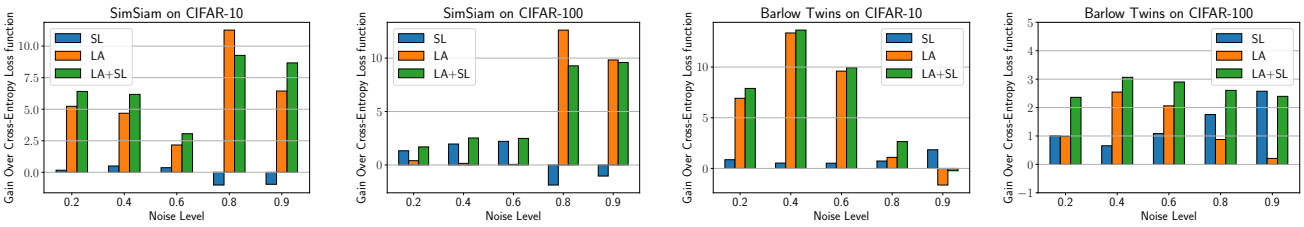


Figure 1: Accuracy gain compared to the cross-entropy fine-tuning baseline for models pretrained using SimSiam and Barlow Twins. Results are averaged over all imbalanced ratios (*i.e.* $\gamma = 50, 100$ for CIFAR-10 and $\gamma = 5, 10$ for CIFAR-100).

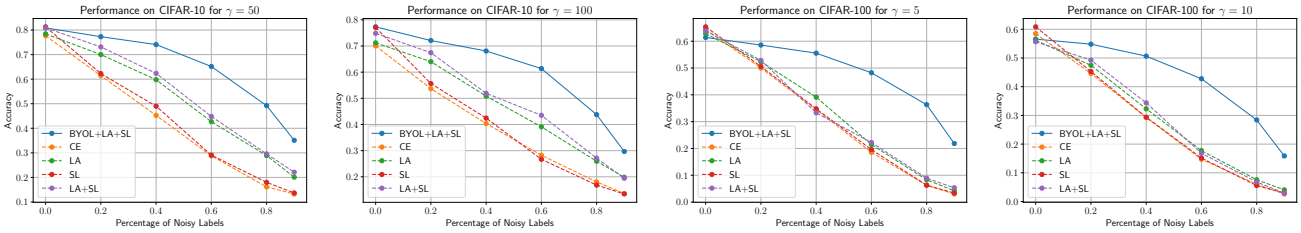


Figure 2: Comparison between two-stage training (here, BYOL with Logit-Adjusted and SuperLoss) versus single-stage methods trained with various class imbalance- and noise-resistant losses (resp. Logit-adjusted and SuperLoss) on imbalanced versions of CIFAR-10 and CIFAR-100

3.2. Clothing1M experiments

We first evaluate the effect of the pre-training duration on the Clothing1M dataset in Figure 3a. We also show the kNN-based proxy metric proposed by Chen *et al.* [11] in Figure 3b. In imbalanced settings, we observe that it only weakly correlates with the much higher performance obtained after finetuning the model. In contrast to the kNN-based accuracy that rapidly stagnates, the actual accuracy after finetuning keeps increasing after 200 epochs, even though it gradually diminishes as is expected with this type of approach [11].

Comparison with SOTA. We benchmark our SimSiam-based self-supervised method on Clothing-1M to compare performance on a real-world noise model. Here, it is important to note that DivideMix and ELR use ImageNet initialization and model ensembling, which significantly contribute to their excellent performance. In comparison, our SimSiam-based model, trained from scratch and without any tricks, reaches an accuracy only 3% below these more complex approaches. Confirming earlier findings, we observe that their performance degrades sharply when imbalance is introduced. In contrast, our SimSiam model yields very similar performance regardless of the imbalance level, in accordance with earlier findings. It significantly outperforms both DivideMix and ELR by more than 2% and 4% for $\gamma = 5$ and $\gamma = 10$, respectively. This shows that self-supervised pre-training is an effective strategy in large-scale datasets with realistic noise patterns.

Method	$\gamma = 1$	$\gamma = 50$	$\gamma = 100$
DivideMix [25]	73.9	67.1	64.9
ELR [29]	**74.2**	63.9	59.6
SimSiam+Logit+SuperLoss	71.1	**69.3**	**68.2**

Table 1: Results on Clothing-1M with varying imbalance.

4. Conclusion

In this work, we jointly tackle the problems of learning from long-tailed distributions and learning with noisy labels. Despite the vast literature that exists on both fields, these issues are usually tackled separately, often by making strong assumptions which are violated in the joint setting. Our proposed solution is inspired by recent findings in the field of semi-supervised learning. It consists of a two-stage learning process that first pre-trains the model using one of the existing self-supervised techniques, followed by fine-tuning using a robust loss function. We surprisingly find that all self-supervised methods that we experiment with are remarkably robust to class imbalance, even though they have not been explicitly designed for this use-case. Overall, the proposed approach shows excellent robustness to both class imbalance and label noise, and set a new state of the art on CIFAR and on the real-world, large-scale dataset, Clothing1M in severe noise and class imbalance conditions. We hope that this serves as strong baseline for future exploration in this topic.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*, pages 312–321. PMLR, 2019. 1

[2] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *CoRR*, abs/1911.05371, 2019. 7

[3] Artem Babenko, Anton Slesarev, Alexander Chigorin, and Victor S. Lempitsky. Neural codes for image retrieval. volume 8689, pages 584–599. Springer, 2014. 1

[4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. pages 1565–1576, 2019. 3, 6

[5] 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)*, pages 132–149, 2018. 7

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. 2020. 7

[7] Thibault Castells, Philippe Weinzaepfel, and Jerome Revaud. Superloss: A generic loss for robust curriculum learning. volume 33, 2020. 1, 2, 3

[8] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 1, 6

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. volume 119, pages 1597–1607, 2020. 2, 7

[10] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. 2020. 1, 2, 7

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *CoRR*, abs/2011.10566, 2020. 2, 4, 7

[12] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. pages 9268–9277, 2019. 3

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *CoRR*, abs/1803.07728, 2018. 7

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. 2020. 2, 7

[15] Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W. Tsang, James T. Kwok, and Masashi Sugiyama. A survey of label-noise representation learning: Past, present and future. *CoRR*, abs/2011.04406, 2020. 1

[16] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. pages 8536–8546, 2018. 1, 6

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

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 2016. 2, 7

[19] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. volume 80, pages 2309–2318, 2018. 1, 6

[20] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. 2020. 1, 6

[21] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Fei-Fei Li. Large-scale video classification with convolutional neural networks. pages 1725–1732, 2014. 1

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 7

[23] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, pages 179–186. Citeseer, 1997. 1, 6

[24] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. pages 5447–5456, 2018. 6

[25] Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. 2020. 3, 4, 6, 8

[26] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *CoRR*, abs/1708.02862, 2017. 1

[27] Yuan-Hong Liao, Amlan Kar, and Sanja Fidler. Towards good practices for efficiently annotating large-scale image classification datasets. *CoRR*, abs/2104.12690, 2021. 1

[28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[29] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. 2020. 1, 3, 4, 6, 8

[30] Dhruv Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. volume 11206, pages 185–201, 2018. 1

[31] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *CoRR*, abs/2007.07314, 2020. 1, 2, 3, 6

[32] 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*, pages 2536–2544, 2016. 7

[33] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. pages 2233–2241. IEEE Computer Society, 2017. 1, 6

[34] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. 2020. 2

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. 7

[36] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. pages 843–852, 2017. 1

[37] Deng-Bao Wang, Yong Wen, Lujia Pan, and Min-Ling Zhang. Learning from noisy labels with complementary loss functions. 2021. 1

[38] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. pages 322–330, 2019. 6

[39] Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. When do curricula work? *CoRR*, abs/2012.03107, 2020. 2

[40] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. pages 2691–2699, 2015. 1, 3

[41] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *CoRR*, abs/1708.03888, 2017. 7

[42] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *CoRR*, abs/2103.03230, 2021. 2, 7

[43] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017. 3

[44] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 7

A. Appendix

A.1. Related Work

We are not aware of any work that jointly tackles long-tailed learning and training with noise and review these two fields separately.

Long-tailed learning, *i.e.* learning with imbalanced classes, is a way to alleviate the performance degradation due to imbalance in the class distribution. Existing works in this field can be divided into three categories. First, several methods focus on changing the *data* that is given as input to the model [8, 23]. The most common strategy here is to over-sample the rare classes, or equivalently to under-sample the dominant classes. A second solution is to modify either the algorithm design or the loss function that is used to train the model. Loss functions designed for imbalanced data include the focal loss [28], the logit adjustment loss [31] as well as the label-distribution aware margin loss [4]. Their common idea is to assign a higher loss to samples from rare classes, thereby providing a stronger supervisory signal for the model to learn these classes. Thirdly, some methods modify the outputs predicted by the model. This family of post-hoc correction can either modify the threshold [31], or change the weights of the final classifier [20] usually using some normalization procedure.

Decoupling the learning procedure into representation learning and classification was proposed for long-tailed data in [20]. The key finding was that training a model using instance-balanced sampling conventionally, followed by training a classifier robust to imbalance works extremely well for long-tailed data. Our approach is very similar in spirit to this idea. However, in the presence of label noise, standard supervised learning collapses due to noise memorization. Therefore, we use self-supervised learning to learn effective representations, and train a classifier using losses that are robust to both data imbalance and label noise.

Learning with label noise is an active research field where existing approaches can be categorized into three groups. First, *label correction* methods aim to relabel the corrupted labels. They try to formulate explicit or implicit noise models to characterize the distribution of noisy and true labels [24]. However, to recover the ground-truth labels, these approaches usually require the support of a small set of clean samples. Second, *loss correction* techniques seek to modify the loss function to achieve robustness, by using pre-calculated Backward or Forward noise transition matrix [33], or combining cross entropy and reverse cross entropy [38]. Lastly, a third group of methods adopts *sample selection* to identify potentially clean samples from a noisy training dataset. MentorNet [19] introduces a data-driven curriculum learning paradigm in which a pre-trained mentor network guides the training of a student network. Co-teaching [16] trains two DNNs simultaneously, and let them teach each other with some selected samples during every mini-batch. DivideMix [25] trains two networks simultaneously and fits a Gaussian Mixture Model (GMM) on its per-sample loss distribution to divide the training samples into a labeled set and an unlabeled set.

Early Learning Regularization (ELR) [29] is a recent advance in learning with noisy labels. The key observation here is that the clean samples are learnt first, followed by the noisy samples in the later epochs. Using this insight, ELR proposed a regularization term to prevent memorization of the noisy samples. All these methods are designed with the underlying assumption that classes are balanced. As a result, this idea of separating noisy and clean samples using different techniques does not work as effectively when the

data is imbalanced, as illustrated in our results in Section 3.

A.1.1 Background on Augmentation-invariant Self-Supervision

Prior work has shown that self-supervision for visual data can be tackled in various ways. In contrast to older approaches that propose a variety of pretext tasks [44, 13, 32], recent approaches all revolves around the principle of learning invariance to random image augmentations (*e.g.* scaling, color jitter, blur, etc.) using a Siamese network architecture [6, 9, 11, 14, 17, 42]. Specifically, the goal is to maximize the similarity between the encoded representations of two augmented versions of the same image. Because this procedure can collapse to trivial solutions, different remedies have been proposed.

Earlier methods like SimCLR [9], SimCLRv2 [10] and MoCo [17], for instance, use negative samples and contrastive losses based on artificially constructed positive and negative pairs. More recent methods like BYOL [14] and SimSiam [11] have shown that it is possible to use only positive pairs. The trick is to rely on asymmetric operations such as propagating gradients through only one of the two Siamese branches. SimSiam [11], for its part, can be seen as a minimalist version of BYOL [14] without momentum encoder. Clustering-based method like SwAV [6], DeepCluster [5] or Sela [2] are instead based on trainable versions of the k-means algorithm that learns image representations leading to clusters stable against random augmentations. Lastly, Barlow Twins [42] is able to prevent collapse without considering image pairs at all nor asymmetric operations thanks to a novel loss function based on redundancy reduction. While these approaches appear very diverse, they all display excellent performance for visual representation learning in situations where data is perfectly balanced (typically, on ImageNet [35]). In this work, we show that the benefits of self-supervised pre-training extend to the imbalanced setting as well.

A.2. Implementation Details

We briefly describe the self-supervised training protocols for each considered method. In all cases, we stay as close as possible to the original protocols, only making minimal changes required to pass from ImageNet-based training to CIFAR and Clothing1M.

Backbone encoder. For all the datasets, we use a ResNet18 [18] backbone. For the CIFAR-10 and CIFAR-100, the first convolutional layer has a stride of 3×3 instead of the usual 7×7 and the first max pooling layer is removed [18]. During SSL pre-training and fine-tuning, we attach a projection head whose configuration depends on the specific SSL method, as described below.

SimCLR: We pre-train SimCLR for 1000 epochs using the Adam optimizer [22] with a learning rate of 0.001, a weight decay of 10^{-6} and a batch size of 512. For fine-tuning the model on the noisy labels, we train a linear classifier on top of the representations extracted by the encoder. For this, we train for 25 epochs using Adam with a learning rate of 0.001 and a weight decay of 10^{-6}.

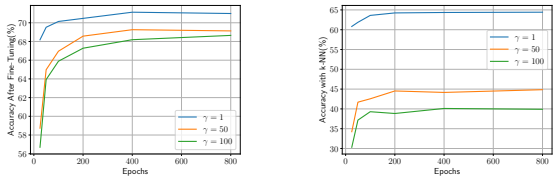
SimSiam: We follow the original SimSiam implementation, except that we use 2 FC layers in the projection head instead of 3 as we find it slightly better in our case. We pre-train the network for 800 epochs using SGD with a learning rate of $lr \times bs/256$, a base $lr = 0.03$ and a batch size $bs = 512$. We use a cosine decay schedule except during the initial warm-up period where it is scaled linearly for 10 epochs. The weight decay is set to 0.0005 and the SGD momentum is 0.9. For fine-tuning, we train the full projection head (2 FC layers) with Adam for 10 epochs and a learning rate of 0.003 without weight decay and with a batch size of 256. When the noise exceeds $\nu_{\text{SimSiam}} = 60\%$, we only fine-tune the last FC layer with a learning rate of 0.01. This strategy is very similar the method proposed in SimCLRv2 [10].

BYOL: We follow the original architecture for the projection and predictor heads [14], but we use the Adam optimizer instead of LARS [41] due to the small size of the CIFAR training sets. We use the same pre-training protocol as for SimSiam, with a base learning rate set to 0.001 and a weight decay of $1.5 \cdot 10^{-6}$. For fine-tuning, we again follow SimSiam, but this time switching from 2 FC layers to 1 FC layer at $\nu_{\text{BYOL}} = 40\%$ noise.

Barlow Twins: As for BYOL, we pre-train using the Adam optimizer instead of LARS as in [42]. The pre-training protocol is identical to SimSiam except the base learning rate is set to 0.003. The λ parameter is kept to 0.005 as in [42] but we find that setting the size of the projection head's hidden and output layers to 2048 improves performance. For fine-tuning, we again use the same protocol as for BYOL and SimSiam except for $\nu_{\text{BarlowTwins}} = 20\%$ noise.

A.3. Additional evaluations

Comparison of self-supervised methods. We present experimental results for each of the considered self-supervised methods in Fig. 4a-4c for CIFAR-10 and Fig. 4d-4f for CIFAR-100.We observe that BYOL significantly outperforms other methods in low-to-moderate noise levels by up to 3~4%. This is consistent with the overall superior results achieved by BYOL in balanced and noiseless settings compared to other self-supervised methods [11, 14]. We hypothesize that this superiority is due to the weight-averaging trick used by BYOL, which is also employed to



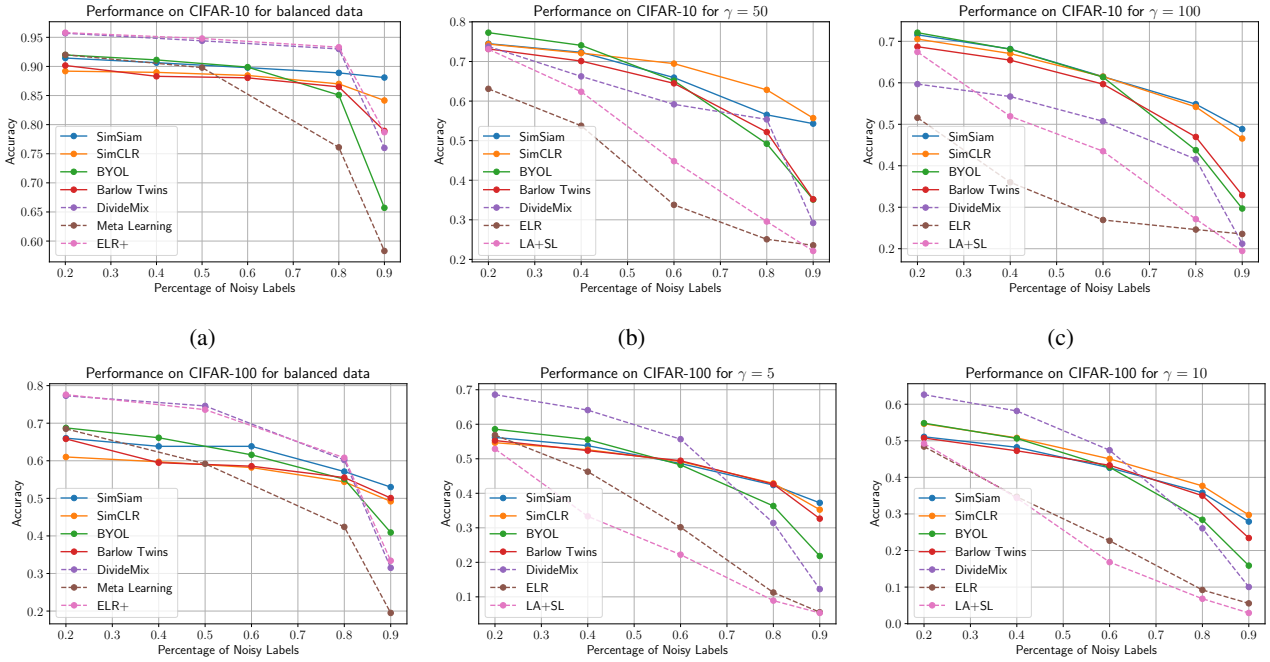
(a) (b)

Figure 3: Results for two-stage learning on the Clothing1M dataset as a function of the imbalance level and length of self-supervised pre-training (in epochs). For all γ values, both the k-NN as well as the accuracy after fine-tuning values achieve their maximum values after 400 epochs.

improve results by ELR [29]. Other methods, in comparison, performs on par at this noise regime. Conversely, we find that BYOL significantly underperforms compared to the other methods at high-noise levels. Interestingly, SimSiam, which is mostly similar to BYOL in principle except for the weight-averaging part, is either on par or significantly better than other self-supervised approaches under severe noise. Overall, we find it interesting and rather unexpected that all self-supervised approaches are robust to strong class imbalance. For instance, all approaches obtains at least 60% accuracy on CIFAR-10 with $\gamma = 100$ at 60% noise, which is unprecedented.

Comparison with SOTA. Experimental comparisons with state-of-the-art methods and other baselines are presented in Figure 4 for CIFAR-10 and CIFAR-100 at several imbalance levels. Self-supervised pre-training leads to significant gains over state-of-the-art methods at any noise level in strongly imbalanced situations (*e.g.* accuracy is 10% to 30% above for CIFAR-10 with $\gamma = 100$).

When the imbalance is moderate, self-supervised models are able to achieve reasonable performance (*e.g.* 90.1% to 92.0% for CIFAR-10 when $\gamma = 1$ and noise $\nu = 20\%$), but they do not match the fully-supervised counterparts of DivideMix [25] and ELR [29], which are able to achieve 95.7% and 95.8% in this setting, respectively. Overall, these specialized approaches still perform better as long as both the imbalance and noise level are not too severe. For instance, all self-supervised methods start outperforming DivideMix on CIFAR-100 at $\gamma = 5$ when the noise level is above 70%. More generally, we observe that the performance of the self-supervised models *degrade much less*, even when the noise is increased to 80% or 90%. In particular, we achieve a new state-of-the-art for both CIFAR-10 and CIFAR-100 at 90% noise by achieving (with SimSiam) 88.1% and 53.0% respectively in the imbalanced case. This is an improvement of 12.1% and 21.5% over DivideMix.



(a) (b) (c)

(d) (e) (f)

Figure 4: Results on CIFAR-10 and CIFAR-100 datasets as a function of symmetric noise.