MoPro: Webly Supervised Learning with Momentum Prototypes

Anonymous authors
Paper under double-blind review

Abstract

We propose a webly-supervised representation learning method that does not suffer from the annotation unscalability of supervised learning, nor the computation unscalability of self-supervised learning. Most existing works on webly-supervised representation learning adopt a vanilla supervised learning method without accounting for the prevalent noise in the training data, whereas most prior methods in learning with label noise are less effective for real-world large-scale noisy data. We propose momentum prototypes (MoPro), a simple contrastive learning method that achieves online label noise correction, out-of-distribution sample removal, and representation learning. MoPro achieves state-of-the-art performance on WebVision, a weakly-labeled noisy dataset. MoPro also shows superior performance when the pretrained model is transferred to down-stream image classification and detection tasks. It outperforms the ImageNet supervised pretrained model by $+10.5$ on 1-shot classification on VOC, and outperforms the best self-supervised pretrained model by $+17.3$ when finetuned on 1% of ImageNet labeled samples. Furthermore, MoPro is more robust to distribution shifts[1].

1 Introduction

Large-scale datasets with human-annotated labels have revolutionized computer vision. Supervised pretraining on ImageNet (Deng et al., 2009) has been the de facto formula of success for almost all state-of-the-art visual perception models. However, it is extremely labor intensive to manually annotate millions of images, which makes it a non-scalable solution. One alternative to reduce annotation cost is self-supervised representation learning, which leverages unlabeled data. However, self-supervised learning methods (Goyal et al., 2019; He et al., 2019; Chen et al., 2020a; Li et al., 2020b) have yet consistently shown superior performance compared to supervised learning, especially when transferred to downstream tasks with limited labels.

With the help of commercial search engines, photo-sharing websites, and social media platforms, there is near-infinite amount of weakly-labeled images available on the web. Several works have exploited the scalable source of web images and demonstrated promising results with webly-supervised representation learning (Mahajan et al., 2018; Sun et al., 2017; Li et al., 2017; Kolesnikov et al., 2020). However, there exists two competing claims on whether weakly-labeled noisy datasets lead to worse generalization performance. One claim argues that the effect of noise can be overpowered by the scale of data, and simply applies standard supervised learning method on web datasets (Mahajan et al., 2018; Sun et al., 2017; Li et al., 2017; Kolesnikov et al., 2020). The other claim argues that deep models can easily memorize noisy labels, resulting in worse generalization (Zhang et al., 2017; Ma et al., 2018). In this paper, we show that both claims are partially true. While increasing the size of data does improve the model's robustness to noise, our method can substantially boost the representation learning performance by addressing noise.

There exists a large body of literature on learning with label noise (Jiang et al., 2018; Han et al., 2018; Guo et al., 2018; Tanaka et al., 2018; Arazo et al., 2019; Li et al., 2020a). However, existing methods have several limitations that make them less effective for webly-supervised representation learning. First, most methods do not consider out-of-distribution (OOD) samples, which is a major source of noise in real-world web datasets. Second, many methods perform computation-heavy

[1] Code and pretrained models will be released.

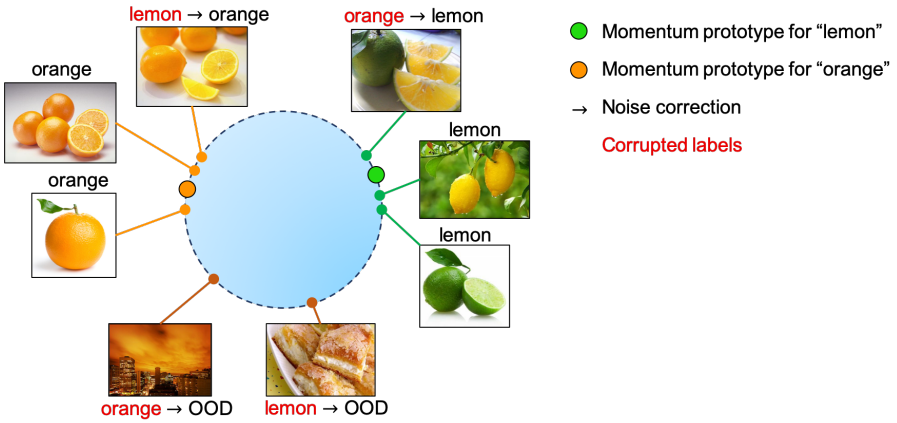

Figure 1: Illustration of the normalized embedding space learned with MoPro. Samples from the same class gather around their class prototype, whereas OOD samples are separated from in-distribution samples. Label correction and OOD removal are achieved based on a sample's distance with the prototypes.

procedures for noise cleaning (Jiang et al., 2018; Li et al., 2019; 2020a), or require access to a set of samples with clean labels (Vahdat, 2017; Veit et al., 2017; Lee et al., 2018), which limit their scalability in practice.

We propose a new method for efficient representation learning from weakly-labeled web images. Our method is inspired by recent developments in contrastive learning for self-supervised learning (He et al., 2019; Chen et al., 2020a; Li et al., 2020b) We introduce *Momentum Prototypes* (MoPro), a simple component which is effective in **label noise correction**, **OOD sample removal**, and **representation learning**. A visual explanation of our method is shown in Figure 1. We use a deep network to project images into normalized low-dimensional embeddings, and calculate the prototype for a class as the moving-average embedding for clean samples in that class. We train the network such that embeddings are pulled closer to their corresponding prototypes, while pushed away from other prototypes. Images with corrupted labels are corrected either as another class or as an OOD sample based on their distance to the momentum prototypes.

We experimentally show that:

- MoPro achieves state-of-the-art performance on the upstream weakly-supervised learning task.
- MoPro substantially improves representation learning performance when the pretrained model is transferred to downstream image classification and object detection tasks. For the first time, we show that weakly-supervised representation learning achieves similar performance as supervised representation learning, under the same data and computation budget. With a larger web dataset, MoPro outperforms ImageNet supervised learning by a large margin.
- MoPro learns a more robust and calibrated model that generalizes better to distribution variations.

2 Related Work

2.1 Webly-supervised representation learning

A number of prior works exploit large web datasets for visual representation learning (Divvala et al., 2014; Chen & Gupta, 2015; Joulin et al., 2016; Mahajan et al., 2018; Sun et al., 2017; Li et al., 2017; Kolesnikov et al., 2020). These datasets contain a considerable amount of noise. Approximately 20% of the labels in the JMT-300M dataset (Sun et al., 2017) are noisy, whereas 34% of images in the WebVision dataset (Li et al., 2017) are considered outliers. Surprisingly, most prior works have chosen to ignore the noise and applied vanilla supervised method, with the claim that the scale of data can overpower the noise (Mahajan et al., 2018; Sun et al., 2017; Li et al., 2017). However, we show that supervised method cannot fully harvest the power of large-scale weakly-labeled datasets. Our method achieves substantial improvement by addressing noise, and advances the potential of webly-supervised representation learning.

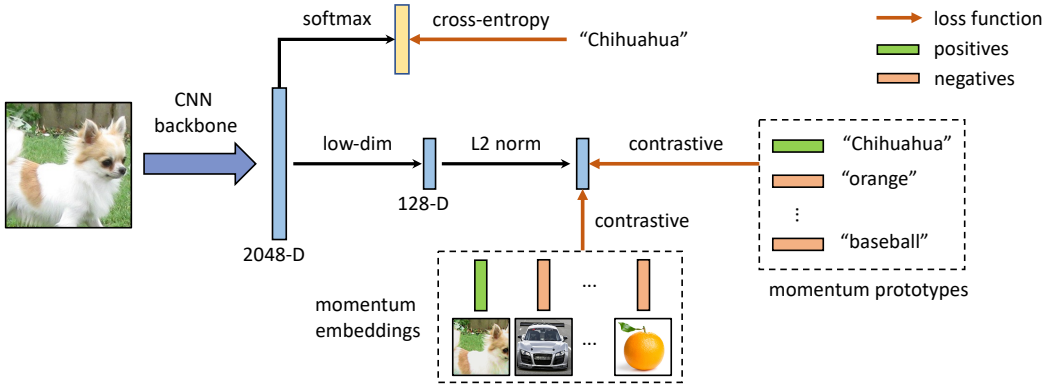

Figure 2: Proposed weakly-supervised learning framework. We jointly optimize a prototypical contrastive loss using momentum prototypes, an instance contrastive loss using momentum embeddings, and a cross-entropy loss using pseudo-labels. The pseudo-label for a sample is generated based on its original training label, the model's prediction, and the sample's distance to the prototypes.

2.2 Learning with label noise

Learning with label noise has been widely studied. Some methods require access to a small set of clean samples (Xiao et al., 2015; Vahdat, 2017; Veit et al., 2017; Lee et al., 2018; Zhang et al., 2020), and other methods assume that no clean labels are available. There exist two major types of approaches. The first type performs label correction using predictions from the network (Reed et al., 2015; Ma et al., 2018; Tanaka et al., 2018; Yi & Wu, 2019; Yang et al., 2020). The second type separates clean samples from corrupted samples, and trains the model on clean samples (Han et al., 2018; Arazo et al., 2019; Jiang et al., 2018; Wang et al., 2018; Chen et al., 2019; Li et al., 2020a). However, existing methods have yet shown promising results for large-scale weakly-supervised representation learning. The main reasons include: (1) most methods do not consider OOD samples, which commonly occur in real-world web datasets; (2) most methods are computational-heavy due to co-training (Han et al., 2018; Li et al., 2020a; Jiang et al., 2018; 2020), iterative training (Tanaka et al., 2018; Yi & Wu, 2019; Wang et al., 2018; Chen et al., 2019), or meta-learning (Li et al., 2019; Zhang et al., 2019).

Different from existing methods, MoPro achieves both label correction and OOD sample removal on-the-fly with a single step, based on the similarity between an image embedding and the momentum prototypes. MoPro also leverages contrastive learning to learn a robust embedding space.

2.3 Self-supervised representation learning

Self-supervised methods have been proposed for representation learning using unlabeled data. The recent developments in self-supervised representation learning can be attributed to contrastive learning. Most methods (He et al., 2019; Chen et al., 2020a; Oord et al., 2018; Wu et al., 2018) leverage the task of instance discrimination, where augmented crops from the same source image are enforced to have similar embeddings. Prototypical contrastive learning (PCL) (Li et al., 2020b) performs clustering to find prototypical embeddings, and enforces an image embedding to be similar to its assigned prototypes. Different from PCL, we update prototypes on-the-fly in a weakly-supervised setting, where the momentum prototype of a class is the moving average of clean samples' embeddings. Furthermore, we jointly optimize two contrastive losses and a cross-entropy loss.

Current self-supervised representation learning methods are limited in (1) inferior performance in low-shot task adaptation, (2) huge computation cost, and (3) inadequate to harvest larger datasets. We show that weakly-supervised learning with MoPro addresses these limitations.

3 Method

In this section, we delineate the details of our method. First, we introduce the components in our representation learning framework. Then, we describe the loss functions. Finally, we explain the

noise correction procedure for label correction and OOD sample removal. A pseudo-code of MoPro is provided in appendix B.

3.1 Representation Learning Framework

Our proposed framework consists of the following components. Figure 2 gives an illustration.

- A noisy training dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i$ is an image and $y_i \in \{1, ..., K\}$ is its class label.
- A pseudo-label $\hat{y}_i$ for each image $\boldsymbol{x}_i$, which is its corrected label. Details for generating the pseudo-label is explained in Sec 3.3.
- An encoder network, which maps an augmented image $\tilde{\boldsymbol{x}}_i$ to a representation vector $\boldsymbol{v}_i \in \mathbb{R}^{d_e}$. We experiment with ResNet-50 (He et al., 2016) as the encoder, where the activations of the final global pooling layer ($d_e = 2048$) are used as the representation vector.
- A classifier (a fully-connected layer followed by softmax) which receives the representation $\boldsymbol{v}_i$ as input and outputs class predictions $\boldsymbol{p}_i$.
- A projection network, which maps the representation $\boldsymbol{v}_i$ into a low-dimensional embedding $\boldsymbol{z}_i \in \mathbb{R}^{d_p}$ ($d_p = 128$). $\boldsymbol{z}_i$ is always normalized to the unit sphere. Following SimCLR (Chen et al., 2020a), we use a MLP with one hidden layer as the projection network.
- Momentum embeddings $\boldsymbol{z}'_i$ generated by a momentum encoder. The momentum encoder has the same architecture as the encoder followed by the projection network, and its parameters are the moving-average of the encoder's and the projection network's parameters. Same as in MoCo (He et al., 2019), we maintain a queue of momentum embeddings of past samples.
- Momentum prototypes $\boldsymbol{C} \in \mathbb{R}^{d_p \times K}$. The momentum prototype of the k-th class, $\boldsymbol{c}_k$, is the normalized moving-average embedding for samples with pseudo-label $\hat{y}_i = k$.

3.2 Contrastive Loss

As illustrated in Figure 1, we aim to learn an embedding space where samples from the same class gather around its class prototype, while samples from different classes are seperated. We achieve it with two contrastive losses: (1) a prototypical contrastive loss $\mathcal{L}_{\text{pro}}$ which increases the similarity between an embedding and its corresponding class prototype, $(\boldsymbol{z}_i, \boldsymbol{c}_{\hat{y}_i})$, in contrast to other prototypes; (2) an instance contrastive loss $\mathcal{L}_{\text{ins}}$ which increases the similarity between two embeddings of the same source image, $(\boldsymbol{z}_i, \boldsymbol{z}'_i)$, in contrast to embeddings of other images. Specifically, the contrastive losses are defined as:

$$\mathcal{L}^i_{\text{pro}} = -\log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{c}_{\hat{y}_i}/\tau)}{\sum_{k=1}^K \exp(\boldsymbol{z}_i \cdot \boldsymbol{c}_k/\tau)}, \qquad \mathcal{L}^i_{\text{ins}} = -\log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}'_i/\tau)}{\sum_{r=0}^R \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}'_r/\tau)}, \tag{1}$$

where τ is a temperature parameter, and $\hat{y}_i$ is the pseudo-label. We use R negative momentum embeddings to construct the denominator of the instance contrastive loss.

We train the classifier with cross-entropy loss, using pseudo-labels as targets.

$$\mathcal{L}^i_{\text{ce}} = -\log(p_i^{\hat{y}_i}) \tag{2}$$

We jointly optimize the contrastive losses and the classification loss. The training objective is:

$$\mathcal{L} = \sum_{i=1}^n (\mathcal{L}^i_{\text{ce}} + \lambda_{\text{pro}} \mathcal{L}^i_{\text{pro}} + \lambda_{\text{ins}} \mathcal{L}^i_{\text{ins}}) \tag{3}$$

For simplicity, we set $\lambda_{\text{pro}} = \lambda_{\text{ins}} = 1$ for all experiments.

3.3 Noise Correction

We propose a simple yet effective method for online noise correction during training, which cleans label noise and removes OOD samples. For each sample, we generate a soft pseudo-label $\boldsymbol{q}_i$ by combining the classifier's output probability $\boldsymbol{p}_i$ with $\boldsymbol{s}_i$, a class probability distribution calculated using the sample's similarity *w.r.t* the momentum prototypes:

$$\begin{aligned} \boldsymbol{q}_i &= \alpha \boldsymbol{p}_i + (1-\alpha)\boldsymbol{s}_i, \\ s_i^k &= \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{c}_k/\tau)}{\sum_{k=1}^K \exp(\boldsymbol{z}_i \cdot \boldsymbol{c}_k/\tau)}. \end{aligned} \tag{4}$$

where the combination weight is simply set as $\alpha = 0.5$ in all experiments.

We convert $\boldsymbol{q}_i$ into a hard pseudo-label $\hat{y}_i$ based on the following rules: (1) if the highest score of $\boldsymbol{q}_i$ is above certain threshold T, use the class with the highest score as the pseudo-label; (2) otherwise, if the score for the original label y_i is higher than uniform probability, use y_i as the pseudo-label; (3) otherwise, label it as an OOD sample.

$$\hat{y}_i = \begin{cases} \arg\max_k q_i^k & \text{if } \max_k q_i^k > T, \\ y_i & \text{elseif } q_i^{y_i} > 1/K, \\ \text{OOD} & \text{otherwise.} \end{cases} \tag{5}$$

We remove OOD samples from both the cross-entropy loss and the prototypical contrastive loss so that they do not affect class-specific learning, but include them in the instance contrastive loss to further separate them from in-distribution samples. Examples of OOD images and corrected pseudo-labels are shown in the appendices.

3.4 Momentum Prototypes

For each class k, we calculate its momentum prototype as a moving-average of the normalized embeddings for samples with pseudo-label k. Specifically, we update $\boldsymbol{c}_k$ by:

$$\boldsymbol{c}_k \leftarrow \text{Normalize}(m\boldsymbol{c}_k + (1-m)\boldsymbol{z}_i), \ \ \forall i \in \{i \mid \hat{y}_i = k\}, \tag{6}$$

where $\text{Normalize}(\boldsymbol{c}) = \boldsymbol{c}/\left\|\boldsymbol{c}\right\|_2$. The momentum coefficient m is set 0.999 in our experiments.

4 Experiments

4.1 Dataset for upstream training

We use the WebVision (Li et al., 2017) dataset as the noisy training data. It consists of images automatically crawled from Google and Flickr, using visual concepts from ImageNet as queries. We experiment with three versions of WebVision with different sizes: (1) WebVision-V1.0 contains 2.44m images with the same classes as the ImageNet-1k (ILSVRC 2012) dataset; (2) WebVision-V0.5 is a randomly sampled subset of WebVision-V1.0, which contains the same number of images (1.28m) as ImageNet-1k; (3) WebVision-V2.0 contains 16m images with 5k classes.

4.2 Implementation details

We follow standard settings for ImageNet training: batch size is 256; total number of epochs is 90; optimizer is SGD with a momentum of 0.9; initial learning rate is 0.1, decayed at 40 and 80 epochs; weight decay is 0.0001. We use ResNet-50 (He et al., 2016) as the encoder. For MoPro-specific hyperparameters, we set $\tau = 0.1, \alpha = 0.5, T = 0.8$ ($T = 0.6$ for WebVision-V2.0). The momentum for both the momentum encoder and momentum prototypes is set as 0.999. The queue to store momentum embeddings has a size of 8192. We apply standard data augmentation (crop and horizontal flip) to the encoder's input, and stronger data augmentation (color changes in MoCo (He et al., 2019)) to the momentum encoder's input. We warm-up the model for 10 epochs by training on all samples with original labels, before applying noise correction.

4.3 Upstream task performance

In Table 1, we compare MoPro with existing weakly-supervised learning methods trained on WebVision-V1.0, where MoPro achieves state-of-the-art performance. Since the training dataset has imbalanced number of samples per-class, inspired by Kang et al. (2020), we perform the following decoupled training steps to re-balance the classifier: (1) pretrain the model with MoPro; (2) perform noise correction on the training data using the pretrained model, following the method in Section 3.3; (3) keep the pretrained encoder fixed and finetune the classifier on the cleaned dataset, using square-root data sampling (Mahajan et al., 2018) which balances the classes. We retrain the classifier for 15 epochs, using a learning rate of 0.01 which is decayed at 5 and 10 epochs. Surprisingly, we also find that a vanilla cross-entropy method with decoupled classifier re-balancing can also achieve competitive performance, outperforming most existing baselines.

Method	Architecture	WebVision		ImageNet	
		top-1	top-5	top-1	top-5
Cross-Entropy (Tu et al., 2020)	ResNet-50	66.4	83.4	57.7	78.4
MentorNet (Jiang et al., 2018)	InceptionResNet-V2	70.8	88.0	62.5	83.0
CurriculumNet (Guo et al., 2018)	Inception-V2	72.1	89.1	64.8	84.9
CleanNet (Lee et al., 2018)	ResNet-50	70.3	87.8	63.4	84.6
CurriculumNet (Guo et al., 2018; Tu et al., 2020)	ResNet-50	70.7	88.6	62.7	83.4
SOM (Tu et al., 2020)	ResNet-50	72.2	89.5	65.0	85.1
Distill (Zhang et al., 2020)	ResNet-50	-	-	65.8	85.8
Cross-Entropy (decoupled)	ResNet-50	72.4	89.0	65.7	85.1
MoPro (ours)	ResNet-50	**73.9**	**90.0**	**67.8**	**87.0**

Table 1: Comparison with state-of-the-art methods on WebVision-V1.0. Numbers denote accuracy (%) on the clean WebVision-V1.0 validation set and the ILSVRC 2012 validation set. CleanNet (Lee et al., 2018) and Distill (Zhang et al., 2020) require data with clean annotations.

5 Transfer Learning

In this section, we transfer weakly-supervised learned models to a variety of downstream tasks. We show that MoPro yields superior performance in image classification, object detection, instance segmentation, and obtains better robustness to domain shifts. Implementation details for the transfer learning experiments are described in appendix C.

5.1 Low-shot image classification on fixed representation

First, we transfer the learned representation to downstream tasks with few training samples. We perform low-shot classification on two datasets: PASCAL VOC2007 (Everingham et al., 2010) for object classification and Places205 (Zhou et al., 2014) for scene recognition. Following the setup by Goyal et al. (2019); Li et al. (2020b), we train linear SVMs using fixed representations from pretrained models. We vary the number k of samples per-class and report the average result across 5 independent runs. Table 2 shows the results. When pretrained on weakly-labeled datasets, MoPro consistently outperforms the vanilla CE method. The improvement of MoPro becomes less significant when the number of web images increases from 2.4m to 16m, suggesting that increasing dataset size is a viable solution to combat noise.

When compared with ImageNet pretrained models, MoPro substantially outperforms self-supervised learning (MoCo v2 (Chen et al., 2020b) and PCL v2 (Li et al., 2020b)), and achieves comparable performance with supervised learning when the same amount of web images (*i.e.* WebVision-V0.5) is used. Our results for the first time show that weakly-supervised representation learning can be as powerful as supervised representation learning under the same data and computation budget.

Method	Pretrain dataset	VOC07					Places205				
		k=1	k=2	k=4	k=8	k=16	k=1	k=2	k=4	k=8	k=16
MoCo v2*	ImageNet	46.3	58.4	64.9	72.5	76.1	11.9	17.0	22.6	28.1	32.4
PCL v2*	ImageNet	47.9	59.6	66.2	74.5	78.3	12.5	17.5	23.2	28.1	32.3
CE (Sup.)	ImageNet	54.3	67.8	73.9	79.6	82.3	14.9	21.0	26.9	32.1	36.0
CE	WebVision-V0.5	49.8	63.9	69.9	76.1	79.2	13.5	19.3	24.7	29.5	33.8
MoPro (ours)	WebVision-V0.5	54.3	67.8	73.5	79.2	81.8	15.0	21.2	26.6	31.8	36.0
CE	WebVision-V1.0	54.5	67.1	72.8	78.4	81.4	15.1	21.5	27.2	32.1	36.4
MoPro (ours)	WebVision-V1.0	59.5	71.3	76.5	81.4	83.7	16.9	23.2	29.2	34.5	38.7
CE	WebVision-V2.0	63.0	73.8	78.7	83.0	85.4	21.8	28.6	35.1	40.0	43.6
MoPro (ours)	WebVision-V2.0	**64.8**	**74.8**	**79.9**	**83.9**	**86.1**	**22.2**	**29.2**	**35.6**	**40.9**	**44.4**

Table 2: **Low-shot image classification** on VOC07 and Places205 using linear SVMs trained on fixed representations. We vary the number of labeled examples per-class (k), and report the average mAP (for VOC) and accuracy (for Places) across 5 independent runs. WebVision-V0.5 has the same number of training samples as ImageNet. The self-supervised learning methods* are trained for 200 epochs, while other methods are trained for 90 epochs. MoPro outperforms vanilla CE pretrained on Web datasets, as well as self-supervised learning and supervised learning methods pretrained on ImageNet.

5.2 LOW-RESOURCE TRANSFER WITH FINETUNING

Next, we perform experiment to evaluate whether the pretrained model provides a good basis for finetuning when the downstream task has limited training data. Following the setup by Chen et al. (2020a), we finetune the pretrained model on 1% or 10% of ImageNet training samples. Table 3 shows the results. MoPro consistently outperforms CE when pretrained on Web datasets. Compared to self-supervised learning methods pretrained on ImageNet, weakly-supervised learning achieves significantly better performance with fewer number of epochs.

Surprisingly, pretraining on the larger WebVision-V2 leads to worse performance compared to V0.5 and V1.0. This is because WebVision-V0.5 and V1.0 contain the same 1k class as ImageNet, whereas V2 also contains 4k extra classes. Hence, the representations learned from V2 are less task-specific and more difficult to adapt to ImageNet, especially with only 1% of samples for finetuning. This suggests that if the classes for a downstream task are known a priori, it is more effective to curate a task-specific weakly-labeled dataset with the same classes.

	Pretrain Method	Pretrain dataset	#Pretrain epochs	Top-1 1%	Top-1 10%	Top-5 1%	Top-5 10%
Random init.	None	None	None	25.4	56.4	48.4	80.4
Self-supervised	PCL	ImageNet	200	48.8	62.9	75.3	85.6
	SimCLR		1000	48.3	65.6	75.5	87.8
	BYOL		1000	53.2	68.8	78.4	89.0
	SwAV		800	53.9	70.2	78.5	89.9
Weakly-supervised	CE	WebVision-V0.5	90	65.9	72.4	87.0	90.9
	MoPro (ours)			69.3	73.3	89.1	91.7
	CE	WebVision-V1.0	90	67.6	73.5	88.3	91.7
	MoPro (ours)			**71.2**	**74.8**	**90.5**	**92.4**
	CE	WebVision-V2.0	90	62.1	72.9	86.9	91.4
	MoPro (ours)			65.3	73.7	88.2	92.1

Table 3: **Low-resource finetuning** on ImageNet. A pretrained model is finetuned with 1% or 10% of ImageNet training data. Weakly-supervised learning with MoPro substantially outperforms self-supervised learning methods: PCL (Li et al., 2020b), SimCLR (Chen et al., 2020a), BYOL (Grill et al., 2020), and SwAV (Caron et al., 2020). Result for random init. is from Zhai et al. (2019).

5.3 OBJECT DETECTION AND INSTANCE SEGMENTATION

We further transfer the pretrained model to object detection and instance segmentation tasks on COCO (Lin et al., 2014). Following the setup by He et al. (2019), we use the pretrained ResNet-50 as the backbone for a Mask-RCNN (He et al., 2017) with FPN (Lin et al., 2017). We finetune all layers end-to-end, including BN. The schedule is the default $1\times$ or $2\times$ in Girshick et al. (2018) Table 4 shows the results. Weakly-supervised learning with MoPro outperforms both supervised learning on ImageNet and self-supervised learning on one billion Instagram images.

5.4 ROBUSTNESS

It has been shown that deep models trained on ImageNet lack robustness to out-of-distribution samples, often falsely producing over-confident predictions. Hendricks *et al*. have curated two benchmark datasets to evaluate models' robustness to real-world distribution variation: (1) ImageNet-R (Hendrycks et al., 2020) which contains various artistic renditions of object classes from the original ImageNet dataset, and (2) ImageNet-A (Hendrycks et al., 2019) which contains natural images where ImageNet-pretrained models consistently fail due to variations in background elements, color, or texture. Both datasets contain 200 classes, a subset of ImageNet's 1,000 classes.

We evaluate weakly-supervised trained models on these two robustness benchmarks. We report both accuracy and the ℓ_2 calibration error (Kumar et al., 2019). The calibration error measures the misalignment between a model's confidence and its accuracy. Concretely, a well-calibrated classifier which give examples 80% confidence should be correct 80% of the time. Results are shown in Table 5. Webly-supervised learning show significantly higher accuracy and lower calibration error. The robustness to distribution shift could come from the higher diversity of samples in Web images. Compared to vanilla CE, MoPro further improves the model's robustness on both datasets. Note that we made sure that the training data of WebVision does not overlap with the test data.

Method	Pretrain dataset	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
random	None	31.0	49.5	33.2	28.5	46.8	30.4
CE (Sup.)	ImageNet	38.9	59.6	42.7	35.4	56.5	38.1
MoCo	Instagram-1B	38.9	59.4	42.3	35.4	56.5	37.9
CE	WebVision-V1.0	39.2	60.0	42.9	35.6	56.8	38.0
MoPro		39.7 (+0.8)	60.9 (+1.3)	43.1 (+0.4)	36.1 (+0.7)	57.5 (+1.0)	38.6 (+0.5)
MoPro	WebVision-V2.0	**40.7** (+1.8)	**61.7** (+2.1)	**44.5** (+1.8)	**36.8** (+1.4)	**58.4** (+1.9)	**39.6** (+1.5)

(a) $1\times$ schedule

Method	Pretrain dataset	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
random	None	36.7	56.7	40.0	33.7	53.8	35.9
CE (Sup.)	ImageNet	40.6	61.3	44.4	36.8	58.1	39.5
MoCo	Instagram-1B	41.1	61.8	45.1	37.4	59.1	40.2
CE	WebVision-V1.0	40.9	61.6	44.7	37.2	58.7	40.1
MoPro		41.2 (+0.6)	62.2 (+0.9)	45.0 (+0.6)	37.4 (+0.6)	58.9 (+0.8)	40.3 (+0.8)
MoPro	WebVision-V2.0	**41.8** (+1.2)	**62.6** (+1.3)	**45.6** (+1.2)	**37.8** (+1.0)	**59.5** (+1.4)	**40.6** (+1.1)

(a) $2\times$ schedule

Table 4: **Object detection and instance segmentation** using Mask-RCNN with R50-FPN fine-tuned on COCO `train2017`. We evaluate bounding-box AP (AP^{bb}) and mask AP (AP^{mk}) on `val2017`. Weakly-supervised learning with MoPro outperforms both supervised learning on ImageNet and self-supervised learning (MoCo (He et al., 2019)) on one billion Instagram images.

Method	Pretrain dataset	ImageNet-R		ImageNet-A	
		Accuracy (↑)	Calib. Error (↓)	Accuracy (↑)	Calib. Error (↓)
CE (Sup.)	ImageNet	36.14	19.66	0.03	62.50
CE	WebVision-V1.0	49.56	10.05	10.24	37.84
MoPro		**54.87**	**5.73**	**11.93**	**35.85**

Table 5: **Evaluation of model robustness** on images with artistic and natural distribution shifts. Weakly supervised learning with MoPro leads to a more robust and well-calibrated model.

	MoPro	w/o $\mathcal{L}_{\text{pro}}$	w/o $\mathcal{L}_{\text{inst}}$	w/o $\boldsymbol{s}_i$ (*i.e.* $\alpha = 1$)	w/o $\mathcal{L}_{\text{pro}}$ & $\mathcal{L}_{\text{inst}}$ & $\boldsymbol{s}_i$	CE
ImageNet acc.	69.3	68.0	68.2	68.4	66.9	65.9

Table 6: Ablation study where different components are removed from MoPro. Models are pre-trained on WebVision-V0.5 and finetuned on 1% of ImageNet data.

6 Ablation Study

We perform ablation study to verify the effectiveness of three important components in MoPro: (1) prototypical contrastive loss $\mathcal{L}_{\text{pro}}$, (2) instance contrastive loss $\mathcal{L}_{\text{ins}}$, (3) prototypical similarity $\boldsymbol{s}_i$ used for noise correction (equation 4). We choose low-resource finetuning on 1% of ImageNet training data as the benchmark, and report the top-1 accuracy for models pretrained on WebVision-V0.5. As shown in Table 6, all of the three components contribute to the efficacy of MoPro.

7 Conclusion

This paper introduces a new contrastive learning framework for webly-supervised representation learning. We propose momentum prototypes, a simple component that is effective in label noise correction, OOD sample removal, and representation learning. MoPro achieves state-of-the-art performance on the upstream task of learning from real-world noisy data, and superior representation learning performance on multiple down-stream tasks. Webly-supervised learning with MoPro does not require the expensive annotation cost in supervised learning, nor the huge computation budget in self-supervised learning. For future work, MoPro could be extended to utilize other sources of free Web data, such as weakly-labeled videos, for representation learning in other domains.

REFERENCES

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *ICML*, pp. 312–321, 2019.

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.

Pengfei Chen, Benben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *ICML*, pp. 1062–1070, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020a.

Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *ICCV*, pp. 1431–1439, 2015.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Santosh Kumar Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, pp. 3270–3277, 2014.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018.

Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *ICCV*, pp. 6391–6400, 2019.

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.

Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R. Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *ECCV*, pp. 139–154, 2018.

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. In *NeurIPS*, pp. 8536–8546, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, pp. 2980–2988, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

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. In *ICML*, pp. 2309–2318, 2018.

Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In *ICML*, 2020.

Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *ECCV*, 2016.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Large scale learning of general visual representations for transfer. In *ECCV*, 2020.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *NeurIPS*, pp. 3787–3798, 2019.

Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *CVPR*, pp. 5447–5456, 2018.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Learning to learn from noisy labeled data. In *CVPR*, pp. 5051–5059, 2019.

Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2020a.

Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven C.H. Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020b.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pp. 740–755, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, pp. 936–944, 2017.

Xingjun Ma, Yisen Wang, Michael E. Houle, Shuo Zhou, Sarah M. Erfani, Shu-Tao Xia, Sudanthi N. R. Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *ICML*, pp. 3361–3370, 2018.

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. In *ECCV*, pp. 185–201, 2018.

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

Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *ICLR*, 2015.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, pp. 843–852, 2017.

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *CVPR*, pp. 5552–5560, 2018.

Yi Tu, Li Niu, Dawei Cheng, and Liqing Zhang. Protonet: Learning from web data with memory. In *CVPR*, 2020.

Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *NIPS*, pp. 5601–5610, 2017.

Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge J. Belongie. Learning from noisy large-scale datasets with minimal supervision. In *CVPR*, pp. 6575–6583, 2017.

Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *CVPR*, pp. 8688–8696, 2018.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pp. 3733–3742, 2018.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, pp. 2691–2699, 2015.

Jingkang Yang, Litong Feng, Weirong Chen, Xiaopeng Yan, Huabin Zheng, Ping Luo, and Wayne Zhang. Webly supervised image classification with self-contained confidence. In *ECCV*, 2020.

Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *CVPR*, 2019.

Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *ICCV*, pp. 1476–1485, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

Weihe Zhang, Yali Wang, and Yu Qiao. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In *CVPR*, 2019.

Zizhao Zhang, Han Zhang, Sercan Ömer Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. In *CVPR*, pp. 9291–9300, 2020.

Bolei Zhou, Àgata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pp. 487–495, 2014.