Distill on the Go: Online knowledge distillation in self-supervised learning

Prashant Bhat, Elahe Arani, Bahram Zonooz

Advanced Research Lab, NavInfo Europe, Eindhoven, The Netherlands

{prashant.bhat, elahe.arani}@navinfo.eu, bahram.zonooz@gmail.com

Abstract

Self-supervised learning solves pretext prediction tasks that do not require annotations to learn feature representations. For vision tasks, pretext tasks such as predicting rotation, solving jigsaw are solely created from the input data. Yet, predicting this known information helps in learning representations useful for downstream tasks. However, recent works have shown that wider and deeper models benefit more from self-supervised learning than smaller models. To address the issue of self-supervised pre-training of smaller models, we propose Distill-on-the-Go (DoGo), a self-supervised learning paradigm using single-stage online knowledge distillation to improve the representation quality of the smaller models. We employ deep mutual learning strategy in which two models collaboratively learn from each other to improve one another. Specifically, each model is trained using self-supervised learning along with distillation that aligns each model's softmax probabilities of similarity scores with that of the peer model. We conduct extensive experiments on multiple benchmark datasets, learning objectives, and architectures to demonstrate the potential of our proposed method. Our results show significant performance gain in the presence of noisy and limited labels, and in generalization to out-of-distribution data.

1. Introduction

Self-supervised learning (SSL) has recently begun to rival the performance of supervised learning on computer vision tasks [14, 5, 4]. SSL learns meaningful representations from data without requiring manually annotated labels. To learn task-agnostic visual representations, SSL solves pretext prediction tasks such as predicting relative position [10] and/or rotation [13], solve jigsaw [24] and image in-painting [34]. Predicting known information helps in learning representations generalizable for downstream tasks such as segmentation and object detection [28].

Recent works have empirically shown that deeper and wider models benefit more from task agnostic use of unlabeled data than their smaller counterparts i.e smaller models
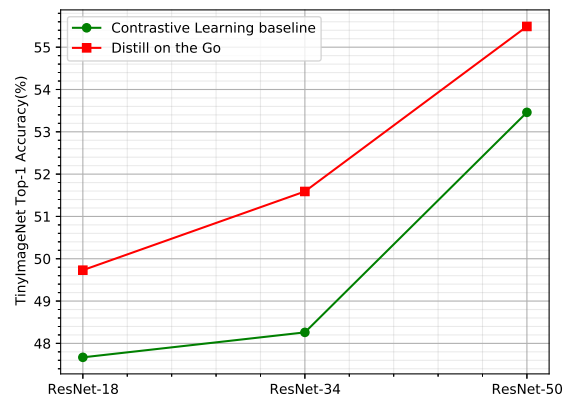

Figure 1. Tiny-ImageNet evaluation: comparison of state-of-the-art self-supervised contrastive learning against our proposed method Distill-on-the-go using linear evaluation. Self-supervised models are trained using SimCLR while Distill-on-the-go models are trained together with ResNet-50.

when trained using self-supervised learning fail to close in the gap with respect to supervised training [11, 6]. For instance level discrimination tasks such as contrastive learning, smaller models with fewer parameters are hard to optimize given large amounts of data. The limitation in the representation quality lies in the difficulty of optimization rather than the model size [2].

To address this issue and improve the representation quality of smaller models, we leverage knowledge distillation (KD) [20]. KD involves training a student model under the supervision of a larger pre-trained teacher model in an interactive manner, similar to how humans learn. Many of the traditional KD methods formulate the student to mimic the softened softmax output of the teacher. Thus, these methods are not directly applicable to self-supervised knowledge distillation in the absence of the labels. Also, offline KD methods require longer training process and significantly large memory and computational resources to pretrain large teacher models [26].

Online knowledge distillation offers a more attractive alternative owing to its one stage training and bidirectional knowledge distillation [47, 15, 26, 37]. These approaches

treat all (typically two) participating models equally, enabling them to learn from each other. In this paper, we propose a fully self-supervised online knowledge distillation framework, *Distill-on-the-go (DoGo)*. In contrast to offline knowledge distillation, our proposed method starts with multiple untrained models which simultaneously learn by solving a pretext task. Specifically, we train two models collaboratively by applying different augmentation for each model using our method. Each model generates two projections corresponding to two randomly augmented views. We then align temperature scaled similarity scores across these projections for KD in which the relative differences in similarity between the reference sample and other samples in a batch can provide additional useful information. The additional supervision signal can assist the optimization of the smaller model. Furthermore, the collaboration between multiple models enables them to explore the feature space more extensively and aids in converging to a more robust (flatter) minima which leads to a better generalization to unseen data [47].

With extensive experiments on different datasets, we empirically demonstrate that learning along with peers is better than learning alone in a conventional self-supervised learning setting. For example, on Tiny-ImageNet dataset [27], our method improves top-1 accuracy(%) of ResNet-50 by 2.7% and ResNet-18 by 4.5% compared to contrastive learning baselines (Figure 1). Our main contributions are as follows:

- We propose a fully self-supervised online KD method DoGo to improve the representation quality of the smaller models. The method is completely self-supervised, i.e. knowledge is distilled during the pre-training stage in the absence of labels.

- With DoGo, we demonstrate a significant improvement in performance for smaller models over contrastive learning baselines.

- We demonstrate the efficacy of DoGo on multiple SSL algorithms, thereby showing that our method is independent of the underlying SSL learning objective.

- We further show the effectiveness of DoGo in solving more common challenges in the real-world problems including learning with noisy and limited labels, and generalization to out-of-distribution data.

2. Related Work

Self-supervised learning: SSL can be broadly categorized into generative and contrastive methods [29]. Generative self-supervised models try to learn meaningful visual representations by re-constructing either a part of an input or a whole of it. For example, Auto-Regressive models use observations from previous time step to predict the joint distribution factorised as a product of conditionals. Pixel-RNN [32] and PixelCNN [40] adapt this idea to model image pixel by pixel.

Contrastive learning, on the other hand, learns to compare through Noise Contrast Estimation [16]. Context-instance contrast methods model belonging relationship between local feature of a sample and its global context. Predict Relative Position (PRP) and Maximize Mutual Information (MI) fall under this category. In PRP, self-supervised methods focus on learning visual representations by predicting relative position of local components. PRP pretext tasks include predict relative position [10] and/or rotation [13], solve jigsaw [24] etc. In MI, Deep InfoMax [21] employs Mutual Information Neural Estimation [3] to learn unsupervised representations by simultaneously maximizing and estimating mutual information between learned high level representations and the input data.

Although MI based contrastive self-supervised methods achieved quite a bit of success, some recent methods outperform them through context-context contrastive learning [29]. These methods discard mutual information, instead learn the relationship between the global representations of different samples akin to metric learning. InstDisc [44] proposed instance discrimination as a pretext task. CMC [38] employed multi-view contrastive learning framework with multiple different views of an image as positive samples and take views of other images as the negatives. MoCo [17] further developed the idea of instance discrimination by leveraging momentum contrast. MoCo addressed two critical issues in dealing with the negative sampling: (i) Momentum contrast prevents the fluctuation of loss convergence in the earlier stages, (ii) MoCo maintains a queue to store negative samples from previous batches thereby significantly improving the negative sample efficiency. However, MoCo adopts too simple augmentations, thus making positive pair far too easy to identify. Hard positive sample strategy plays a key role in instance discrimination as outlined in SimCLR [5]. Therefore, SimCLR relinquishes momentum contrast overall but retains the siamese structure and introduces augmentations of 10 forms with an end-to-end training framework. SimCLRv2 [6] outlined that bigger models benefit more from a task agnostic use of unlabeled data for visual representation learning. Owing to larger modeling capacity, bigger self-supervised models are far more label efficient and perform better than smaller models on downstream tasks.

Knowledge distillation: KD is an effective technique for improving the performance of compact models either using the supervision of larger pre-trained model or with a cohort of smaller models trained collaboratively. In a typical "teacher-student" knowledge distillation, smaller student model is trained under the supervision of a pre-trained

$$\langle z_1', z_1'' \rangle \;\; \gg \;\; \langle z_1', k_i \rangle$$

$$P = \sigma(sim(Z_1', Z_1'')/\tau_{kd})$$

$$KL(P\|Q) \qquad KL(Q\|P)$$

$$Q = \sigma(sim(Z_2', Z_2'')/\tau_{kd})$$

$$\langle z_2', z_2'' \rangle \;\; \gg \;\; \langle z_2', k_j \rangle$$
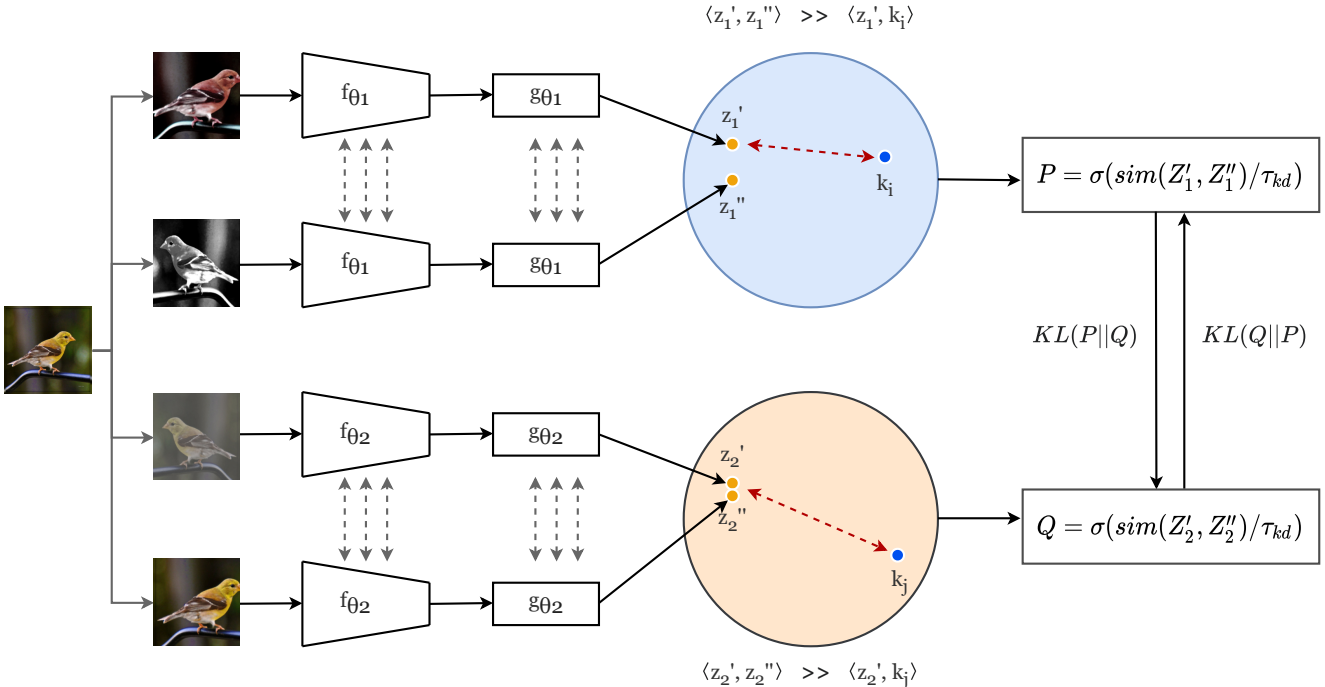
Figure 2. Distill-on-the-go: The proposed method consists of a stochastic augmentation module, followed by shared encoder and a two-layer MLP projection head. The participating models primarily learn visual representations by contrastive loss. $\langle .,. \rangle$ represents cosine similarity between two embeddings. The extra learning experience is provided by way of aligning softmax probabilities of temperature-scaled similarity scores across augmentations of two peer models.

larger teacher model, producing a higher representation quality compared to training student model from scratch. In the original formulation, Hinton *et al*. [20] proposed a representation distillation by way of mimicking softened softmax output of the teacher. Better generalization can be achieved by emulating the latent feature space in addition to mimicking the output of the teacher [37, 39, 45, 36, 33].

To circumvent the associated computational costs of pre-training a teacher, deep mutual learning (DML) [47] proposed online knowledge distillation using Kullback–Leibler (KL) divergence. Alongside a primary supervised cross-entropy loss, DML involves training each participating model using a distillation loss that aligns the class posterior probabilities of the current model with that of the other models in the cohort. Inspired by the recent progress made by online knowledge distillation methods, we propose a fully self-supervised online knowledge distillation method to improve the representation quality of the smaller models. To the best of our knowledge, we are the first to propose an online knowledge distillation method in the absence of labels.

3. Proposed Method

We begin with the core idea of unsupervised contrastive representation learning. We wish to learn visual representations by contrasting semantically similar (positive) and dis-similar (negative) pairs of data samples such that similar pairs have the maximum agreement via a contrastive loss in the latent space. Contrastive loss encourages the representations of similar pairs to be close, while those of dissimilar pairs to be more orthogonal in the latent space [8]. Empirically, similar pairs are obtained by randomly augmenting the same sample using rotation, gaussian noise and color jittering [5, 44, 17].

Inspired by the recent advancements in contrastive representation learning, DoGo comprises of the following major components: It consists of a stochastic augmentation module resulting in two highly correlated views I' and I'' of the same input sample I. The correlated views are then fed into $f_\theta(.)$, typically an encoder network such as ResNet-50 [18] and subsequently to $g_\theta(.)$, a two-layer perceptron with ReLU non-linearity. To learn the visual representations, the network $g_\theta(f_\theta(.))$ should learn to maximize the similarity between the positive embedding pair $\langle z', z'' \rangle$ while simultaneously pushing away the negative embedding pairs $\langle z', k_i \rangle$, where $i = (1, ..., K)$ are the embeddings of augmented views of other samples in a batch and K is the number of negative samples. Contrastive representation learning can thus be cast as an instance level discrimination task. Instance level discrimination objective is typically formulated using a softmax criterion. However, the cost of computing non-parametric softmax is prohibitively large especially

3

when the number of instances are very large [44]. Popular techniques to reduce computation include hierarchical softmax [31], noise contrast estimation [16] and negative sampling [30]. Following [5, 6], We use noise contrast estimation for a positive embedding pair $\langle z_i', z_i'' \rangle$ where $i \in \{1, 2\}$ indicates the two models as follows:

$$L_{cl,i} = -log \frac{e^{sim(z_i', z_i'')/\tau_c}}{e^{sim(z_i', z_i'')/\tau_c} + \sum_{j=1}^{K} e^{sim(z_i', k_j)/\tau_c}} \quad (1)$$

L_{cl} is a normalized temperature scaled cross entropy loss [5]. Wang *et al.* [43] provided an in-depth understanding of necessity of normalization when using dot product of feature vectors in a cross entropy loss. Therefore, we use cosine similarity (l2 normalized dot product) in the computation of the contrastive loss L_{cl}.

Smaller models find it hard to optimize and find right set of parameters in instance level discrimination tasks, attributing to difficulty of optimization rather than the model size. We believe that the additional supervision in KD regarding the relative differences in similarity between the reference sample and other sample pairs and the collaboration between multiple models can assist the optimization of the smaller model. Therefore to improve generalizability of smaller model $g_{\theta 1}(f_{\theta 1}(.))$, we propose to utilize another peer model $g_{\theta 2}(f_{\theta 2}(.))$. Given a new sample, each participating peer model generates embeddings z', z'' of two different augmented views. Let $Z', Z'' \in R^{N \times m}$ be a batch of z', z'' where N is batch size and m is the length of the projection vector. Let $P = \sigma(sim(Z_1', Z_1'')/\tau)$ and $Q = \sigma(sim(Z_2', Z_2'')/\tau)$ be softmax probabilities of temperature-scaled similarity scores across augmentations of two peer models. We employ KL divergence to distill the knowledge across peers by aligning the distributions P and Q. The distillation losses are defined as follows:

$$L_{kd,1} = D_{KL}(Q||P)$$
$$= \sigma \left(\frac{sim(Z_2', Z_2'')}{\tau_{kd}} \right) log \frac{\sigma(sim(Z_2', Z_2'')/\tau_{kd})}{\sigma(sim(Z_1', Z_1'')/\tau_{kd})} \quad (2)$$

$$L_{kd,2} = D_{KL}(P||Q)$$
$$= \sigma \left(\frac{sim(Z_1', Z_1'')}{\tau_{kd}} \right) log \frac{\sigma(sim(Z_1', Z_1'')/\tau_{kd})}{\sigma(sim(Z_2', Z_2'')/\tau_{kd})} \quad (3)$$

The final learning objective for the two participating models can be written as:

$$L_{\theta_1} = L_{cl,1} + \lambda L_{kd,1} \quad (4)$$

$$L_{\theta_2} = L_{cl,2} + \lambda L_{kd,2} \quad (5)$$

where λ is a regularization parameter for adjusting the magnitude of the knowledge distillation loss. Our method can also be extended to more than two peers by simply computing distillation loss with all the peers. Another strategy is to create an ensemble distribution from all participating peers and use it as a single teacher for creating learning experience. We reserve training with more than two peers as a possible future research direction.

4. Experimental Results

4.1. Implementation details

We use SimCLR [5] as our baseline network. It consists of a feature extractor and an additional projector. The projector is a two layer MLP with ReLU non-linearity generating embeddings of size 128. Since Tiny-ImageNet images are much smaller than ImageNet, we replace the first 7×7 Conv of stride 2 with a 3×3 Conv of stride 1 in ResNet backbones. We also remove max pooling layer from the first Conv block. Our stochastic augmentation module employs random resized crop, random horizontal flip followed by random color distortions. Since images are small in size, we leave out the Gaussian blur so as not to risk overaugmentation. We use Adam optimizer with a fixed learning rate of $3e^{-4}$ and weight decay of $1e^{-6}$ with a batch size of 256 across all pre-training experiments. To make a fair comparison, we use $\lambda = 100$ keeping other hyperparameters intact in knowledge distillation experiments.

Datasets: We use Tiny-ImageNet [27] as our primary dataset for self-supervised pre-training. It has 200 classes and 500 samples per class each 64×64 in size. The test set contains 10,000 images. Tiny-ImageNet pre-trained models are evaluated on six different out-of-distribution datasets including CIFAR-10 [25], CIFAR-100 [25], STL-10 [9] and three DomainNet [35] datasets. DomainNet is a domain adaptation dataset with six domains and 0.6 million images distributed across 345 categories. We have used ClipArt, Sketch and QuickDraw domain datasets in this work.

4.2. Evaluation metrics

Linear Evaluation: To evaluate the learned representations, we follow the widely used linear evaluation metric [5, 41] as our primary evaluation protocol. A linear layer is trained on the labeled training set on top of frozen encoder $f_\theta(.)$. Top-1 test accuracy(%) is used as a quality measure for learned representations. To reduce hyperparameter tuning per experiment, we standardize the linear evaluation experiments as follows: We use Adam optimizer with a fixed learning rate of $3e^{-4}$ and weight decay of $1e^{-6}$ with a batch size of 64. All linear evaluations are run for 100 epochs except out-of-distribution experiments, which are evaluated for 50 epochs.

Nearest Neighbor Evaluation: We also evaluate learned representations using nearest neighbour classification with an efficient similarity search library Faiss GPU

Baseline	ResNet-50 (53.46)	ResNet-34 (48.26)	ResNet-18 (47.58)
ResNet-50 (53.46)	55.49 \ 54.93 +3.8% +2.7%	55.24 \ 51.59 +3.3% +6.9%	54.93 \ 49.73 +2.7% +4.5%
ResNet-34 (48.26)	51.59 \ 55.24 +6.9% +3.3%	50.86 \ 51.32 +5.4% +6.3%	50.26 \ 49.10 +4.1% +3.1%
ResNet-18 (47.67)	49.73 \ 54.93 +4.5% +2.7	49.10 \ 50.26 +3.1% +4.1%	48.77 \ 48.74 +2.3% +2.2%

Baseline	WRN-28-2 (37.10)	WRN-16-2 (34.28)	WRN-10-2 (30.75)
WRN-28-2 (37.10)	38.16 \ 37.73 +2.9% +1.7%	37.79 \ 36.30 +1.9% +5.9%	36.49 \ 33.55 -1.7% +9.1%
WRN-16-2 (34.28)	36.30 \ 37.79 +5.9% +1.9%	35.92 \ 35.68 +4.8% +4.1%	34.69 \ 33.04 +1.2% +7.4%
WRN-10-2 (30.75)	33.55 \ 36.49 +9.1% -1.7%	33.04 \ 34.69 +7.4% +1.2%	32.53 \ 32.60 +5.8% +6.0%

Table 1. Tiny-ImageNet top-1 accuracy(%) under linear evaluation for various ResNet and Wide-ResNet models. Baseline measures the linear evaluation accuracy when the models are pre-trained using contrastive learning alone. Each box contains DoGo results: left value corresponds to model in the corresponding row, right value corresponds to model in the corresponding column. Percentage change in accuracy between baseline and our method are highlighted in green.

Label(%)	CL Baseline	DoGo
1	15.00	**15.46**
5	29.49	**29.64**
10	34.23	**36.07**
20	40.23	**42.43**
50	49.81	**50.46**
100	54.76	**55.57**

Table 2. Tiny-ImageNet top-1 accuracy(%) for ResNet-18 with varied available labels during fine-tuning. Labels are sampled in a class balanced manner. CL baseline is trained using contrastive loss while DoGo is jointly trained with ResNet-50. DoGo consistently outperforms the contrastive learning baseline.

[22]. We extract the features from frozen encoder $f_\theta(.)$ and index features along with their labels. We report the average of nearest neighbour classification with $k = 1, 2, 4, 8$ neighbors. The results can be found in the supplementary material.

4.3. Evaluation on in-distribution data

Table 1 compares the top-1 accuracy(%) of our proposed method on Tiny-ImageNet dataset obtained by ResNet and Wide-ResNet architectures. Following observations can be made from the table: (i) All participating models under ResNet family consistently see an improvement over the contrastive learning baseline. (ii) Although ResNet-50 is a much larger network compared to ResNet-18, ResNet-50 still sees a significant improvement. It is therefore clear that all participating models can benefit from joint training unless there is a drastic difference in modeling capacity. (iii) Wide-ResNets also display a similar trend with an exception of WRN-28-2 owing to a large difference in model capacity between WRN-10-2 and WRN-28-2. However, WRN-10-2 achieves a significant improvement of over 9%. (iv) Smaller models benefit more from joint training than their larger counterparts. Given the difficulty of training smaller models under contrastive learning regime, joint training under our proposed method can significantly improve their performance. (v) Smaller models gain a consistent improvement with the increase in modeling capacity of their counterparts. For example, the performance gain in ResNet-18 monotonically increases as we increase the capacity of the peer model from ResNet-18 to ResNet-34 and ResNet-50. A similar trend can be seen for WRN-10-2, WRN-16-2 and ResNet-34. (vi) Contrary to conventional wisdom, even joint training of two models with exact same architecture yields a discernible improvement over the baseline.

4.4. Evaluation under varying label quantities

We sample 1%, 5%, 10%, 20%, 50% and 100% of the Tiny-ImageNet in a class balanced manner and simply fine-tune the whole encoder $f_\theta(.)$ with a linear layer on top on the labeled data. We use 15% of the training data as a validation set and top-1 accuracy(%) are reported on full test

		CIFAR-10	CIFAR-100	STL-10	Clip Art	Quick Draw	Sketch
ResNet-50	CL Baseline	78.59	53.54	76.78	37.82	47.31	31.04
	DoGo	**80.26**	**56.36**	**77.93**	**42.20**	**51.88**	**33.62**
ResNet-18	CL Baseline	72.50	45.98	72.90	29.10	39.97	23.68
	DoGo	**75.19**	**48.52**	**74.73**	**33.51**	**43.43**	**26.25**

Table 3. Linear evaluation on out-of-distribution datasets. CL baseline is trained using contrastive loss while DoGo is a joint training of ResNet-50 and ResNet-18.
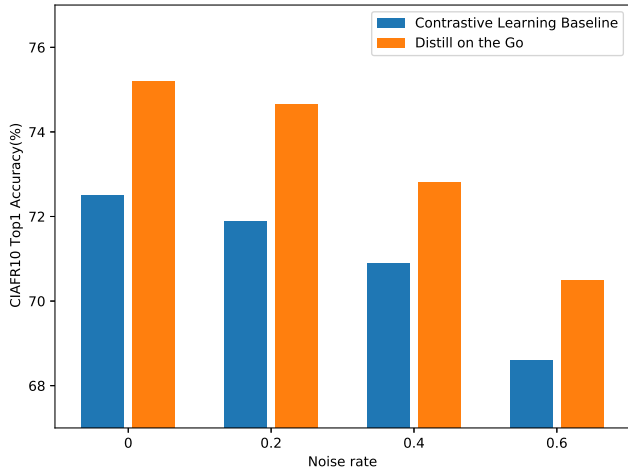

Figure 3. CIFAR-10 top-1 accuracy(%) for ResNet-18 under linear evaluation with simulated noisy labels. CL baseline is trained using contrastive learning while DoGo is jointly trained with ResNet-50. DoGo significantly outperforms the CL baseline across all noise rates.

set. Table 2 shows the comparison of fine-tuning pre-trained models on fewer labels against contrastive learning baseline. It can be clearly seen that our method outperforms the contrastive learning baseline across all limited label scenarios. The difference is more lucid when the available labels are 10% or higher. The difference is higher in the lower mid range (10%, 20%) and saturates thereafter. It is therefore clear that online knowledge distillation brings discernible benefits when pre-trained models are fine-tuned on the downstream task.

4.5. Generalization on out-of-distribution data

Self-supervised contrastive learning learns the visual representations in a task agnostic manner. It is therefore essential that these learned representations generalize well across downstream tasks. To evaluate the models pre-trained on Tiny-ImageNet, we consider six different out-of-distribution datasets. Table 3 compares the top-1 test accuracy(%) under linear evaluation protocol of pre-trained models with and without online knowledge distillation. ResNet-50 and ResNet-18 generalize well across all datasets when trained together using our method compared

to the baseline. Results show that DoGo generalizes well across datasets with drastically different distribution than Tiny-ImageNet indicating DOGo learns better feature representations.

4.6. Evaluation under noisy labels

The success of supervised learning often hinges on the availability of huge amounts of high quality annotations. Quite often, large datasets contain noisy labels due to difficulty of manual annotation. It is pertinent to have robust training procedures to offset the impact of noisy labels as studies have shown that deep neural networks can memorize noisy labels [1]. Self-supervised learning decouples representation learning and classifier, thus, is more robust to noisy labels than supervised learning [19, 46]. We hypothesize that learned representations in DoGo are more robust to noisy labels. To test our hypothesis, we simulate label corruption on CIFAR-10. With a given probability (noise rate), we corrupt every true label by randomly sampling from a uniform distribution over number of classes. Figure 3 presents the results of linear evaluation under noisy labels. Our method consistently outperforms the contrastive learning baseline across different noise rates.

Baseline	Contrastive loss L_{cl} (47.67)		SimSiam loss L_{ss} (46.80)	
L_{cl} (47.67)	**48.77** \ **48.74**		**48.61** \ **49.03**	
	+2.3%	+2.2%	+2.0%	+4.8%
L_{ss} (46.80)	**49.03** \ **48.61**		**48.95** \ **48.38**	
	+4.8%	+2.0%	+4.6%	+3.4%

Table 4. Tiny-ImageNet top-1 accuracy(%) for ResNet-18 under linear evaluation for different learning objectives. DoGo is jointly trained with a ResNet-18 peer model. Each box contains DoGo results: left value corresponds to model in the corresponding row, right value corresponds to model in the corresponding column. Percentage change in accuracy between baseline and our method are highlighted in green. DoGo surpasses the baselines on all combinations of learning objectives.
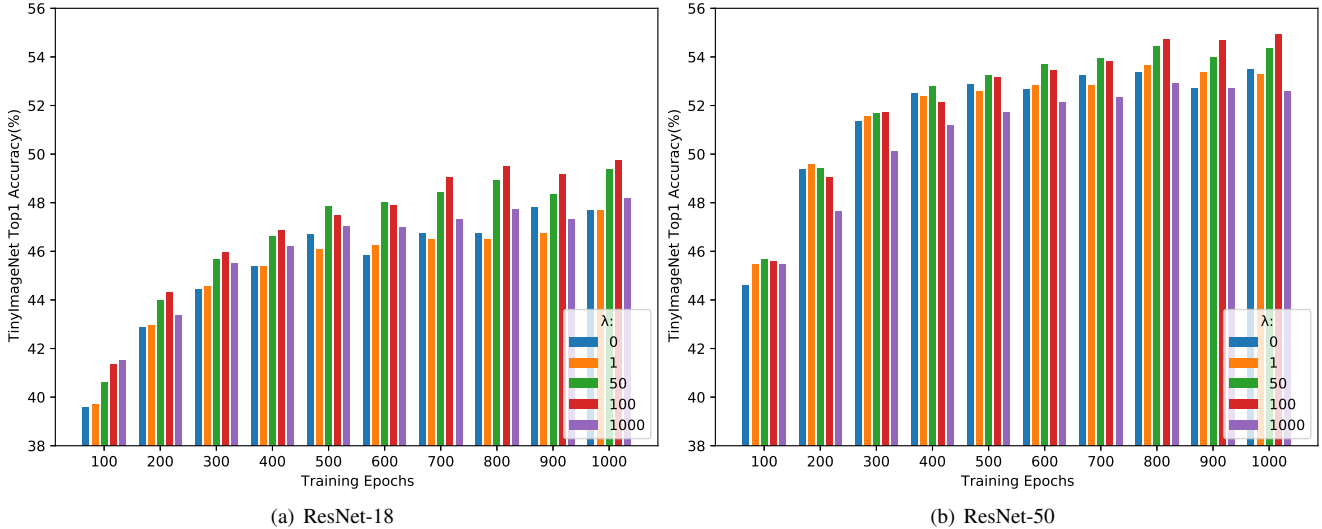
(a) ResNet-18

(b) ResNet-50

Figure 4. Hyperparameter tuning of regularization parameter λ

4.7. Distillation across learning objectives

The experimental results shown thus far have had contrastive learning as their primary learning objective. In this section, we explore whether it is possible to distill knowledge across peers with different learning objective. In contrast to contrastive learning, SimSiam [7] is one such learning objective that can learn meaningful visual representations without negative sample pairs and large batches. Sim-Siam can best be described as BYOL [14] without momentum encoder, SimCLR [5] without negative pairs and Swav [4] without online clustering. Essentially, it covers all leading self-supervised learning methods and makes an ideal candidate for our research. In addition to encoder $f_\theta(.)$ and projection MLP $g_\theta(.)$, SimSiam has a prediction MLP $h_\theta(.)$ that transforms the output of one view and matches it to the other view. Denoting two output vectors $p' = h_\theta(g_\theta(f_\theta(I')))$ and $z'' = g_\theta(f_\theta(I''))$, the learning objective is to minimize the negative cosine similarity,

$$L_{ss} = D(p', z'') = -\frac{p'}{||p'||_2} \cdot \frac{z''}{||z''||_2} \qquad (6)$$

To evaluate our method across learning objectives, we train one model using contrastive loss (Eq. 1) and the other by SimSiam loss (Eq. 6). Since SimSiam model has both projections ($\in R^{b \times 2048}$) and predictions ($\in R^{b \times 2048}$), we use predictions for creating learning experience during joint training. Table 4 shows the generalizability of our method beyond contrastive learning. Our method surpasses the baseline across all combinations of learning objectives. Results reinforce our earlier proposition that DoGo is independent of underlying SSL objective.
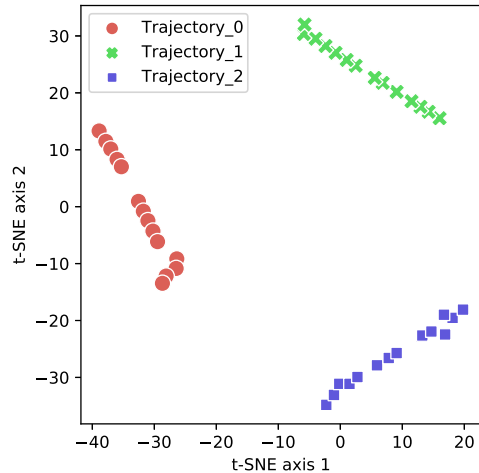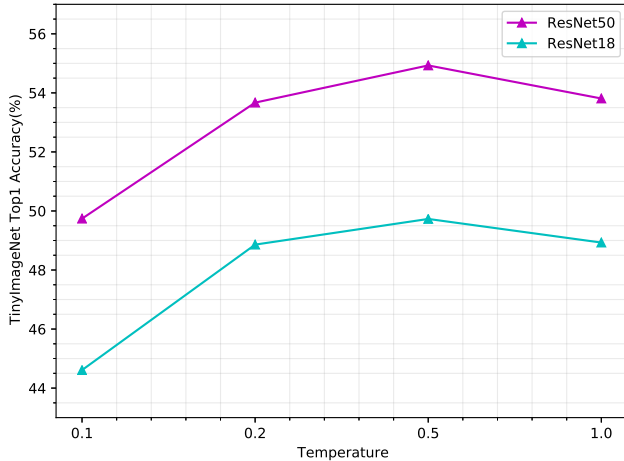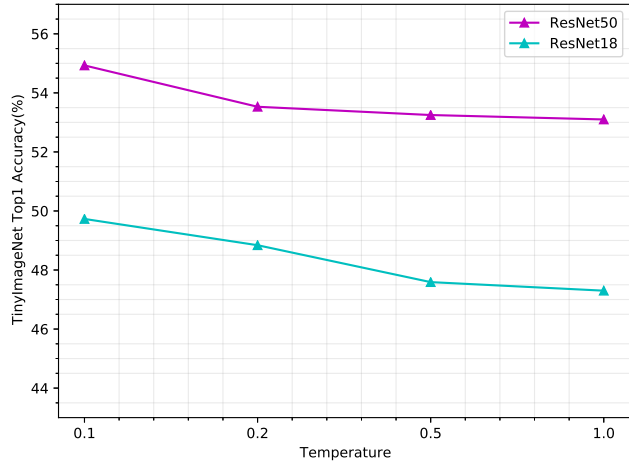

Figure 5. t-SNE plot of projections from checkpoints corresponding to 3 different randomly initialized trajectories.

4.8. Visualizing function space similarity

In this section, we attempt to provide additional intuition about why our method works. Motivated by [12], we plot t-SNE [42] of projections of different checkpoints along the individual training trajectories from multiple initializations on Tiny-ImageNet trained with contrastive learning. Trajectory_i in Figure 5 shows the variation in the function space during the training of model i. We can see that a single model only explores a limited function space whereas models initialized differently explore considerably different function spaces even though they provide similar performance. Therefore, we believe that the collaborative learning between multiple models in DoGo enables the models to collectively explore broader region of the function space re-

(a) Temperature in L_{cl} (b) Temperature in L_{kd}

Figure 6. Hyperparameter tuning of scalar temperature parameters (a) τ_c in contrastive loss and (b) τ_{kd} in knowledge distillation loss

sulting in more generalizable feature representations. DoGo enables the exchange of information across different family of functions resulting in improved downstream performance.

5. Effect of hyperparameters

We now explore the influence of λ and scalar temperature parameter τ on knowledge distillation.

Regularization parameter λ: Figure 4 summarizes the effect of λ parameter when ResNet-18 and ResNet-50 are trained together using our method. We draw several conclusions from the graph: (i) We see a clear performance improvement as the training progresses i.e. online knowledge distillation reaps more benefits from longer training. This is in line with similar empirical results on self-supervised learning [5]. (ii) Smaller models benefit early in the training while larger counterparts see a performance improvement in the later stages of the training. The smaller model ResNet-18, sees an improvement early in the training while ResNet-50 is benefited only in the later stages. We attribute this behaviour to large modeling capacity difference between these two models. (iii) For both ResNet-18 and ResNet-50, $\lambda = 1$ yields either comparable or slightly worse performance. This is due to the difference in magnitudes of contrastive loss and distillation loss functions. To compensate, we increase the weight λ for the distillation loss. We find that $\lambda = 100$ yields optimum performance for both participating models. Increasing λ further is detrimental for the downstream performance.

Temperature parameters τ_c and τ_{kd}: Several prior works have shown that an appropriately adjusted temperature parameter benefits contrastive cross entropy loss [5, 23, 44]. We can see from Eq. 1 that the range of simi-

larity can be scaled using an appropriate temperature value. Since the gradients of Eq. 1 will be scaled by $1/\tau_c$, tuning this parameter could effectively tune the hardness of the negative samples and speed of learning [23]. In our implementation, we use two decoupled scalar temperature parameters τ_c and τ_{kd}. Figure 6 shows the linear evaluation of ResNet-18 and ResNet-50 trained together using online knowledge distillation. We find that when training to convergence, the optimal τ_c is 0.5 and τ_{kd} is 0.1. The optimal value of $\tau_c = 0.5$ is in accordance with the empirical results in contrastive learning. We also find that decoupling of τ_c and τ_{kd} yields a much better result in our experiments.

6. Conclusion and future work

We proposed Distill-on-the-Go, a self-supervised learning paradigm using single stage online knowledge distillation to improve the representation quality of the smaller models. Through extensive experiments across multiple network architectures, SSL algorithms and datasets, we demonstrated the efficacy of DoGo over self-supervised learning baselines resulting in compact yet accurate models. Performance of compact models can be boosted even further by training them with much larger models. We further demonstrated the additional benefits of using online distillation in SSL in enabling the models to learn efficiently under varying degrees of label noise and different levels of annotations data availability. DoGo also enables the models to explore the function space more extensively and learn better generalizable features. Joint training with more than two peers, online knowledge distillation on bigger datasets such as ImageNet, downstream performance on other tasks such as object detection and segmentation are some of the useful future research directions for this work.

References

[1] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks, 2017. 6

[2] L.J. Ba and R. Caruana. Do deep nets really need to be deep? *Advances in Neural Information Processing Systems*, 3:2654–2662, 01 2014. 1

[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: Mutual information neural estimation, 2018. 2

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2021. 1, 7

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 1, 2, 3, 4, 7, 8

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners, 2020. 1, 2, 4

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning, 2020. 7

[8] Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning, 2020. 3

[9] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 215–223, 2011. 4

[10] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction, 2015. 1, 2

[11] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation, 2021. 1

[12] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective, 2020. 7

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018. 1, 2

[14] 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, 2020. 1, 7

[15] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11017–11026. IEEE, 2020. 1

[16] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. 2, 4

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2019. 2, 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 3

[19] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty, 2019. 6

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 1, 3

[21] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization, 2018. 2

[22] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017. 5

[23] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning, 2020. 8

[24] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Learning image representations by completing damaged jigsaw puzzles, 2018. 1, 2

[25] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012. 4

[26] xu lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 7517–7527. Curran Associates, Inc., 2018. 1

[27] Y. Le and X. Yang. Tiny imagenet visual recognition challenge. 2015. 2, 4

[28] Jason D. Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning, 2020. 1

[29] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive, 2020. 2

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119. Curran Associates, Inc., 2013. 4

[31] Frederic Morin and Yoshua Bengio. Hierarchistic neural network language model. In *AISTATS'05*, pages 246–252, 2005. 4

[32] Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1747–1756, New York, New York, USA, 20–22 Jun 2016. PMLR. 2

[33] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. *CoRR*, abs/1904.05068, 2019. 3

[34] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting, 2016. 1

[35] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. *CoRR*, abs/1812.01754, 2018. 4

[36] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. *arXiv*, 12 2014. 3

[37] Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Knowledge distillation beyond model compression, 2020. 1, 3

[38] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2019. 2

[39] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. *CoRR*, abs/1907.09682, 2019. 3

[40] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4790–4798. Curran Associates, Inc., 2016. 2

[41] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 4

[42] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 7

[43] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface. *Proceedings of the 25th ACM international conference on Multimedia*, Oct 2017. 4

[44] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination, 2018. 2, 3, 4, 8

[45] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[46] Hui Zhang and Quanming Yao. Decoupling representation and classifier for noisy label learning, 2020. 6

[47] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning, 2017. 1, 2, 3