PSEUDO-LABELING AND CONFIRMATION BIAS IN DEEP SEMI-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Semi-supervised learning, i.e. jointly learning from labeled an unlabeled samples, is an active research topic due to its key role on relaxing human annotation constraints. In the context of image classification, recent advances to learn from unlabeled samples are mainly focused on consistency regularization methods that encourage invariant predictions for different perturbations of unlabeled samples. We, conversely, propose to learn from unlabeled data by generating soft pseudo-labels using the network predictions. We show that a naive pseudo-labeling overfits to incorrect pseudo-labels due to the so-called confirmation bias and demonstrate that mixup augmentation and setting a minimum number of labeled samples per mini-batch are effective regularization techniques for reducing it. The proposed approach achieves state-of-the-art results in CIFAR-10/100 and Mini-ImageNet despite being much simpler than other state-of-the-art. These results demonstrate that pseudo-labeling can outperform consistency regularization methods, while the opposite was supposed in previous work. Code will be made available.

1 INTRODUCTION

Convolutional neural networks (CNNs) have become the dominant approach for many computer vision tasks (He et al., 2016a; Lin et al., 2017; Liu et al., 2018; Kim et al., 2018; Xie et al., 2018). The main requirement to best exploit them is the availability of vast amounts of labeled data. Obtaining such volumes of data, however, is not trivial, and the research community is exploring alternatives to alleviate this (Li et al., 2017; Wang & Deng, 2018; Oliver et al., 2018; Liu et al., 2019).

Knowledge transfer via deep domain adaptation (Wang & Deng, 2018) is a popular alternative that seeks to learn more transferable representations from source to target domains by embedding domain adaptation in the learning pipeline. Other approaches focus exclusively on learning useful representations from scratch in a target domain when annotation constraints are relaxed (Oliver et al., 2018; Arazo et al., 2019; Gidaris et al., 2018). Semi-supervised learning (Oliver et al., 2018) focuses on scenarios with sparsely labeled data and extensive amounts of unlabeled data; learning with label noise (Arazo et al., 2019) seeks robust learning when labels are obtained automatically and may not represent the image content; and self-supervised learning (Gidaris et al., 2018) uses data supervision to learn from unlabeled data in a supervised manner. This paper focuses on semi-supervised learning for image classification, a recently very active research area (Li et al., 2019).

Semi-supervised learning is a transversal task for different domains including images (Oliver et al., 2018), audio (Zhang et al., 2016), time series (González et al., 2018), and text (Miyato et al., 2016). Recent approaches in image classification primarily focus on exploiting the consistency in the predictions for the same sample under different perturbations (consistency regularization) (Sajjadi et al., 2016; Li et al., 2019), while other approaches directly generate labels for the unlabeled data to guide the learning process (pseudo-labeling) (Lee, 2013; Iscen et al., 2019). Consistency regularization and pseudo-labeling approaches apply different strategies such as a warm-up phase using labeled data (Sajjadi et al., 2016; Laine & Aila, 2017; Tarvainen & Valpola, 2017; Qiao et al., 2018; Iscen et al., 2019), uncertainty weighting (Shi et al., 2018; Li et al., 2019), adversarial attacks (Miyato et al., 2018; Qiao et al., 2018; Iscen et al., 2019), or graph-consistency (Luo et al., 2018; Iscen et al., 2019). These strategies deal with confirmation bias (Li et al., 2019), also known as the noise accumulation problem (Zhang et al., 2016). This bias stems from using incorrect predictions

on unlabeled data for training in subsequent epochs and, thereby increasing confidence in incorrect predictions and producing a model that will tend to resist new changes.

This paper explores pseudo-labeling for semi-supervised deep learning from the network predictions and shows that, contrary to previous evidences on pseudo-labeling capabilities (Oliver et al., 2018), simple modifications to prevent confirmation bias lead it to state-of-the-art performance. We adapt the approach proposed by Tanaka et al. (2018) in the context of label noise and apply it exclusively on unlabeled samples. Experiments show that this naive pseudo-labeling is limited by confirmation bias as prediction errors are fit by the network. To deal with this issue, we propose to use mixup augmentation (Zhang et al., 2018) as an effective regularization that helps calibrate deep neural networks (Thulasidasan et al., 2019) and, therefore, alleviates confirmation bias. We find that mixup alone does not guarantee robustness against confirmation bias when reducing the amount of labeled samples or using certain network architectures (see Subsection 4.3), and show that, when properly introduced, dropout regularization (Srivastava et al., 2014) and data augmentation mitigates this issue. Our purely pseudo-labeling approach achieves state-of-the-art results (see Subsection 4.4) without requiring multiple networks (Tarvainen & Valpola, 2017; Qiao et al., 2018; Li et al., 2019; Verma et al., 2019), nor does it require over a thousand epochs of training to achieve peak performance in every dataset (Athiwaratkun et al., 2019; Berthelo et al., 2019), nor needs many (ten) forward passes for each sample (Li et al., 2019). Compared to other pseudo-labeling approaches, the proposed approach is simpler in that it does not require graph construction and diffusion (Iscen et al., 2019) or combination with consistency regularization methods (Shi et al., 2018), but still achieves state-of-the-art results. Additionally, we are the first to show that pseudo-labeling is a viable alternative for semi-supervised learning, as opposed to previous results in the state-of-the-art (Oliver et al., 2018).

2 Related work

Semi-supervised learning for image classification is an active research topic (Oliver et al., 2018); this section focuses on reviewing closely related methods, i.e. those using deep learning with mini-batch optimization over large image collections. Previous work on semi-supervised deep learning differ in whether they use consistency regularization or pseudo-labeling to learn from the unlabeled set (Iscen et al., 2019), while they all share the use of a cross-entropy loss (or similar) on labeled data.

Consistency regularization Imposes that the same sample under different perturbations must produce the same output. This idea was used in (Sajjadi et al., 2016) where they apply randomized data augmentation, dropout, and random max-pooling while forcing softmax predictions to be similar. A similar idea is applied by the so-called Π-model (Laine & Aila, 2017), which also extends the perturbation to different epochs during training, i.e. the current prediction for a sample has to be similar to an ensemble of predictions of the same sample in the past. Here the different perturbations come from networks at different states, dropout, and data augmentation. In (Tarvainen & Valpola, 2017), the temporal ensembling method is interpreted as a teacher-student problem where the network is both a teacher that produces targets for the unlabeled data as a temporal ensemble, and a student that learns the generated targets by imposing the consistency regularization. Tarvainen & Valpola (2017) naturally re-define the problem to deal with confirmation bias by separating the teacher and the student. The teacher is defined as a different network with similar architecture whose parameters are updated as an exponential moving average of the student network weights during training. This method is extended in (Li et al., 2019), where they apply an uncertainty weight over the unlabeled samples to incrementally learn from the unlabeled samples with low uncertainty, with uncertainty defined as the variance or entropy of the predictions for each sample under random perturbations. Additionally, Miyato et al. (2018) use Virtual Adversarial Training (VAT) to carefully introduce perturbations to data samples as adversarial noise and later impose consistency regularization on the predictions. More recently, Luo et al. (2018) propose to use a contrastive loss on the predictions as a regularization that forces predictions to be similar (different) when they are from the same (different) class. This method extends the consistency regularization previously considered only in-between the same data samples to in-between different samples. Their method can naturally be combined with (Tarvainen & Valpola, 2017) or (Miyato et al., 2018) to boost their performance. Similarly, Verma et al. (2019) propose Interpolation Consistency Training (ICT), a method inspired by (Zhang et al., 2018) that encourage predictions at interpolated unlabeled samples to be consistent with the interpolated predictions of individual samples. Also, authors in (Berthelo et al., 2019) apply

consistency regularization by guessing low-entropy labels, generating data-augmented unlabeled examples and mixing labeled and unlabeled examples using mixup (Zhang et al., 2018). Both (Verma et al., 2019) and (Berthelo et al., 2019) adopt (Tarvainen & Valpola, 2017) to estimate the targets used in the consistency regularization of unlabeled samples.

Co-training (Qiao et al., 2018) combines several ideas from the previous works, using two (or more) networks trained simultaneously to agree on their predictions (consistency regularization) and disagree on their errors. Errors are defined as different predictions when exposed to adversarial attacks, thus forcing different networks to learn complementary representations for the same samples. Recently, Chen et al. (2018) measure the consistency between the current prediction and an additional prediction for the same sample given by an external memory module that keeps track of previous representations of a sample. They additionally introduce an uncertainty weighting of the consistency term to reduce the contribution of uncertain predictions given by the memory module. Consistency regularization methods such as Π-model (Laine & Aila, 2017), mean teachers (Tarvainen & Valpola, 2017), and VAT (Miyato et al., 2018) have all been shown to benefit from stochastic weight averaging (SWA) method (Athiwaratkun et al., 2019). SWA averages network parameters at different training epochs to move the SGD solution on borders of flat loss regions to their center and improve generalization.

Pseudo-labeling Seeks the generation of labels or pseudo-labels for unlabeled samples to guide the learning process. An early attempt at pseudo-labeling proposed in (Lee, 2013) uses the network predictions as labels. However, they constrain the pseudo-labeling to a fine-tuning stage, i.e. there is a pre-training or warm-up, as with the consistency regularization approaches. A recent pseudo-labeling approach proposed in (Shi et al., 2018) uses the network class prediction as hard labels for the unlabeled samples. They also introduce an uncertainty weight for each sample loss, it being higher for samples that have distant k-nearest neighbors in terms of feature representation distance. They further include a loss term to encourage intra-class compactness and inter-class separation, and a consistency term between samples with different perturbations. They combine their method with mean teachers (Tarvainen & Valpola, 2017) to achieve state-of-the-art performance. Finally, a recently published work (Iscen et al., 2019) implements pseudo-labeling through graph-based label propagation. The method alternates between two steps: training from labeled and pseudo-labeled data and using the representations of the network to build a nearest neighbor graph where label propagation is applied to refine hard pseudo-labels on unlabeled images. They further add an uncertainty score for every sample (softmax prediction entropy based) and class (class population based) to deal, respectively, with the unequal confidence in network predictions over unlabeled samples and class-imbalance.

3 PSEUDO-LABELING

We formulate semi-supervised image classification as the task to learn a model $h_\theta(x)$ from a set of N training examples $\mathcal{D}$. These samples are split into the unlabeled set $\mathcal{D}_u = \{x_i\}_{i=1}^{N_u}$ and the labeled set $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$ with $y_i \in \{0,1\}^C$ being the one-hot encoding ground-truth label for C classes corresponding to x_i and $N = N_l + N_u$. In our case, h_θ is a CNN and θ represents the model parameters (weights and biases). As we seek to perform pseudo-labeling for the N_u unlabeled samples, we assume that a pseudo-label $\tilde{y}$ is available for these samples. We can then reformulate the problem as training using $\tilde{\mathcal{D}} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$, being $\tilde{y} = y$ for the N_l labeled samples.

The CNN parameters θ can be optimized using categorical cross-entropy:

$$\ell^*(\theta) = -\sum_{i=1}^{N} \tilde{y}_i^T \log\left(h_\theta(x_i)\right), \tag{1}$$

where $h_\theta(x)$ are the softmax probabilities produced by the model and $\log(\cdot)$ is applied element-wise. A key decision is how to generate the pseudo-labels $\tilde{y}$ for the N_u unlabeled samples. Previous approaches have used hard pseudo-labels (i.e. one-hot vectors) directly using the network output class (Lee, 2013; Shi et al., 2018) or the class estimated using label propagation on a nearest neighbor graph (Iscen et al., 2019). We adopt the former approach, but use soft pseudo-labels, as we have seen this outperforms hard labels, confirming the observations noted in (Tanaka et al., 2018) in the context of relabeling when learning with label noise. In particular, we store the softmax predictions $h_\theta(x_i)$ of the network in every mini-batch of an epoch and use them to modify the soft pseudo-label $\tilde{y}$ for

the N_u unlabeled samples at the end of the epoch. We proceed as described from the second to the last training epoch, while in the first epoch we use the softmax predictions for the unlabeled samples from a model trained in a 10 epochs warm-up phase using the labeled data subset $\mathcal{D}_u$.

We use the two regularizations applied in (Tanaka et al., 2018) to improve convergence. The first regularization deals with the difficulty of converging at early training stages when the network's predictions are mostly incorrect and the CNN tends to predict the same class to minimize the loss. Assignment of all samples to a single class is discouraged by adding:

$$R_A = \sum_{c=1}^{C} p_c \log\left(\frac{p_c}{\overline{h}_c}\right), \tag{2}$$

where p_c is the prior probability distribution for class c and $\overline{h}_c$ denotes the mean softmax probability of the model for class c across all samples in the dataset. As in (Tanaka et al., 2018), we assume a uniform distribution $p_c = 1/C$ for the prior probabilities (R_A stands for all classes regularization) and approximate $\overline{h}_c$ using mini-batches. The second regularization is needed to concentrate the probability distribution of each soft pseudo-label on a single class, thus avoiding the local optima in which the network might get stuck due to a weak guidance:

$$R_H = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} h_\theta^c(x_i)\log\left(h_\theta^c(x_i)\right), \tag{3}$$

where $h_\theta^c(x_i)$ denotes the c class value of the softmax output $h_\theta(x_i)$ and again using mini-batches (i.e. N is replaced by the mini-batch size) to approximate this term. This second regularization is the average per-sample entropy (R_H stands for entropy regularization), a well-known regularization in semi-supervised learning (Grandvalet & Bengio, 2004). Finally, the total semi-supervised loss is:

$$\ell = \ell^* + \lambda_A R_A + \lambda_H R_H, \tag{4}$$

where λ_A and λ_H control the contribution of each regularization term. We stress that this pseudo-labeling approach adapted from Tanaka et al. (2018) is far from the state-of-the-art for semi-supervised learning (see Subsection 4.2), and are the mechanisms proposed in Subsection 3.1 which make pseudo-labeling a suitable alternative.

3.1 Confirmation bias

Network predictions are, of course, sometimes incorrect. This situation is reinforced when incorrect predictions are used as labels for unlabeled samples, as it is the case in pseudo-labeling. Overfitting to incorrect pseudo-labels predicted by the network is known as confirmation bias. It is natural to think that reducing the confidence of the network on its predictions might alleviate this problem and improve generalization. Recently, mixup data augmentation (Zhang et al., 2018) introduced a strong regularization technique that combines data augmentation with label smoothing, which makes it potentially useful to deal with this bias. Mixup trains on convex combinations of sample pairs (x_p and x_q) and corresponding labels (y_p and y_q):

$$x = \delta x_p + (1-\delta)x_q, \tag{5}$$

$$y = \delta y_p + (1-\delta)y_q, \tag{6}$$

where $\delta \in \{0,1\}$ is randomly sampled from a beta distribution $\mathcal{B}e\left(\alpha,\beta\right)$, with $\alpha = \beta$ (e.g. $\alpha = 1$ uniformly selects δ). This combination regularizes the network to favor linear behavior in-between training samples, reducing oscillations in regions far from them. Additionally, Eq. 6 can be re-interpreted in the loss as $\ell^* = \delta\ell_p^* + (1-\delta)\ell_q^*$, thus re-defining the loss ℓ^* used in Eq. 4 as:

$$\ell^* = -\sum_{i=1}^{N} \delta\left[\tilde{y}_{i,p}^T \log\left(h_\theta(x_i)\right)\right] + (1-\delta)\left[\tilde{y}_{i,q}^T \log\left(h_\theta(x_i)\right)\right]. \tag{7}$$

As shown in (Thulasidasan et al., 2019), overconfidence in deep neural networks is a consequence of training on hard labels and it is the label smoothing effect from randomly combining y_p and y_q during mixup training that reduces prediction confidence and improves model calibration. In the semi-supervised context with pseudo-labeling, using soft-labels and mixup reduces overfitting to

model predictions, which is especially important for unlabeled samples whose predictions are used as soft-labels. Note that training with mixup generates softmax outputs $h_\theta(x)$ for mixed inputs x, thus requiring a second forward pass with the original images to compute unmixed predictions. Subsections 4.2 and 4.3 experimentally show that mixup and other techniques (minimum number of samples per mini-batch, dropout, and data augmentation) reduce confirmation bias and make pseudo-labeling a suitable alternative to consistency regularization methods.

4 EXPERIMENTAL WORK

4.1 DATASETS AND TRAINING

We use three image classification datasets, CIFAR-10/100 (Krizhevsky & Hinton, 2009), and Mini-ImageNet (Vinyals et al., 2016), to validate our approach. Part of the training images are labeled and the remaining are unlabeled. Following (Oliver et al., 2018), we use an independent validation set of 5K samples for CIFAR-10/100 for studying hyperparameters in Subsections 4.2 and 4.3. However, as done in (Athiwaratkun et al., 2019), we add the 5K samples back to the training set for comparison with the state-of-the-art in Subsection 4.4, where we report test results (model from the best epoch).

CIFAR-10 and CIFAR-100 These datasets contain 10 and 100 classes, both with 50K color images for training and 10K for testing with resolution 32×32. We perform experiments with a number of labeled images $N_l = 0.25$K, 0.5K, 1K, and 4K (4K and 10K) for CIFAR-10 (CIFAR-100). We use the well-known "13-CNN" architecture (Athiwaratkun et al., 2019) for CIFAR-10/100. We also experiment with a Wide ResNet-28-2 (WR-28) (Oliver et al., 2018) and a PreAct ResNet-18 (PR-18) (Zhang et al., 2018) in Subsection 4.3 to study the generalization to different architectures.

Mini-ImageNet We emulate the semi-supervised learning setup Mini-ImageNet (Vinyals et al., 2016) (a subset of the well-known ImageNet (Deng et al., 2009) dataset) used in (Iscen et al., 2019). Train and test sets of 100 classes and 600 color images per class with resolution 84 × 84 are selected from ImageNet, as in (Ravi & Larochelle, 2017). 500 (100) images per-class are kept for train (test) splits. The train and test sets therefore contain 50k and 10k images. As with CIFAR-100, we experiment with a number of labeled images $N_l = 4$K and 10K. Following (Iscen et al., 2019), we use a ResNet-18 (RN-18) architecture (He et al., 2016a).

Hyperparameters We use the typical configuration for CIFAR-10/100 (Laine & Aila, 2017) and the same for Mini-ImageNet. Image normalization using dataset mean and standard deviation together and subsequent data augmentation (Laine & Aila, 2017) by random horizontal flips and random 2 (6) pixel translations for CIFAR (Mini-ImageNet). Additionally, color jitter is applied as in (Asano et al., 2019) in Subsections 4.3 and 4.4 for higher robustness against confirmation bias. We train using SGD with a momentum of 0.9, a weight decay of 10^{-4}, and batch size of 100. Training always starts with a high learning rate (0.1 in CIFAR and 0.2 in Mini-ImageNet), dividing it by ten twice during training. We always train the model 400 epochs (reducing learning rate in epochs 250 and 350) and use 10 epoch warm-up with labeled data. We do not attempt careful tuning of the regularization weights λ_A and λ_H and just set them to 0.8 and 0.4 as done in (Tanaka et al., 2018). When using dropout, it is introduced between consecutive convolutional layers of ResNet blocks in WR-28, PR-18, and RN-18, while for 13-CNN we introduce it as in Laine & Aila (2017). Following (Athiwaratkun et al., 2019)[1], we use weight normalization (Salimans & Kingma, 2016) in all networks.

4.2 EFFECT OF MIXUP ON CONFIRMATION BIAS

This section demonstrates that carefully regularized pseudo-labeling is a suitable alternative for semi-supervised learning. Figure 1 illustrates our approach on the "two moons" toy data. Figure 1 (left) shows the limitations of a naive pseudo-labeling adapted from (Tanaka et al., 2018), which fails to adapt to the structure in the unlabelled examples and results in a linear decision boundary. Figure 1 (middle) shows the effect of mixup, which alleviates confirmation bias to better model the structure and gives a smoother boundary. Figure 1 (right) shows that combining mixup with a minimum number of labeled samples k per mini-batch improves the semi-supervised decision boundary.

[1] `https://github.com/benathi/fastswa-semi-sup`

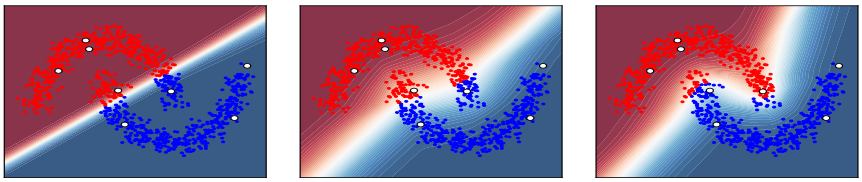

Figure 1: Pseudo-labeling in the "two moons" data (4 labels/class) for 1000 samples. From left to right: no mixup, mixup, and mixup with a minimum number of labeled samples per mini-batch. We use an NN classifier with one hidden layer with 50 hidden units as in (Miyato et al., 2018).

Table 1: Confirmation bias alleviation using mixup and a minimum number of k labeled samples per mini-batch. Left: Validation error for naive pseudo-labeling without mixup (C), mixup (M), and alternatives with minimum k. Right: Study of the effect of k on the validation error.

	CIFAR-10		CIFAR-100
Labeled images	500	4000	4000
C	52.44	11.40	48.54
C* ($k = 16$)	35.08	10.90	46.60
M	32.10	7.16	41.80
M *($k = 16$)	**13.68**	**6.90**	**38.78**

	CIFAR-10		CIFAR-100
Labeled images	500	4000	4000
$k = 8$	**13.14**	7.18	42.32
$k = 16$	13.68	**6.90**	**38.78**
$k = 32$	14.58	7.06	39.62
$k = 64$	19.40	8.20	46.28

Naive pseudo-labeling leads to overfitting the network predictions and high training accuracy in CIFAR-10/100. Table 1 (left) reports mixup effect in terms of validation error. Naive pseudo-labeling leads to an error of 11.40/48.54 for CIFAR-10/100 when training with cross-entropy (C) loss for 4000 labels. This error can be greatly reduced when using mixup (M) to 7.16/41.80. However, when further reducing the number of labels to 500 in CIFAR-10, M is insufficient to ensure low-error (32.10). We propose to set a minimum number of samples k per mini-batch to tackle the problem. Table 1 (right) studies this parameter k when combined with mixup, showing that 16 samples per mini-batch works well for both CIFAR-10 and CIFAR-100, dramatically reducing error in all cases (e.g. in CIFAR-10 for 500 labels error is reduced from 32.10 to 13.68).

Confirmation bias causes dramatic increase the certainty of incorrect predictions during training. To demonstrate this behavior we compute the average cross-entropy of the softmax output with a uniform $\mathcal{U}$ across the classes in every epoch t for all incorrectly predicted samples $\{x_{m_t}\}_{m_t=1}^{M_t}$ as: $r_t = -\frac{1}{M_t}\sum_{m_t=1}^{M_t} \mathcal{U}^T \log\left(h_\theta(x_{m_t})\right)$. Figure 2 shows that mixup and minimum k are effective regularizers for reducing r_t, i.e. confirmation bias is reduced. We also experimented with using label noise regularizations (Xie et al., 2016), but setting a minimum k proved more effective.

4.3 Generalization to different architectures

There are examples in the recent literature (Kolesnikov et al., 2019) where moving from one architecture to another modifies the belief of which methods have a higher potential. Kolesnikov et al. (2019) show that skip-connections in ResNet arquitectures play a key role on the quality of learned representations, while most approaches in previous literature were systematically evaluated using AlexNet (A. Krizhevsky, 2012). Ulyanov et al. (2018) showed that different architectures lead different and useful image priors, highlighting the importance of exploring different networks. We, therefore, test our method with two more architectures: a Wide ResNet-28-2 (WR-28) (S. Zagoruyko, 2016) typically used in semi-supervised learning (Oliver et al., 2018) (1.5M parameters) and a PreAct ResNet-18 (PR-18) (He et al., 2016b) used in the context of label noise (Zhang et al., 2018) (11M parameters). Table 2 presents the results for the 13-CNN (AlexNet-type) and these network architectures (ResNet-type). Our pseudo-labeling with mixup and $k = 16$ (M*) works well for 4000 and 500 labels across architectures, except for 500 labels for WR-28 where there is large error increase (29.50). This is due to a stronger confirmation bias in which labeled samples are not properly learned, while incorrect pseudo-labels are fit. Interestingly, PR-18 (11M) is more robust to confirmation bias than WR-28 (1.5M), while the 13-layer network (3M) has fewer parameters than PR-18 and achieves

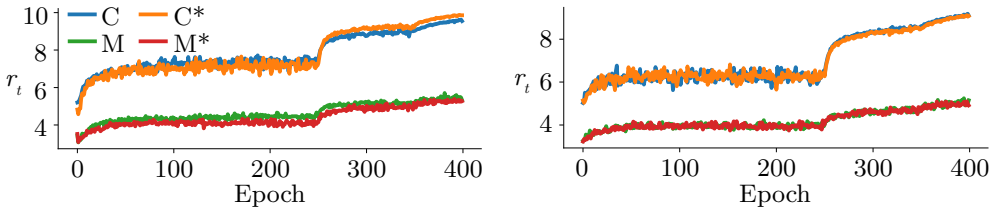

Figure 2: Example of certainty of incorrect predictions r_t during training when using 500 (left) and 4000 (right) labeled images in CIFAR-10. Moving from cross-entropy (C) to mixup (M) reduces r_t, whereas adding a minimum number of samples per mini-batch (*) also helps in 500 labels, where M* (with slightly lower r_t than M) is the only configuration that converges, as shown in Table 1 (left).

Table 2: Validation error across architectures is stabilized using dropout p and data augmentation (A).

	M*		M* + $p = 0.1$		M* + $p = 0.3$		M* + $p = 0.1$ + A	
Labeled images	500	4000	500	4000	500	4000	500	4000
13-layer	13.68	6.90	12.62	6.58	11.94	6.66	**9.16**	**6.22**
WR-28	29.50	**6.40**	14.14	7.06	30.56	11.44	**10.94**	6.74
PR-18	**13.90**	**5.94**	14.78	5.90	14.78	6.62	14.96	6.32

better performance. This suggests that the network architecture plays and important role, being a relevant prior for semi-supervised learning with few labels.

We found that dropout (Srivastava et al., 2014) and data augmentation is needed for good performance across all architectures. Table 2 shows that dropout $p = 0.1, 0.3$ helps in achieving better convergence in CIFAR-10, whereas adding color jitter as additional data augmentation (details in Subsection 4.1) further contributes to error reduction. Note that the quality of pseudo-labels is key, so it is essential to disable dropout to prevent corruption when computing these in the second forward pass. We similarly disable data augmentation in the second forward pass, which consistently improves performance. This configuration is used for comparison with the state-of-the-art in Subsection 4.4.

4.4 Comparison with the state-of-the-art

We compare our pseudo-labeling approach against related work that makes use of the 13-CNN (Tarvainen & Valpola, 2017) in CIFAR-10/100: Π model (Laine & Aila, 2017), TE (Laine & Aila, 2017), MT (Tarvainen & Valpola, 2017), Π model-SN (Luo et al., 2018), MA-DNN (Chen et al., 2018), Deep-Co (Qiao et al., 2018), TSSDL (Shi et al., 2018), LP (Iscen et al., 2019), CCL (Li et al., 2019), fast-SWA (Athiwaratkun et al., 2019) and ICT (Verma et al., 2019). The table divides methods into those based on consistency regularization and pseudo-labeling. Note that we include pseudo-labeling approaches combined with consistency regularization ones (e.g. MT) in the consistency regularization set. The proposed approach clearly outperforms consistency regularization methods, as well as other purely pseudo-labeling approaches and their combination with consistency regularization methods in CIFAR-10/100. These results demonstrate the generalization of the proposed approach compared to other methods that fail when decreasing the number of labels. Furthermore, Table 4 (left) demonstrates that the proposed approach successfully scales to higher resolution images, obtaining an over 10 point margin on the best related work in Mini-ImageNet. Note that all supervised baselines are reported using the same data augmentation and dropout as in the proposed pseudo-labeling.

Table 4 (right) compares against recent consistency regularization approaches that use mixup. We achieve better performance than ICT (Verma et al., 2019), while being competitive with MM (Berthelo et al., 2019) for 500 and 4000 labels using WR-28. Regarding PR-18, we converge to reasonable performance for 4000 and 500 labels, whereas for 250 we do not. Finally, the 13-CNN robustly converges even for 250 labels where we obtain 9.37 test error (see Appendix A.1 for some details on different architectures convergence). Therefore, these results suggest that it is worth exploring the relationship between number of labels, dataset complexity and architecture type. As shown in Subsection 4.3, dropout and additional data augmentation help with 500 labels/class across

Table 3: Test error in CIFAR-10/100 for the proposed approach using the 13-CNN network. (*) denotes that we have run the algorithm. Bold indicates lowest error. We report average and standard deviation of 3 runs with different labeled/unlabeled splits.

	CIFAR-10			CIFAR-100	
Labeled images	500	1000	4000	4000	10000
Supervised (C)*	43.64 ± 1.21	34.83 ± 1.15	19.26 ± 0.26	54.49 ± 0.53	41.14 ± 0.26
Supervised (M)*	37.60 ± 0.65	28.59 ± 1.21	15.94 ± 0.26	52.70 ± 0.28	39.42 ± 0.37
Consistency regularization methods					
Π model	-	-	12.36 ± 0.31	-	39.19 ± 0.36
TE	-	-	12.16 ± 0.24	-	38.65 ± 0.51
MT	27.45 ± 2.64	19.04 ± 0.51	11.41 ± 0.25	45.36 ± 0.49	36.08 ± 0.51
Π model-SN	-	21.23 ± 1.27	11.00 ± 0.13	-	37.97 ± 0.29
MA-DNN	-	-	11.91 ± 0.22	-	34.51 ± 0.61
Deep-Co	-	-	9.03 ± 0.18	-	38.77 ± 0.28
MT-TSSDL	-	18.41 ± 0.92	9.30 ± 0.55		
MT-LP	24.02 ± 2.44	16.93 ± 0.70	10.61 ± 0.28	43.73 ± 0.20	35.92 ± 0.47
MT-CCL	-	16.99 ± 0.71	10.63 ± 0.22	-	34.81 ± 0.52
MT-fast-SWA	-	15.58 ± 0.12	9.05 ± 0.21	-	34.10 ± 0.31
ICT	-	15.48 ± 0.78	7.29 ± 0.02	-	-
Pseudo-labeling methods					
TSSDL	-	21.13 ± 1.17	10.90 ± 0.23	-	-
LP	32.40 ± 1.80	22.02 ± 0.88	12.69 ± 0.29	46.20 ± 0.76	38.43 ± 1.88
Ours*	**8.80 ± 0.45**	**6.85 ± 0.15**	**5.97 ± 0.15**	**37.55 ± 1.09**	**32.15 ± 0.5**

Table 4: Test error in Mini-ImageNet (left) and CIFAR-10 with few labeled samples (right). (*) denotes that we have run the algorithm. Bold indicates lowest error. We report average and standard deviation of 3 runs with different labeled/unlabeled splits.

Labeled images	4000	10000
Supervised (C)*	75.69 ± 0.24	63.24 ± 0.33
Supervised (M)*	72.03 ± 0.21	59.96 ± 0.40
Consistency regularization methods		
MT	72.51 ± 0.22	57.55 ± 1.11
MT-LP	72.78 ± 0.15	57.35 ± 1.66
Pseudo-labeling methods		
LP	70.29 ± 0.81	57.58 ± 1.47
Ours*	**56.49 ± 0.51**	**46.08 ± 0.11**

Labeled images	250	500	4000
MM (WR-28)	**11.08 ± 0.87**	**9.65 ± 0.94**	**6.24 ± 0.06**
ICT* (WR-28)	52.19 ± 1.54	42.33 ± 0.08	7.26 ± 0.04
Ours* (WR-28)	24.81 ± 5.35	14.25 ± 0.86	6.28 ± 0.3
Ours* (13-CNN)	**9.37 ± 0.12**	**8.80 ± 0.45**	5.97 ± 0.15
Ours* (PR-18)	23.86 ± 4.82	12.16 ± 1.06	**5.86 ± 0.17**

architectures, but are insufficient for 250 labels. Better data augmentation (Ho et al., 2019) or self-supervised pre-training (Rebuffi et al., 2019) might overcome this challenge. Furthermore, hyperparameters such as the regularization weights $\lambda_A = 0.8$ and $\lambda_H = 0.4$ from Eq. 4 and the mixup α require further study. However, it is already interesting that a straightforward modification of pseudo-labeling, designed to tackle confirmation bias, gives a competitive semi-supervised learning approach, without any consistency regularization, and that future work should take this into account.

5 Conclusions

This paper presented a semi-supervised learning approach for image classification based on pseudo-labeling. We proposed to directly use the network predictions as soft pseudo-labels for unlabeled data together with mixup augmentation, a minimum number of labeled samples per mini-batch, dropout and data augmentation to alleviate confirmation bias. This conceptually simple approach outperforms related work in CIFAR-10/100 and Mini-ImageNet datasets, demonstrating that pseudo-labeling is a suitable alternative to the dominant approach in recent literature: consistency-regularization. The proposed approach is, to the best of our knowledge, both simpler and more accurate than most recent approaches. Future research will explore careful hyperparameter selection and larger-scale datasets.

REFERENCES

I. Sutskever G. Hinton A. Krizhevsky. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2012. 4.3

E. Arazo, D. Ortego, P. Albert, N. O'Connor, and K. McGuinness. Unsupervised Label Noise Modeling and Loss Correction. In *International Conference on Machine Learning (ICML)*, 2019. 1

Y.M. Asano, C. Rupprecht, and A. Vedaldi. Surprising Effectiveness of Few-Image Unsupervised Feature Learning. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 4.1

B. Athiwaratkun, M. Finzi, P. Izmailov, and A.G. Wilson. There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 2, 4.1, 4.1, 4.1, 4.4

D. Berthelo, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel. MixMatch: A Holistic Approach to Semi-Supervised Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1, 2, 4.4

Y. Chen, X. Zhu, and S. Gong. Semi-Supervised Deep Learning with Memory. In *European Conference on Computer Vision (ECCV)*, 2018. 2, 4.4

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 4.1

S. Gidaris, P. Singh, and N. Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. In *International Conference on Learning Representations (ICLR)*, 2018. 1

M. González, C. Bergmeir, I. Triguero, Y. Rodríguez, and J.M. Benítez. Self-labeling techniques for semi-supervised time series classification: an empirical study. *Knowledge and Information Systems*, 55(2):493–528, 2018. 1

Y. Grandvalet and Y. Bengio. Semi-supervised Learning by Entropy Minimization. In *International Conference on Neural Information Processing Systems (NIPS)*, 2004. 3

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016a. 1, 4.1

K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. In *European Conference on Computer Vision (ECCV)*, 2016b. 4.3

D. Ho, E. Liang, X. Chen, I. Stoica, and P. Abbeel. Population Based Augmentation: Efficient Learning of Augmentation Policy Schedules. In *International Conference on Machine Learning (ICML)*, 2019. 4.4

A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Label Propagation for Deep Semi-supervised Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 2, 3, 4.1, 4.4

C. Kim, F. Li, and J.M. Rehg. Multi-object Tracking with Neural Gating Using Bilinear LSTM. In *European Conference on Computer Vision (ECCV)*, 2018. 1

A. Kolesnikov, X. Zhai, and L. Beyer. Revisiting Self-Supervised Visual Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4.3

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 4.1

S. Laine and T. Aila. Temporal Ensembling for Semi-Supervised Learning. In *International Conference on Learning Representations (ICLR)*, 2017. 1, 2, 4.1, 4.4

D.H. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *International Conference on Machine Learning Workshops (ICMLW)*, 2013. 1, 2, 3

W. Li, L. Wang, W. Li, E. Agustsson, and L. Van Gool. WebVision Database: Visual Learning and Understanding from Web Data. *arXiv: 1708.02862*, 2017. 1

Y. Li, L. Liu, and R.T Tan. Certainty-Driven Consistency Loss for Semi-supervised Learning. *arXiv: 1901.05657*, 2019. 1, 2, 4.4

T.Y Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal Loss for Dense Object Detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

W. Liu, R. Lin, Z. Liu, L. Liu, Z. Yu, B. Dai, and L. Song. Learning towards Minimum Hyperspherical Energy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1

X. Liu, J. Van De Weijer, and A. D. Bagdanov. Exploiting Unlabeled Data in CNNs by Self-supervised Learning to Rank. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

Y. Luo, J. Zhu, M. Li, Y. Ren, and B. Zhang. Smooth Neighbors on Teacher Graphs for Semi-Supervised Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 4.4

T. Miyato, A.M. Dai, and I. Goodfellow. Adversarial Training Methods for Semi-Supervised Text Classification. *arXiv: 1605.07725*, 2016. 1

T. Miyato, S. Maeda, S. Ishii, and M. Koyama. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 1, 2, 1

A. Oliver, A. Odena, C.A. Raffel, E.D. Cubuk, and I. Goodfellow. Realistic Evaluation of Deep Semi-Supervised Learning Algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1, 2, 4.1, 4.1, 4.3

S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille. Deep Co-Training for Semi-Supervised Image Recognition. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 4.4

S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations (ICLR)*, 2017. 4.1

S.-A. Rebuffi, S. Ehrhardt, K. Han, A. Vedaldi, and A. Zisserman. Semi-Supervised Learning with Scarce Annotations. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 4.4

N. Komodakis S. Zagoruyko. Wide Residual Networks. In *British Machine Vision Conference (BMVC)*, 2016. 4.3

M. Sajjadi, M. Javanmardi, and T. Tasdizen. Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 1, 2

T. Salimans and D.P. Kingma. Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 4.1

W. Shi, Y. Gong, C. Ding, Z. MaXiaoyu Tao, and N. Zheng. Transductive Semi-Supervised Deep Learning using Min-Max Features. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 3, 4.4

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15: 1929–1958, 2014. 1, 4.3

D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa. Joint Optimization Framework for Learning with Noisy Labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3, 3, 3, 4.1, 4.2

A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1, 2, 2, 4.4

S. Thulasidasan, G. Chennupati, J. Bilmes, T. Bhattacharya, and S. Michalak. On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks. *arXiv: 1905.11001*, 2019. 1, 3.1

D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep Image Prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4.3

V. Verma, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz. Interpolation Consistency Training for Semi-Supervised Learning. In *Journal of Artificial Intelligence Research (IJCAI)*, 2019. 1, 2, 4.4

O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra. Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 4.1, 4.1

M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 1

L. Xie, J. Wang, Z. Wei, M. Wang, and Q. Tian. DisturbLabel: Regularizing CNN on the Loss Layer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4.2

S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. In *European Conference on Computer Vision (ECCV)*, September 2018. 1

H. Zhang, M. Cisse, Y.N. Dauphin, and D. Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 2, 3.1, 4.1, 4.3

Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schüller. Enhanced semi-supervised learning for multimodal emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016. 1

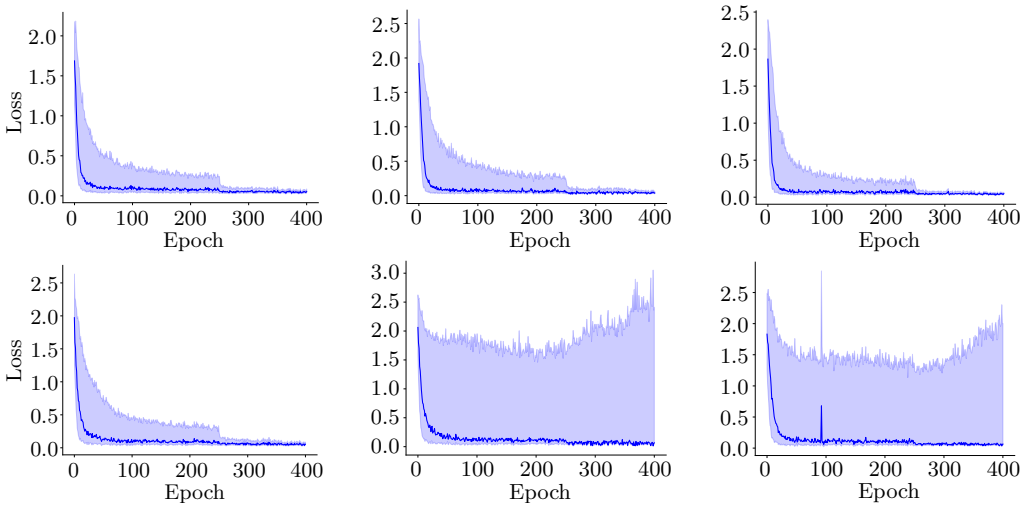

Figure 3: Cross-entropy loss for labeled samples. First (second) row show 500 (250) labels in CIFAR-10. From left to right: 13-CNN, WR-28 and PR-18. The heavy lines represent the median losses and the shaded areas are the interquartile ranges.

A APPENDIX

A.1 CONVERGENCE FOR FEW LABELS

Figure 3 presents the cross-entropy loss for labeled samples when training with 13-CNN, WR-28 and PR-18 and using 500 and 250 labels in CIFAR-10. This loss is a good indicator of a robust convergence to reasonable performance as the interquartile range for cases failing (250 labels for WR-28 and PR-18) is much higher.