Blind Knowledge Distillation for Robust Image Classification

Timo Kaiser , **Lukas Ehmann** , **Christoph Reinders** and **Bodo Rosenhahn**
Institute for Information Processing, Leibniz University Hannover
{kaiser, ehmannlu, reinders, rosenhahn}@tnt.uni-hannover.de

Abstract

Optimizing neural networks with noisy labels is a challenging task, especially if the label set contains real-world noise. Networks tend to generalize to reasonable patterns in the early training stages and overfit to specific details of noisy samples in the latter ones. We introduce *Blind Knowledge Distillation* - a novel teacher-student approach for learning with noisy labels by masking the ground truth related teacher output to filter out potentially corrupted 'knowledge' and to estimate the tipping point from generalizing to overfitting. Based on this, we enable the estimation of noise in the training data with Otsu's algorithm. With this estimation, we train the network with a modified weighted cross-entropy loss function. We show in our experiments that *Blind Knowledge Distillation* detects overfitting effectively during training and improves the detection of clean and noisy labels on the recently published CIFAR-N dataset. Code is available at GitHub[1].

1 Introduction

Learning with noisy labels is a challenging task in image classification. It is well known that label noise leads to heavy performance drops with standard classification methods [Song *et al.*, 2022]. The goal of learning with noisy labels is therefore to train a classification model with labelled training images and achieve high classification performance on unseen test images, even if the labels for training are noisy and corrupted. Labels are noisy because humans are naturally unable to classify images perfectly due to ambiguous images, individual human bias, pressure of time, or various other reasons. Many modern methods [Liu *et al.*, 2022b; Rawat and Wang, 2017] are trained on large and potentially noisy datasets and thus it is an interest of the community to make classification robust against noisy labels.

To evaluate the robustness of methods for learning with noisy labels, clean image datasets like *CIFAR* [Krizhevsky and Hinton, 2009], *Clothing1M* [Xiao *et al.*, 2015], or *WebVision* [Li *et al.*, 2017] are synthetically corrupted by randomly flipping label annotations either symmetrically without constraints or asymmetrically with predefined rules to mimic realistic label noise. However, Wei *et al.* [2022] shows that synthetic label noise has different behaviour compared to real-world label noise and is thus not an ideal choice to evaluate robust learning. To close this gap, Wei *et al.* have made great efforts and presented *CIFAR-N* with multiple newly annotated ground truth labels for *CIFAR* with human-induced label noise. With these new annotations, robust learning can be evaluated more realistically.

In this paper, we introduce a novel method to detect the beginning of overfitting on sample details during training, that is usually roughly estimated as in [Li *et al.*, 2020], and present a simple but effective method to detect most likely corrupted la-

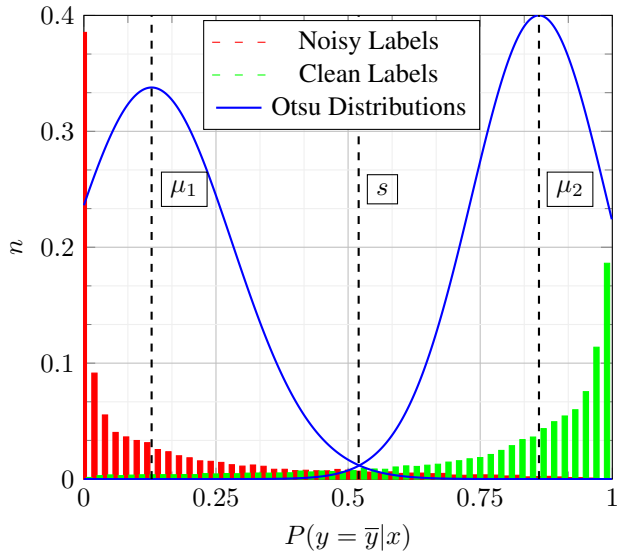


Figure 1: Distribution of ground truth label related probabilities $P_A(y = \overline{y}|x)$ at beginning of overfitting (tipping point) and the resulting gaussian distributions after Otsu's algorithm for the dataset *Worst*. Red bars show the normalized distribution of noisy labels and green bars of clean labels, respectively. Note that the gaussian distributions (blue) are scaled for visualization purposes. Our presented Blind Knowledge Distillation enables an adaptive noise estimation via the thresholds μ_1, s, and μ_2 and a robust learning with noisy labels.

[1]https://github.com/TimoK93/blind_knowledge_distillation

bels. Our method is inspired by *Knowledge Distillation* [Hinton *et al.*, 2015] for neural networks which extracts 'knowledge' from a teacher network to train a student network. Differently than usual, our student network is just trained with a subset of the teachers 'knowledge'. Specifically, it does not 'see' the 'knowledge' about the given and potentially corrupted ground truth labels by utilizing the teachers ground truth complementary logits. Therefore we call it *Blind Knowledge Distillation*. Based on the detected noisy labels, we propose a simple but effective loss-correction method to train the teacher model robustly with label noise. We perform extensive experiments on CIFAR-10N and the results show that *Blind Knowledge Distillation*

- successfully estimates the tipping point from fitting to general patterns to (over)fitting to sample details,
- is an effective method to estimate the likelihood of labels being noisy,
- and improves the classification accuracy while training with high noise levels.

2 Related Work

Methods in the field of robust learning to tackle noisy labels can be divided into label correction, loss correction, and refined strategies [Song *et al.*, 2022; Wang *et al.*, 2019]. In this section, we contextualize the latest methods of the *CIFAR-N* leaderboard based on the aforementioned categories.

Label correction is an approach in which the given ground truth labels are dynamically changed during optimization to obtain labels of higher quality. *SOP* [Liu *et al.*, 2022a] performs label correction by optimizing the ground truth labels with Stochastic Gradient Descent (SGD) w.r.t. the classification loss. It alternates between the update of model weights and the update of additional soft-label weights.

Another approach is loss correction which is usually applied by weighting the loss term or adding a new loss for each sample in the training dataset. The methods *CORES* [Cheng *et al.*, 2021] and *ELR* [Liu *et al.*, 2020] add a regularization term to the standard cross-entropy (CE) loss to penalize likely corrupted labels. *PeerLoss* [Liu and Guo, 2020] introduces and minimizes peer loss functions between randomly selected samples. *CAL* [Zhu *et al.*, 2021] extends this approach and estimates the covariances between noise rates and their bayes optimal label.

The last category tackles noisy labels by using refined strategies. *CoTeaching* [Han *et al.*, 2018] trains a neural network with samples with high confident predictions of a second network, and vice versa. *DivideMix* [Li *et al.*, 2020] and *PES* [Bai *et al.*, 2021] split the dataset in clean and corrupted subsets and apply semi-supervised learning methods. In detail, *DivideMix* trains two independent neural networks and splits the set of one network based on the predictions of the other network to avoid confirmation bias. In contrast to this, *PES* applies early stopping of the optimization to every network layer independently, instead of applying it to the whole network simultaneously, as usual.

Our method combines a refined strategy to detect most likely corrupted labels in the first stage and performs loss-correction in the second stage while incorporating the estimation of likely corrupted labels. While other methods manually define warm-up epochs, we adapt to the dataset and estimate the optimal stopping point for the standard CE training. Instead of applying extensive semi-supervised augmentation methods, we apply a simple sample dependent loss correction.

3 Preliminaries

Given a set of annotated image samples X and a set of classes C, the task of image classification is to assign every sample x from $(x, \overline{y}) \in X$ to the correct class label $y = c \in C$ without prior knowledge of the correct class label y and a potentially noisy annotation $\overline{y}$. Modern methods use neural networks $f(\Phi, x)$ to estimate the probability distribution $P(y = c|x)$ for every class c [Liu *et al.*, 2022b; Li *et al.*, 2020; Liu *et al.*, 2022a; Cheng *et al.*, 2021; He *et al.*, 2016a], in which Φ denotes a set of trainable network parameters. More specifically, neural networks predict a logit vector $\vec{l} \in \mathbb{R}^{|C|}$ with a logit l_c for every class and transform it into probabilities with the softmax function

$$P(y = c|x) = \frac{e^{l_c}}{\sum_{i \in C} e^{l_i}}. \tag{1}$$

Finally, the class c with the highest probability $P(y = c|x)$ is assumed to be the correct label y.

The task is to define the network architecture of f and the training strategy to optimize Φ, so that $f(\Phi, x)$ predicts a satisfying distribution $P(y = c|x)$ in which the correct class has the highest probability. Most methods optimize Φ with large manually annotated image datasets and minimize the categorical cross entropy (CE) loss objective

$$L_{\text{CE}} = \frac{1}{|X|} \sum_{(x,\overline{y}) \in X} -\log\big(P(y = \overline{y}|x)\big) \tag{2}$$

or one of its derivatives.

Extending the task of image classification, the challenging task of learning with noisy labels addresses the problem that the given ground truth labels $\overline{y}$ could be noisy and not the true labels $\overline{y} \neq y$. False ground truth labels dramatically impede the optimization of Φ. Thus, the goal is to train classifiers with an accuracy that is comparable to classifiers that would be optimized with clean labels $\overline{y} = y$. A second goal is to identify noisy labels $\overline{y} \neq y$ in the dataset.

The approach proposed in this paper addresses both tasks. Note that the method is iteratively trained with random sampled batches $X' \subset X$. We keep the notation of X in the next sections for simplicity, e.g. in Eq. (3).

4 Method

To enlarge the robustness of neural networks against label noise and to detect noisy labels, we present a novel training strategy to estimate the likelihood of every label being noisy and apply a weighted loss based on this.

First, we adapt the student-teacher architecture [Gou *et al.*, 2021] and introduce *Blind Knowledge Distillation* to extract generalized patterns from the data. Then, we present

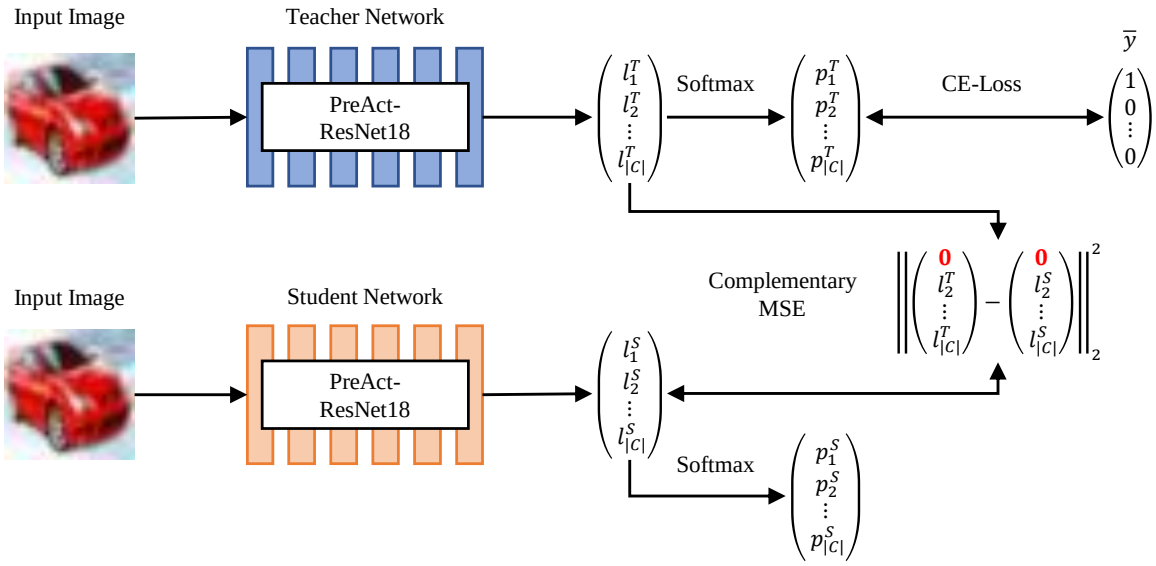


Figure 2: Our proposed *Blind Knowledge Distillation* framework. The teacher and student network share the same topology but have different weights. While the teacher network is trained by optimizing the standard CE loss, the student network is trained with the ground truth complementary logits of the teacher network and the mean squared error loss. The teacher network predicts the class membership probabilities $P_T(y = c|x)$ and the student network predicts the probabilities $P_S(y = c|x)$.

a method to detect the beginning of overfitting with the student network and enable the detection of noisy labels by estimating four confidence levels of being noisy. Finally, we optimize Φ with a robust training strategy to train the final classifier. All three steps are described in the following sections.

4.1 Blind Knowledge Distillation

Neural networks with large number of parameters Φ can memorize the training examples, so that $P(y = \overline{y}|x) \approx 1$ for every sample x in the training set X. Also, neural networks adapt simple patterns during the early optimization epochs and overfit to specific image details in the latter ones. As shown in [Liu *et al.*, 2020], valid patterns are learned by maximizing the logit $l_{y=\overline{y}}$ for clean samples in the first training stages. Subsequently to this early generalization, maximizing the logits $l_{y\neq\overline{y}}$ of corrupted samples degrades the classification accuracy. More important for this method is the phenomenon that the ground truth complementary logits $l_{c\neq\overline{y}}$ are also minimized in the latter stages.

To avoid the maximization of $l_{y\neq\overline{y}}$ for noisy labels, we create a new student-teacher architecture, in which the student only learns generalized patterns. In the student-teacher architecture, a student model is trained with the output of a teacher model. This method is called *Knowledge Distillation* [Hinton *et al.*, 2015] and transfers the patterns that are encoded in the teacher model to the student. Model bias or wrong patterns from the teacher can also be transferred. To avoid this undesired transfer, we introduce the ground truth annotation complementary 'knowledge' by removing all information that is immediately connected to a potentially corrupted ground truth label $\overline{y}$.

Unlike the usual knowledge distillation architecture, our models share the same topology $f(\Phi, x)$ but have different weights Φ_T (teacher) and Φ_S (student). The teacher model $f(\Phi_T, x)$ is trained with the standard CE loss (Eq. (2)). The student model is trained with the unlabelled ground truth complementary logits $l^T_{c\neq\overline{y}}$ derived from the teacher model and an extended but simple mean squared error loss

$$L_{\text{Stud}} = \frac{1}{|X|} \sum_{(x,\overline{y})\in X} \frac{1}{|C|-1} \sum_{\substack{c\in C \\ c\neq\overline{y}}} (l^T_c - l^S_c)^2 \qquad (3)$$

in which l^S_c denotes the logits of the student model. This loss function transfers the generalized patterns by imitating the output of the teacher model but without taking the potentially corrupted ground truth label $\overline{y}$ into account. The training architecture is visualized in Fig. 2.

Since high valued complementary logits $l_{c\neq\overline{y}}$ are minimized in the latter optimization stages after general features are learned, also the resulting probabilities after softmax (Eq. (1)) converge to a uniform distribution. Thus, we can identify approximately the training epoch, in which the neural network starts overfitting to specific sample details by monitoring the mean maximal probability

$$\hat{p}_{\text{max}} = \frac{1}{|X|} \sum_{x\in X} \max_{c\in C} \big(P(y = c|x)\big) \qquad (4)$$

of the students network. During training in epoch i, the *fitting-epoch* in which $\hat{p}^i_{\text{max}}$ is maximal can be certainly identified online with a delay of k epochs by checking if $\hat{p}^{i-k}_{\text{max}}$ is the maximum of the last $2k + 1$ epochs.

Furthermore, it shows that the student model has the ability to classify images comparable to the teacher model before the detail fitting starts. Thus, we modify our final classification probability by combining the teacher's

prediction $P_T(y = c|x)$ and the student's prediction $P_S(y = c|x)$ to the *agreement* probability

$$P_A(y = c|x) = \frac{P_T(y = c|x) \cdot P_S(y = c|x)}{\sum_{i \in C} P_T(y = i|x) \cdot P_S(y = i|x)}. \quad (5)$$

P_A enables the noise estimation in the dataset described in the following. The classification accuracy P_T, P_S, and P_A are elaborated more in detail in the experiments (see Sec. 5.4).

4.2 Adaptive Noise Estimation

The knowledge about the presence of noise can be used to apply loss correction. Unfortunately, this knowledge is not given, so we estimate the probability of a data sample to be noisy. We split the dataset into four subsets based on Otsu's algorithm [Otsu, 1979], in which the membership to a subset indicates the likelihood of being noisy. Given a set of data samples $(x, \overline{y}) \in X$ with their corresponding agreement probability $P_A(y = \overline{y}|x)$, the first step is to find a threshold s that splits X into two distributions $X_1 = \{x \in X | P_A(y = \overline{y}|x) \leq s\}$ and $X_2 = \{x \in X | P_A(y = \overline{y}|x) > s\}$, in which X_1 contains images with likely noisy and X_2 images with likely clean labels. To find s, we assume that X_1 and X_2 can be approximated by two gaussian distributions (μ_1, σ_1) and (μ_2, σ_2). The optimal threshold s maximizes the objective

$$Q(s) = \frac{n_1(s)\big(\mu_1(s) - \mu\big)^2 + n_2(s)\big(\mu_2(s) - \mu\big)^2}{n_1(s)\sigma_1(s)^2 + n_2(s)\sigma_2(s)^2}, \quad (6)$$

where $n_1(s)$ and $n_2(s)$ denote the cardinality of X_1 and X_2 depending on s, and μ is the mean probability $P_A(y = \overline{y}|x)$ of all samples in X. The optimal s minimizes the inter-class variance and can be found by calculating $Q(s)$ for all s with a reasonable step size $\Delta s = 0.001$.

Using Otsu's algorithm, we preserve a threshold s to split the data into noisy and clean samples, and furthermore thresholds μ_1 and μ_2 to subdivide the subsets into more fine-grained subsets. A finer distinction w.r.t. the likelihood of being noisy allows a more precise weighting of the samples in the following steps. Depending on the requirements of the application, the task of label noise detection can be solved by classifying a sample x by comparing $P_A(y = \overline{y}|x)$ with one of the thresholds s, μ_1, and μ_2. While using μ_1 is more liberal to classifying noisy labels into the clean dataset than s, μ_2 is more conservative. A visualization of a distribution of $P_A(y = \overline{y}|x)$ and the estimated noise is shown in Fig. 1.

4.3 Robust Optimization

After splitting up the dataset into potentially clean and corrupted data, we use simple robust training techniques to train the final classification model. Based on the idea of label smoothing [Szegedy *et al.*, 2016], we extend the CE loss (Eq. (2)) and combine the ground truth label with the student's prediction and a sample dependent α_x:

$$L_{\text{Robust}} = -\frac{1}{|X|} \sum_{(x,\overline{y}) \in X} \sum_{c \in C} \mathcal{S}(\beta_x^c) \log\big(P_T(y = c|x)\big)$$
$$\text{with } \beta_x^c = (1 - \alpha_x)\mathbb{1}[c = \overline{y}] + \alpha_x P_S(y = c|x) \quad (7)$$

Acc [%]	Aggre	Rand1	Rand2	Rand3	Worst
SOP	**95.61**	**95.28**	**95.31**	**95.39**	**93.24**
CORES	95.25	94.45	94.88	94.47	91.66
DivideMix	95.01	95.16	95.23	95.21	92.56
ELR+	94.83	94.43	94.20	94.34	91.09
PES	94.66	95.06	95.19	95.22	92.68
ELR	92.38	91.46	91.61	91.41	83.58
CAL	91.97	90.93	90.75	90.74	85.36
CE	87.77	85.02	86.14	85.16	77.69
Ours	93.68	92.50	92.63	92.54	86.64

Table 1: Classification accuracy of our method compared to standard CE-loss framework and state-of-the-art methods *SOP* [Liu *et al.*, 2022a], *CORES* [Cheng *et al.*, 2021], *DivideMix* [Li *et al.*, 2020], *PES* [Bai *et al.*, 2021], *ELR* [Liu *et al.*, 2020], and *CAL* [Zhu *et al.*, 2021].

and *Sharpening* $\mathcal{S}$ that is explained later.

While the teacher network is trained by L_{Robust}, the student network is still trained with the student loss L_{Stud} (Eq. (3)). As larger as the instance dependent α_x gets, the less the ground truth of a sample x is trusted. We adapt α_x for every sample individually, depending on the cluster membership after Otsu. We define four fixed alpha values with $\alpha^1 < \alpha^2 < \alpha^3 < \alpha^4$ where α^1 gets assigned to samples with $P_A(y = \overline{y}|x) \geq \mu_2$, α^2 to samples with $\mu_2 > P_A(y = \overline{y}|x) \geq s$, α^3 to samples with $s > P_A(y = \overline{y}|x) \geq \mu_1$, and α^4 otherwise.

Since a larger α_x enlarges the entropy in the objective, we use a modified *Sharpening* method

$$\mathcal{S}(\beta_x^c) = \frac{(\beta_x^c)^{1+\alpha_x}}{\sum_{i \in C} (\beta_x^i)^{1+\alpha_x}} \quad (8)$$

as used by [Li *et al.*, 2020] to minimize the entropy. The Sharpening function is applied stronger for insecure samples by reusing the above mentioned alpha.

5 Experiments

We perform several experiments to evaluate our proposed method. The experimental setup and the used metrics are explained first. Then we present evaluation metrics on the recently released dataset *CIFAR-10N* [Wei *et al.*, 2022] and show details and observations of our core method *Blind Knowledge Distillation*.

5.1 Experimental Setup

We evaluate our method on the noise levels provided in the CIFAR-10N dataset. To be comparable to other methods, we utilize the same model setup as used in [Li *et al.*, 2020]. We use a 18-layer PreAct ResNet [He *et al.*, 2016b] and *Stochastic Gradient Descent* with momentum of 0.9 and weight decay of 0.0005 as optimizer. The networks are trained for 300 epochs beginning with a learning rate of 0.02 and reduce it to 0.002 after 150 epochs. We train the network with randomly sampled batches of 128 image samples. In the first stage, the teacher network optimizes the standard CE loss (Eq. (2)) until

[%]		Aggre			Rand1			Rand2			Rand3			Worst		
		F_1	Pr	Re	F_1	Pr	Re	F_1	Pr	Re	F_1	Pr	Re	F_1	Pr	Re
P_T	μ_1	69.2	58.0	85.9	**82.9**	82.7	83.0	**83.4**	83.6	83.3	**83.1**	82.9	83.2	75.0	**95.9**	61.6
	s	46.6	30.7	97.1	67.2	51.3	97.5	67.9	52.1	97.6	67.3	51.3	97.6	85.9	79.6	93.3
	μ_2	28.8	16.8	99.7	44.8	28.8	99.8	45.9	29.8	99.8	45.1	29.1	99.8	70.8	55.0	99.6
P_S	μ_1	50.9	35.7	88.4	74.5	65.7	86.0	75.8	67.9	85.9	74.8	66.1	86.0	75.8	92.3	64.4
	s	31.9	19.0	98.3	55.0	38.2	98.2	56.6	39.7	98.1	55.4	38.6	98.5	81.9	72.2	94.4
	μ_2	22.3	12.5	**99.8**	39.8	24.8	**99.9**	41.3	26.0	**99.9**	40.4	25.3	**99.9**	68.9	52.6	**99.7**
P_A	μ_1	**71.2**	**63.7**	80.5	82.5	**84.1**	81.0	83.1	**85.4**	80.9	82.7	**84.9**	80.7	**77.8**	95.8	65.5
	s	55.7	39.6	94.1	73.7	60.1	95.2	75.2	62.1	95.3	74.6	61.3	95.1	87.3	83.4	91.5
	μ_2	37.0	22.8	98.9	54.2	37.3	99.4	55.4	38.4	99.2	54.9	37.9	99.4	75.7	61.3	99.0

Table 2: F_1-Score, Precision, and Recall on the label noise detection task with different probability sets (P_T, P_S, P_A) and different thresholds provided by Otsu's method (s, μ_1, μ_2) to split the dataset into clean and corrupted subsets. Best metrics are presented in bold.

the detection of the tipping point induces the start of the second stage, in which the teacher network optimizes the modified loss (Eq. (7)). The hyperparameters introduced by our method are set to $\alpha^1 = 0.3$, $\alpha^2 = 0.45$, $\alpha^3 = 0.55$, $\alpha^4 = 0.7$. For the noisy detection task, we use the probabilities P_A and the threshold μ_1 after Otsu. We repeated the experiments at least five times with random seeds and report the averaged metrics.

The method proposed in this paper is evaluated on *CIFAR-10N* [Wei *et al.*, 2022]. CIFAR-10N manually re-labelled the CIFAR-10 [Krizhevsky and Hinton, 2009] by multiple humans to investigate the impact of realistic label noise compared to synthetically induced ones. The dataset contains 50K training images and 10K test images with a size of 32×32. For the training set, there are five label sets with realistic human label noise with a ratio of approx. 9%, 17%, 18%, 18%, and 40% label noise. In the same order of the noise ratios, we denote them as *Aggre*, *Rand1*, *Rand2*, *Rand3*, and *Worst* in our experiments.

The tasks for the dataset are twofold: First, the classifier should be trained robust to achieve a high test accuracy even with high noisy rates and second, noisy labels in the training data should be detected and marked as noisy. The metrics to evaluate the tasks are given in the next section.

5.2 Metrics

We evaluate the performance of image classification with the commonly used *Accuracy* (Acc) metric. It measures the classification accuracy on the test dataset X_{Test} using the ratio of correct classified test samples compared to the dataset size:

$$\text{Acc} = \frac{\sum_{(x,\overline{y})\in X_{\text{Test}}} \mathbb{1}\left[\arg\max_{c\in C}\left(P(y=c|x)\right) = \overline{y}\right]}{|X_{\text{Test}}|} \tag{9}$$

The task of noisy label detection is evaluated with the well-known F_1-score, *Precision* (Pr), and *Recall* (Re) metrics, in which *Precision* decreases if clean labels are classified as noisy and *Recall* decreases if noisy labels are classified as clean. The F_1-score harmonizes both aspects. With the subsets of true ($X_{\text{Noise}} \subset X$) and predicted ($X'_{\text{Noise}} \subset X$) noisy labels from the training set X, the metrics are defined as:

$$\text{Pr} = \frac{\sum_{x\in X'_{\text{Noise}}} \mathbb{1}\left[x \in X_{\text{Noise}}\right]}{|X'_{\text{Noise}}|}, \tag{10}$$

$$\text{Re} = \frac{\sum_{x\in X_{\text{Noise}}} \mathbb{1}\left[x \in X'_{\text{Noise}}\right]}{|X_{\text{Noise}}|}, \tag{11}$$

$$\text{and} \quad F_1 = \frac{2}{\text{Pr}^{-1} + \text{Re}^{-1}}. \tag{12}$$

5.3 CIFAR-10N

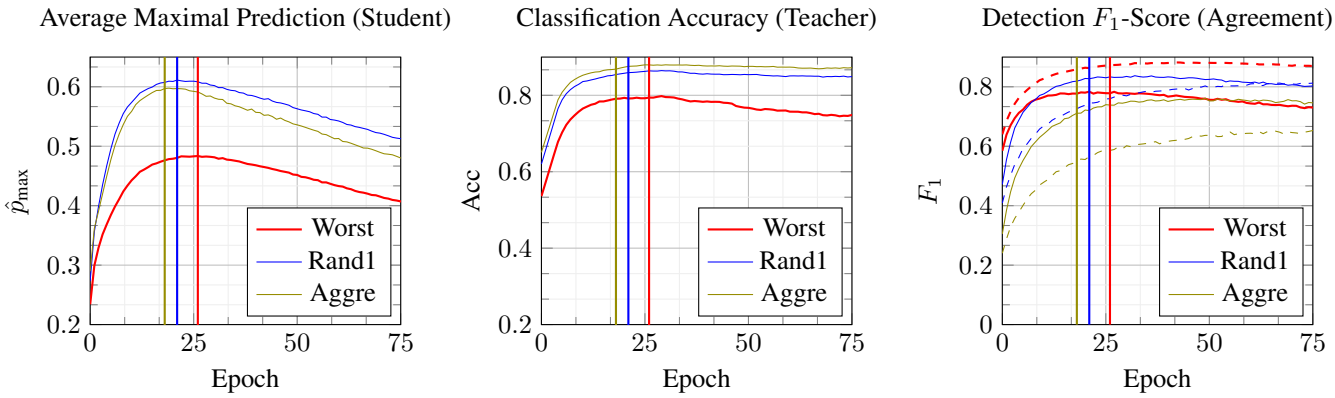
This section elaborates the results for the tasks of robust training and noise detection.

Robust Training We compare our results to the latest six state-of-the-art methods and the standard CE baseline on the CIFAR-10N Leaderboard in Tab. 1. Our method achieves the performance to be listed on the new sixth position of the leaderboard outperforming *CAL* and standard *ELR*. We want to mention that *ELR+* and *DivideMix* apply multiple models and high performance semi-supervised strategies such as *MixMatch* [Berthelot *et al.*, 2019].

Noise Detection The detection performance is shown in Tab. 2. We present F_1, precision, and recall for all five noise levels in CIFAR-10N (10 classes). The split to classify clean and corrupted labels is performed based on one of the probability sets P_T, P_S, and P_A and the three threshold s, μ_1, μ_2 provided by Otsu's method. Intuitively, the precision is higher for the lower threshold μ_1 and recall for the higher threshold μ_2, respectively. The experiments show that the harmonized metric F_1 performs best for μ_1. Thus, μ_1 is used to solve the task of Noise detection. The combined probability P_A performs better or on par w.r.t. the F_1-score, confirming the improved classification accuracy that is also visible in Fig. 3. An exemplary distribution of ground truth label probabilities for clean and noisy labels with the subsequent split based on Otsu is shown for the *Worst* dataset in Fig. 1.

5.4 Blind Knowledge Distillation

The core contribution of our method is *Blind Knowledge Distillation*. This section analyzes its ability to detect



(a) The average maximal probability of the students network prediction. Vertical lines denote the maxima during the training process which can be interpreted as tipping point at which the model start fitting to individual sample details.

(b) Test accuracy of the teacher network during training. Vertical lines denote the tipping points from Fig. 3a. With increasing noise rates, the detected tipping points fit approximately to the maxima of the classification test accuracy.

(c) Noise detection F_1-score on the training set with student maxima from Fig. 3a. Dashed lines indicate split by Otsu threshold s, while solid lines indicate the split by μ_1. The student maxima fit approximately to the maxima of the detection performance for μ_1 for higher noise rates.

Figure 3: Relation between the student's prediction behavior and the classification and noise detection performance. The maximum of the student network's average maximum probability indicates the start of fitting to sample details and thus can be used for early stopping to avoid overfitting. The models for this figure are trained without robust training for 75 epochs to show the standard training behavior.

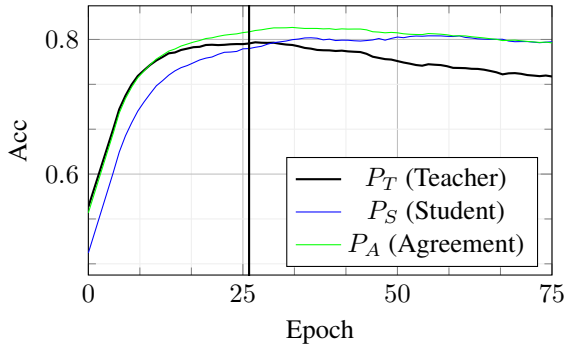


Figure 4: Test accuracy w.r.t. training epoch based on the predicted probabilities of the teacher, student, and the proposed combined agreement. The agreement probability combines the strengths and outperforms the teacher and student probability. Note that we train the framework for 75 epochs and without robust optimization (Sec. 4.3).

(over)fitting on sample details and how it can be used to improve the classification accuracy. For the experiments in this section, we trained our student-teacher framework without detection of the tipping point and robust optimization (Sec. 4.3) after generalization for 75 epochs to show the teacher's and student's learning behavior.

Fig. 3a shows the average maximal probability prediction of the student network over training time. The probability strongly increases in the first training epochs and degrades after a tipping point. It is notable that high noise rates decrease the absolute mean probability in general (see *Worst*). Our explanation of this behavior is that classifiers are not able to clearly predict a class based on simple and generalized but ambiguous image patterns. Thus, the classifier produces multiple predictions $P(y = c|x) \gg 0$ during the first generalization stage. An example pattern could be the coarse shape which is often ambiguous, e.g. for classes *dog* and *cat*. In the second training stage after the tipping point, the teacher network adapts detailed sample patterns to maximize $P(y = \overline{y}|x)$ which also minimizes $P(y \neq \overline{y}|x)$

The test classification accuracy of the teacher network is shown in Fig. 3b. While early stopping of standard optimization is not important for clean datasets or low noise levels (*Aggre*), fitting on sample details leads to overfitting and decreases the classification accuracy on high noise rates (*Worst*). Therefore, the choice of an early stopping epoch is highly important. It shows that the tipping point from Fig. 4 is a good indicator to detect overfitting. It proposes an accurate estimation to stop optimizing on high noise levels without stopping too early on low noise levels.

Using the tipping point in Fig. 3a to split the data into potentially corrupted and clean subsets is intuitively, due to the beginning overfitting and decreasing classification accuracy. Fig. 3c shows the detection ability with the F_1-score if splitting the dataset based on Otsu's algorithm and P_A at every epoch. Similar to Fig. 3b, the tipping point gives a guess for a suitable epoch to split the dataset. The estimation for high noise rates is sufficient, especially due to the decreasing F_1-score after the tipping point. Since low noise rates does not seem to affect the F_1 negatively in latter training stages, the tipping point estimate leads to a slightly too early splitting epoch.

An interesting insight about the teacher and student classification accuracy P_T and P_S is shown in Fig. 4 on a high

noise level. While the teachers accuracy decreases during overfitting, the students accuracy persists. We claim that using the complementary student loss from Eq. (3) prevents the student from fitting to misleading image details by removing the ground truth related logits $l_{c=\overline{y}}$. Also interesting is that the combined probability P_A suits as the overall best probability for classification. While the combined probability is quite similar to the teachers probability $P_A \approx P_T$ during the early stage, it converges to the students accuracy $P_A \approx P_S$ in the latter ones. Near the tipping point, it outperforms both.

Overall, *Blind Knowledge Distillation* is a better choice to automatically detect overfitting rather than to stop training after predefined and fixed periods (e.g. in [Li *et al.*, 2020]). Combining P_T and P_S to P_A can be used to improve the overall classification accuracy.

6 Conclusion

This paper introduces *Blind Knowledge Distillation* that is able to transfer simple and general image patterns that are not based on individual image details. We show that our framework is able to identify the tipping point from fitting to simple but general image patterns to fitting to image details and use it for early stopping in standard classification frameworks and furthermore to estimate the likelihood of samples in the training data of being clean or corrupted.

Our method performs on par with state-of-the-art methods that are not extended with high performance semi-supervised training strategies. Compared to them, we do not rely on manually predefined warm-up phases and adapt it online during training. However, the intention of this paper is to provide new insights about general learning behavior rather than to tune our method with known strategies. We hope that *Blind Knowledge Distillation* helps researchers to improve the handling of under- and overfitting.

Acknowledgments

This work was supported by the Federal Ministry of Education and Research (BMBF), Germany under the project LeibnizKILabor (grant no. 01DD20003) and the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy within the Cluster of Excellence PhoenixD (EXC 2122).

References

[Bai *et al.*, 2021] Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021.

[Berthelot *et al.*, 2019] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[Cheng *et al.*, 2021] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *International Conference on Learning Representations*, 2021.

[Gou *et al.*, 2021] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6), 2021.

[Han *et al.*, 2018] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[He *et al.*, 2016a] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[He *et al.*, 2016b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*. Springer, 2016.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Krizhevsky and Hinton, 2009] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[Li *et al.*, 2017] 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.

[Li *et al.*, 2020] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

[Liu and Guo, 2020] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning*. PMLR, 2020.

[Liu *et al.*, 2020] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020.

[Liu *et al.*, 2022a] Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. *arXiv preprint arXiv:2202.14026*, 2022.

[Liu *et al.*, 2022b] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng

Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[Otsu, 1979] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1), 1979.

[Rawat and Wang, 2017] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29(9), 2017.

[Song *et al.*, 2022] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[Wang *et al.*, 2019] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[Wei *et al.*, 2022] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022.

[Xiao *et al.*, 2015] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *20155 IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[Zhu *et al.*, 2021] Zhaowei Zhu, Tongliang Liu, and Yang Liu. A second-order approach to learning with instance-dependent label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.