Stochastic Label Refinery: Toward Better Target Label Distribution

Xi Fang
Shanghai Jiao Tong University
Shanghai 200240, China
Email: seefun@sjtu.edu.cn

Jiancheng Yang
Shanghai Jiao Tong University
Shanghai 200240, China
Email: jekyll4168@sjtu.edu.cn

Bingbing Ni *
Shanghai Jiao Tong University
Shanghai 200240, China
Email: nibingbing@sjtu.edu.cn

***Abstract*—This paper proposes a simple yet effective strategy for improving deep supervised learning, named Stochastic Label Refinery (SLR), by refining training labels to more informative labels. When training a neural network, target distributions (or ground-truth) are typically "hard", which means the target label of each category consists of only 0 and 1. However, the fixed "hard" target distributions do not capture association between categories or that between objects. In this study, instead of using the hard target distributions, we iteratively generate "soft" target label distributions for training the neural networks, which leads to better performances. The soft target distributions are obtained via an Expectation-Maximization (EM) iteration, where the "true" target distributions and the learned models are regarded as hidden variables. In E step, the models are optimized to approximate the target distributions on stochastic splits of training data; In M step, the target distributions are updated with predicted pseudo-label on leave-out splits. Extensive experiments on classification and ordinal regression tasks, empirically prove that the refined target distribution consistently leads to considerable performance improvements even applied on competitive baselines. Notably, in DeepDR 2020 Diabetic Retinopathy Grading (DeepDRiD) challenge, our method improves the quadratic weighted kappa on official validation set from 0.8247 to 0.8348 and achieves a state-of-the-art score on online test set. The proposed SLR technique is easy to implement and practically applicable.**

I. Introduction

How to improve the neural network is a common research problem in deep supervised learning. Most studies concentrate on data augmentation [1]–[4], model structure [5]–[7] or loss function [8], [9]. In this study, we aim at a simple yet effective strategy focusing on the label aspect. In deep supervised learning, usually the training procedure is to minimize cross-entropy loss between model output and target label [10], [11]. The quality of ground-truth label determines the upper limit of the model performance. The common target label distribution is one-hot or other forms of hard label. At a time when model performance is increasingly difficult to improve, we should start to rethink whether the hard label is the most appropriate target label distribution for deep learning. In several recent studies [12]–[18], we find that numerous works based on soft target label could achieve better performance than hard label. We can naturally think of a question: how to get even better target label distributions?

* corresponding author: Bingbing Ni.

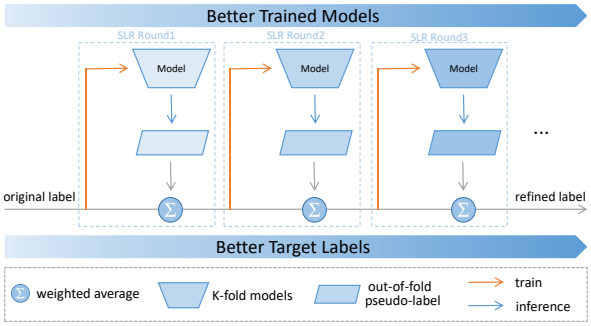


Fig. 1: Framework of Stochastic Label Refinery (SLR). By constantly using out-of-fold pseudo label to refine target label, it is able to get better refined labels and use them to train better models. This figure shows 3 rounds of SLR iterations, and the detail of each round SLR is shown in Fig. 2.

Traditional target label distribution has caused great limitations to the further improvement of model capability. Guo *et al.* [19] have shown that one-hot and other forms of hard label may bring risks of over-confidence to the modern neural network. In practical application faced with noise labeling or real difficult samples, the use of online hard example mining [20] and focal loss [8] increases this potential risk. Please refer to Appendix A for illustrations and empirical results on toy experiments. Fine-grained image classification problems are more and more common nowadays, and the hierarchy or grading relationship between many categories also makes the problem seems more difficult, especially when categories are not orthogonal. The typical hard supervised label has the following drawbacks: a) loss of association between categories (*i.e.*, similar categories should have similarly target confidence in different categories), b) no difference between different samples in the same category, and c) the vulnerability to label noise.

There have been pioneer works focusing on improving the target label distribution. To replace hard label and handle the over-confidence issue, label smoothing is introduced by Szegedy *et al.* [12] After recommended by Mu Li *et al.* [21], label smoothing has been a common trick in image classification problem and used in many state-of-the-art works [7], [21].

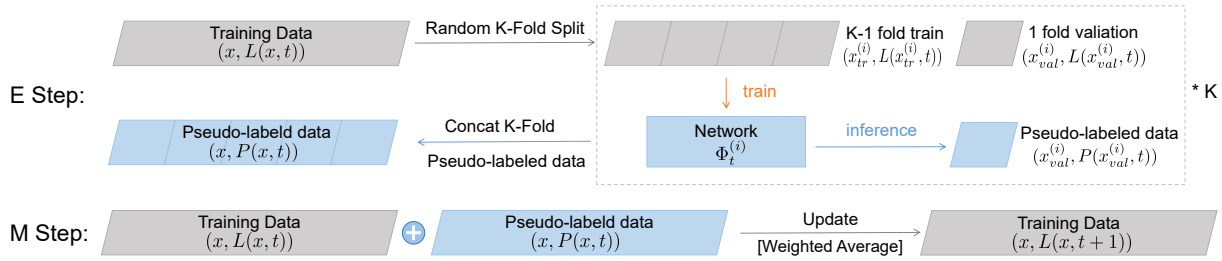


Fig. 2: The algorithm structure of Stochastic Label Refinery (SLR). For each round of refinery, it consists of two phases. E step is stochastic K-fold training and gets out-of-fold inference result as pseudo-labeled data which has label $P(x,t)$. M step is refining the label $L(x,t)$ in round t and gets the label $L(x,t+1)$ as the target distribution for the next round. The label refinery process is done by weighted average. Through the iteration of this EM algorithm, we continue to get better soft target label and better trained models.

Rafael *et al.* [13] also use many experiments to explore label smoothing on handling over-confidence and prove its ability. However, label smoothing lacks flexibility and cannot be used in regression tasks, and the smoothed label is just the soft form of one-hot labeling which may not be the best target label distribution. Some other works provide us with some ideas for generating soft label. Some are based on soft pseudo-label [22] in semi-supervised learning [23], [24]. Pseudo-label is generated by training models on labeled data and then inference on unlabeled data. The process of finding better pseudo-label (hard pseudo-label in common) through constant iteration is called self-training [23]. Xie *et al.* [25] find that using soft pseudo-label in self-training is able to obtain better score than hard pseudo-label in ImageNet [26] dataset in some experiment. Besides, label refinery proposed by Bagherinezhad *et al.* [16] , which is developed upon knowledge distillation introduced by Hinton *et al.* [27], only uses labeled data to generate the soft pseudo-label on itself and use it as the new target label.

Here we propose a simple algorithm named Stochastic Label Refinery (SLR) to generate better soft target label distribution, which makes the model learn better and leads to better performances. In our method, we randomly k-fold split the data and use k-fold training data to train k models and inference them in k-fold validation data respectively to get pseudo-label on all labeled dataset, which commonly known as out-of-fold (oof) pseudo-label, and then use oof pseudo-label to refine the target label distribution by weighted average. This refinery process can be repeated for several rounds. In each iteration, we optimize the KL divergence between network output and target distribution. Through the iteration of this EM algorithm, we get better models and better target label distribution. Figure 1 and 2 show the framework and principle of our algorithm.

In our studies we show significant gain using SLR in several types of image classification tasks. Our method is able to get improvement over strong baselines in some latest datasets which are more suitable for practical application. On CIFAR10 [28] image classification, we achieve 96.53% top-1 accuracy only using one SE-ResNet56 model from the baseline of 96.16% using SLR; On fine-grained categorization dataset (FGVC7) Plant Pathology 2020 [29], SLR improves accuracy from 0.9676 to 0.9747; for grading or ordinal regression problem [30], we use DeepDR Diabetic Retinopathy Grading (DeepDRiD) dataset [31], and SLR improves the quadratic weighted kappa [32] on validation set from 0.8247 to 0.8348 and achieves the state-of-the-art score on online test set. SLR significantly and consistently improve the model performance over competitive baselines, including strong backbones equipped with label smoothing or other methods based on soft label.

II. Stochastic Label Refinery

Figure 2 gives an overview of our Stochastic Label Refinery (SLR) algorithm. SLR could be regarded as an Expectation-Maximization (EM) algorithm, where models and target label distributions are optimized alternatively in an easy-to-implement procedure. As all EM algorithms, SLR is an iterative algorithm. For round t of SLR, the input training data is defined as x with the label $L(x,t)$.

In E step, firstly we shuffle the data and randomly split the training data into k splits (k folds). For each split (fold): take the split as leave-out for validation and take the remaining splits as training data, just the same setting as cross validation [33]. Suppose this is round t, for the i-th fold, the training data is $x_{tr}^{(i)}$ with the target label $L(x_{tr}^{(i)}, t)$. The validation data is $x_{val}^{(i)}$ with the target label $L(x_{val}^{(i)}, t)$. have the following formula to express the k-fold division:

$$x = x_{tr}^{(i)} \cup x_{val}^{(i)} \tag{1}$$

$$x_{tr}^{(i)} = \{x_{val}^{(j)} | j \in [1,k], j \neq i\} \tag{2}$$

After that we train model $\Phi_t^{(i)}$ on training data by optimizing the KL divergence between model output and round t target label $L(x_{tr}^{(i)}, t)$ on this training data. Then we inference the model on leave-out validation data split $x_{val}^{(i)}$ to obtain the pseudo-label $P(x_{val}^{(i)}, t)$. After concatenating of k splits of pseudo-labeled data, we get the out-of-fold pseudo-labeled data for the whole dataset with the label of $P(x,t)$.

$$P(x,t) = \{P(x_{val}^{(i)}, t), i \in [1,k]\} \\ = \{\Phi_t^{(i)}(x_{val}^{(i)}), i \in [1,k]\} \quad (3)$$

In M step, we should refine the target label $L(x,t)$ for round t to get new target label $L(x,t+1)$ for round $t+1$. We implement this label refinery process through the following formula, n is a hyper-parameter which controls the intensity of label refinery. By default, we take $n = 3$ in this study. When we use $n = t$, this formula is equivalent to the formula in stochastic weight average [34]. The difference is that we do not average the model weight, but the target label.

$$L(x,t+1) \leftarrow \frac{n * L(x,t) + P(x,t)}{n+1} \quad (4)$$

Finally, we iterate the process by putting back the refined labels as new target label and start the next EM process. Usually, after 2 to 6 iterations, the trained models can perform best with the refined labels. We will show in Section III that how the number of rounds affects the performance of the model.

The algorithm could be regarded as a variant of self-training [23], a method in semi-supervised learning. We adapt the core idea of self-training and conduct it on supervised learning. In each iteration of EM optimization, the k models which generate the pseudo-labels can be regarded as teachers, and models trained in next iteration using refined labels can be regarded as students. SLR is essentially an teacher-student structure, and we constantly put back student networks as teacher networks in the next iteration. We will discuss how our method is related to prior works like self-training and knowledge distillation in Section V. The key idea of our study is to constantly using stochastic splits of data to refine the labels of the rest of data as a regularization method. All the pseudo-labels are soft in our method. All of this is to achieve our goal: to avoid the risk of over-confidence brought by hard label and to obtain better target label distribution which leads to better model performance.

III. Experiments

We demonstrate the effectiveness of SLR in three different dataset. We first run several experiments on CIFAR10 to prove that our method is effective in traditional image classification dataset and prove that better target label is able to improve the model performance. After that, we selected two more datasets which are closer to the actual deep learning application to further test our algorithm and do ablation studies on them. Fine-grained categorized dataset FGVC7 Plant Pathology is used. SLR is also appropriate for regression problem, we use the DeepDRiD (Diabetic Retinopathy Grading) dataset to do more experiment and analyses by handle it as a regression problem.

Results on CIFAR10. On the traditional classification task dataset, we use CIFAR10 as the representative to carry on the experiment. Firstly, we build a baseline using SE-ResNet56 as model structure [6]. To further improve performance, we use AutoAugment (AA) [3], an method in AutoML [35], with one-cycle learning rate policy [36] to build a stronger baseline. Since the purpose of our experiment on CIFAR10 is only to test whether it is effective, and to pave the way for further experiments, we choose the simplest condition: only using 1 round SLR. The Table I shows the results. In all of the following experiments, the hyper parameters we choose in SLR is $n = 3$ and fold number $k = 5$ by default.

TABLE I: Performance comparison in CIFAR10. AA refers to AutoAugmentations. SLR for one round vanilla SLR. SLR-a: SLR with test-time augmentation in pseudo-label, SLR-e: SLR with model ensemble in pseudo-label, SLR-ae: SLR-a+SLR-e. All controlled trials share the same training protocal.

Method	Top-1 accuracy
VGG16 [37]	92.64%
ResNet101 [38]	93.75%
DenseNet121 [39]	95.04%
PreResNet56 [40]	95.51%
SE-ResNet56 [6]	95.87%
SE-ResNet56 [6] + AA [3]	96.16%
SE-ResNet56 [6] + AA [3] + Label Smoothing [12]	96.16%
SE-ResNet56 [6] + AA [3] + SLR	96.41%
SE-ResNet56 [6] + AA [3] + SLR-a	96.44%
SE-ResNet56 [6] + AA [3] + SLR-e	96.42%
SE-ResNet56 [6] + AA [3] + **SLR-ae**	**96.53%**

As shown in Table I, when compared with the strong baseline (SE-ResNet56 [6] with AutoAugment [3]), label smoothing [12] can not bring about a significant improvement. But after just one round of SLR, we improved the top-1 accuracy in CIFAR10 by 0.25%. This boost means that SLR is indeed a viable way to get better soft labels than label smoothing to further improve model performance.

To get better target label distribution, we can also use test time augmentation when generate pseudo-label, which means for each image, we use a variety of data augmentation method to generate several augmented images and get their pseudo-labels using teacher models, then average them to get the final pseudo-label to do label refinery. This method is abbreviated as SLR-a. Better pseudo-label can also be obtained by multi-model blending, which is abbreviated as SLR-e. Three different models are used to do average blending in our experiment of SLR-e. Combining these two methods, we get SLR-ae, which uses both test time augmentation and model blending. We can expect that pseudo-label generated by more ensemble learning methods can help us get better soft target label. As shown in Table I, all these three method are able to improve the top-1 accuracy in CIFAR10 from vanilla SLR. As expected, SLR-ae achieves the best model performance, indicates that better refined target label could bring better model performance.

The experiment on CIFAR10 shows that the soft label produced by SLR can improve the model performance. It also indicates that better target label distribution is able to bring better model performance. This proves our hypothesis and lays the foundation for our further experiments.

Results on FGVC7 Plant Pathology dataset. FGVC7 Plant Pathology dataset [29] is a fine-grained classified dataset. The aim of this dataset is to distinguish between leaves which are healthy or which disease they have. Since the images of different categories are somewhat similar, there is an association between the categories, which leads to the possibility of some noise in the annotation. For positive samples of the same category, the disease severity is also different. But the ground truth is hard label, which does not reflect the difference and relationship between different samples. Hard label is not able to reflect the severity of one disease. Our method will be very suitable for this kind of dataset. SLR could use soft label to indicate the severity of the disease and the association of different diseases. Getting better target label distribution will also improve the performance of model trained by refined label. We conduct a controlled experiment and compared several different softening label method like label smoothing [12], label refinery [16] and knowledge distillation [27] with our SLR. We have trained a good baseline with EfficientNet-B5 [7] backbone, using strong data augmentations and cosine annealing learning rate schedule with warmup [21]. All experiments are based on this strong baseline. The result is shown in Table II. We report mean $\pm$ std over 5 runs using EfficientNet-B5 [7] model. All controlled trials share the same environment, training tricks and hyper parameter settings. Both vanilla label refinery [16] and our SLR have carried out two rounds of label refinement to get decent scores.

TABLE II: Performance comparison in Plant Pathology dataset. We report mean $\pm$ std over 5 runs using EfficientNet-B5 model. All controlled trials share the same training protocol based on the first row baseline.

Method	Top-1 Accuracy	Average AUC
Baseline	0.9676±0.0056	0.974±0.000
Focal Loss [8]	0.9665±0.0064	0.968±0.000
OHEM [20]	0.9670±0.0071	0.974±0.000
Label Smoothing [12]	0.9736±0.0106	0.974±0.000
Knowledge Distillation [27]	0.9731±0.0090	0.973±0.000
Label Refinery [16]	0.9720±0.0079	0.961±0.000
Stochastic Label Refinery	**0.9747±0.0083**	**0.976±0.000**

We can observe that methods such as online hard example mining (OHEM) [20] and focal loss [8] which are sensitive to hard examples have worse performance while methods based on modified soft target label distribution perform better. Among them, our algorithm has the best performance. By observing the soft label generated by SLR algorithm, we can find that samples with different severity of the same disease have different tag values, and samples with multiple diseases also have multiple peak responses in these categories. This reflects the significant advantages of soft label and our SLR algorithm.

Results on DeepDRiD dataset. In Table III, we present our results with SLR on DeepDR Diabetic Retinopathy Grading (DeepDRiD) dataset [31]. DeepDRiD is one of the largest databases of DR patient population, and provide more than 1,000 patients data [31]. The task of this dataset is to grade the diabetic fundus image into five levels, which can be seen as a ordinal regression problem. Test metric is quadratic weigted kappa [32] in DeepDRiD. This dataset was accompanied by a ISBI 2020 grand challenge.

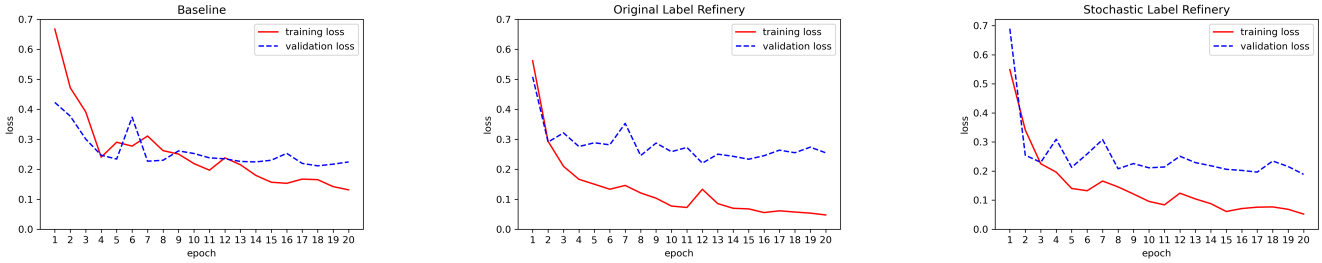
We treat this task as a regression problem and use thresholds to divide the prediction into five levels. All the ablation studies are done based on a high baseline with multiple training tricks like cosine annealing learning rate with warmup [21], strong data augmentation and ImageNet pretrained model of EfficientNet-B5 [7] with general mean pooling [41]. In the experiment, we use AdamW [42] with learning rate of 1e-3 and weight decay of 2e-4 and train for 20 epochs. Five experiments have been done for each method and we show the mean and standard deviation (mean $\pm$ std) of validation quadratic weighted kappa in the table (higher is better). It should be noted that in order to ensure the rigor of the experiment, we only adjust the hyper-parameters in the training set and do not use the validation set. All controlled trials except the first row share the same environment, training tricks and hyper parameter settings based on baseline (w/ tricks). Both vanilla label refinery method and our SLR have run for 4 rounds. Label smoothing does not appear on the list because it can not be trivially adapted for ordinal regression.

TABLE III: Performance comparison in DeepDRiD dataset. Higher is better. We report mean $\pm$ std over 5 runs using EfficientNet-B5 model. All controlled trials except the first row share the same training protocal with several training tricks.

Method	Quadratic Weighted Kappa
Baseline (w/o tricks)	0.8036±0.0214
Baseline (w/ tricks)	0.8247±0.0125
SWA [34]	0.8119±0.0234
OHEM [20]	0.8061±0.0174
Knowledge Distillation [27]	0.8128±0.0100
Label Refinery [16]	0.7527±0.0152
Stochastic Label Refinery	**0.8348±0.0053**

In the Table III, several methods based on changing target distribution (soft label) can not get points under high baseline, whereas our method is able to do that, and even more stable (with smaller variance on score). We also show the impact of SLR optimization rounds n and the loss curve between several different methods in Figure 4 and Figure 3. Our SLR is able to get points from a high baseline. With the increase of refinery rounds, the performance of SLR is gradually improved and reaches the maximum in 3-4 rounds. However, the vanilla label refinement method tends to over-fitting.

From the experiments on the DeepDRiD dataset, using SLR not only make the convergence faster and better, but also make the loss curve of the training set more consistent with loss curve on validation set. Our strong baseline has made the values of train loss and validation loss close to each other, but lost the synchronization of loss changes, which indicates that is not the best generalization. Since only the label of the training set is refined, the training loss becomes

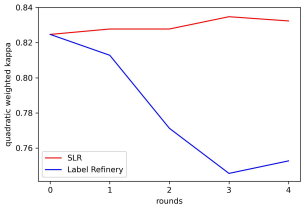


(a) baseline (b) vanilla label refinery (c) stochastic label refinery

Fig. 3: Training and validation loss curve in DeepDRiD dataset. Our Stochastic Label Refinery method has the best performance (smallest validation loss) at validation set. As the label becomes softer, both SLR and label refinery can make the training loss smaller. However, compared with vanilla label refinery, training loss and validation loss of SLR show the better consistency.

Method	Public Test	Private Test
Ours	**0.9303**	**0.9215**
Team1	0.9262	0.9211
Team2	0.9232	0.9097
Team3	0.9202	0.8946
Team4	0.9088	0.8890

(a)



(b)

Fig. 4: Performance of our method. Table (a) shows the leader board of ISBI 2020 DeepDR Diabetic Retinopathy Grading (DeepDRiD) challenge. Using the SLR, our solution get the top-1 score in the leaderboard. Figure (b) denotes the influence of different optimization rounds on SLR. On DeepDRiD validation set, SLR get the highest score of 0.8348 in the third round. Compared with label refinery, SLR can continue to improve the model capability at a high baseline.

smaller in both label refinery and SLR. vanilla label refinery's performance is even worse in model generalization (getting bigger validation loss). From the numerical value of validation loss and synchronicity with the training loss, the experimental results show that SLR does achieve a good regularization effect compared with vanilla label refinery. The value gap between training and validation loss is due to the noisy label. SLR method refine the target distribution in training which reduce the label noise, while the original noisy ground truth is still used in validation. So this gap contains the noise correction, which means that in many case of validation sets, our models trained by SLR make the prediction even better than ground-truth label. Through observation, we also verify it by noting that some of the error annotations have been corrected to a less erroneous soft label. As shown in Table 4 (a), by blending three models, we get 0.9303 and 0.9215 quadratic weighted kappa in DeepDRiD testset (public test and private test), which is the state-of-the-art score in this dataset.

IV. Discussion

Prior arts have focused on a small number of academic datasets. When the dataset is closer to the real application scenario or the baseline score is high enough, the traditional label refinery or knowledge distillation method tends to over-fitting, which makes it difficult to improve the score or even decrease the score as shown in Table III and Figure 4.

From the experiments above, we can observe that our Stochastic Label Refinery (SLR) can continue to improve the model performance under high baseline, and make the convergence more stable and faster. Compared with other target distribution regulation strategies such as label smoothing, there is a more obvious improvement. We will analyze the principle of SLR in detail below:

SLR is a regularization strategy. The loss changes of the training set validation set are more synchronized, which means better generalization performance. Even if the dataset is labeled carefully and correctly, SLR can also make the improvement by working as a regulation strategy. SLR does not lead to the over-fitting when softening the target label distribution which is common in knowledge distillation [27] and vanilla label refinery [16]. SLR is more like a model-based label smoothing [12] method, which let the label of each part of data be refined or regularized by the model trained using other data from the same dataset.

SLR also works as a label correction method. In some aspects, SLR makes the mislabeling less serious. SLR uses out-of-fold inference result to refine or correct the wrong target label annotations. In the last two experiments, we observe that the training proceeds of SLR are usually more stable and lead to faster convergence because of robustness to noise labels. While the performance of focal loss [8] and online hard example mining [20] are even worse than baseline method. We use some toy experiments in appendix A to further prove this hypothesis.

V. Related Works

In Section I, we mentioned how SLR was motivated by previous works. Here we categorize the most relevant works, as well as their similarities and differences with our SLR.

Label Smoothing. Softening labels has been used to improve model generalization in label smoothing [12]. Label smoothing uniformly redistributes 10% of the weight from

the hard ground-truth label to other classes to help regularize during training. Rafael *et al.* [13] point that label smoothing loses the correlation between categories, which is not conducive to further knowledge distillation of the model. On the opposite, our method makes full use of the knowledge of the training data, and mining the correlation between categories and samples to get better soft label and better trained models. Such soft label is more conducive to knowledge distillation and is able to be improved gradually through EM iteration. Our experiments also prove that SLR is a better regularization strategy added to target label distribution to reduce the risk of over-fitting and over-confidence.

Knowledge Distillation. The main use case of knowledge distillation [15], [27] is model compression by making the student model smaller. SLR is also can be regarded as a self-ensemble or self-distillation method. SLR uses the results of multiple teachers of the same model to implement a teacher-student structure which is similar to knowledge distillation. The main difference between our method and knowledge distillation is that knowledge distillation does not aim to improve the ability of teacher model. Our method uses the same model as teacher and student. Our aim is to get better target label, at the same time to make model trained better.

Self-training. Self-training [23], [25] first uses labeled data to train a good teacher model, then use the teacher model to label unlabeled data and finally joint labeled and unlabeled data to train a final student model. SLR has no need of unlabelled data. It takes part of the data as unlabeled samples and generate pseudo-label on it. The correlation between categories is established by using soft pseudo-label, which helps the model to find a more globally appropriate convergence direction in mini-batch training. While looking for a more meaningful and optimal target label distribution for machine learning tasks, we can constantly train better and better models using refined target distribution.

Label refinery. Label refinery [16] is a special form of knowledge distillation [27] applied in label aspect. SLR and label refinery use a similar iterable approach which alternately optimizes model and target label distribution. The difference is that label refinery is not a regularization strategy, but increasing the degree of over-fitting on the training set or validation set, which makes it difficult to achieve better results upon a strong baseline. Our approach, however, is a regularization strategy that does more than just fix noise labels, and is able to improve model performance from a very high baseline.

VI. Conclusion

In our paper, we point out the shortcomings of hard label distribution: (a) the risk of over-confidence, (b) easily affected by noise annotation and (c) lost intra-class and inter-class association; We propose a new regularized strategy of generating soft target label distribution: Stochastic Label Refinery (SLR). SLR is a method to progressive softens the label, establishes the association between the sample or categories, and finds better target label distribution which is more suitable for the deep learning task. At the same time, the better target label can be used to train better models. Extensive experiments prove that SLR is able to improve model performance upon high baseline without additional unlabeled data. Compared with the previous works based on changing target label distribution, SLR has higher performance and better generalization ability, and it is also proved that SLR is robust to label noise. We believe that SLR has great application potential in future deep supervised learning tasks.

Appendix A
Toy Experiments

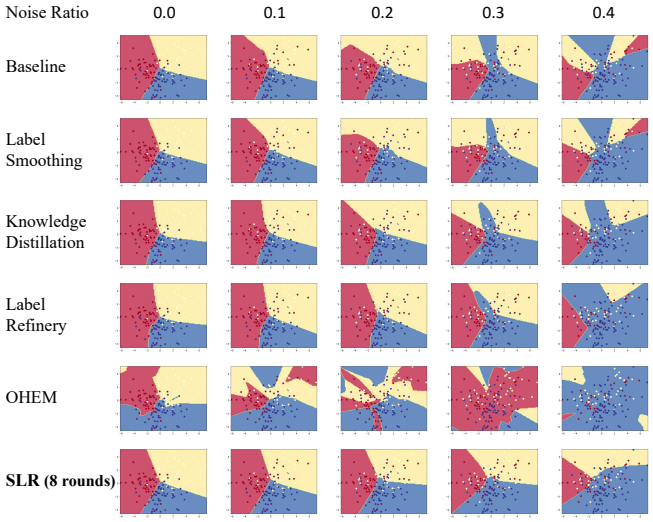


Fig. 5: A toy classification experiment. In a three categories classification problem, we inject different proportions of noise to ground truth labels by randomly changing the class label and test the robustness of different methods to noise. Different rows represent different methods, and different columns represent different label noise levels. Our Stochastic Label Refinery method shows strong noise robustness after 8 rounds of SLR.

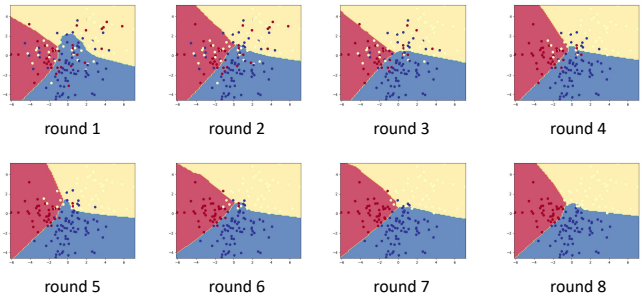


Fig. 6: The refinery process of Stochastic Label Refinery (SLR) algorithm. In a three categories classification problem, we inject 30% of noise to ground truth labels and use the noisy labels to train models using SLR. With the iteration of SLR, the wrong label is being corrected.

ACKNOWLEDGMENT

This work was supported by National Science Foundation of China (U20B200011, 61976137).

REFERENCES

[1] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[2] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017.

[3] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 113–123.

[4] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical data augmentation with no separate search," *arXiv preprint arXiv:1909.13719*, 2019.

[5] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[6] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[7] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[9] M. Berman, A. Rannen Triki, and M. B. Blaschko, "The lovász-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4413–4421.

[10] E. B. Baum and F. Wilczek, "Supervised learning of probability distributions by neural networks," in *Neural information processing systems*, 1988, pp. 52–61.

[11] E. Levin and M. Fleisher, "Accelerated learning in layered neural networks," *Complex systems*, vol. 2, pp. 625–640, 1988.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[13] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" in *Advances in Neural Information Processing Systems*, 2019, pp. 4696–4705.

[14] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," *arXiv preprint arXiv:1701.06548*, 2017.

[15] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers." in *Interspeech*, 2017, pp. 3697–3701.

[16] H. Bagherinezhad, M. Horton, M. Rastegari, and A. Farhadi, "Label refinery: Improving imagenet classification through label progression," *arXiv preprint arXiv:1805.02641*, 2018.

[17] H. Pham, Q. Xie, Z. Dai, and Q. V. Le, "Meta pseudo labels," *arXiv preprint arXiv:2003.10580*, 2020.

[18] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4320–4328.

[19] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1321–1330.

[20] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 761–769.

[21] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 558–567.

[22] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, 2013, p. 2.

[23] H. Scudder, "Probability of error of some adaptive pattern-recognition machines," *IEEE Transactions on Information Theory*, vol. 11, no. 3, pp. 363–371, 1965.

[24] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *33rd annual meeting of the association for computational linguistics*, 1995, pp. 189–196.

[25] Q. Xie, E. Hovy, M.-T. Luong, and Q. V. Le, "Self-training with noisy student improves imagenet classification," *arXiv preprint arXiv:1911.04252*, 2019.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.

[27] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[28] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[29] K. C. Dataset, "Plant pathology 2020 fgvc7," https://www.kaggle.com/c/plant-pathology-2020-fgvc7/data, 2020.

[30] C. Winship and R. D. Mare, "Regression models with ordinal variables," *American sociological review*, pp. 512–525, 1984.

[31] D. C. Dataset, "The 2nd diabetic retinopathy – grading and image quality estimation challenge," https://isbi.deepdr.org/data.html, 2020.

[32] A. Ben-David, "Comparison of classification accuracy using cohen's weighted kappa," *Expert Systems with Applications*, vol. 34, no. 2, pp. 825–832, 2008.

[33] J. Shao, "Linear model selection by cross-validation," *Journal of the American statistical Association*, vol. 88, no. 422, pp. 486–494, 1993.

[34] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," *arXiv preprint arXiv:1803.05407*, 2018.

[35] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Auto-weka: Combined selection and hyperparameter optimization of classification algorithms," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 847–855.

[36] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006. International Society for Optics and Photonics, 2019, p. 1100612.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[41] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[42] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," 2018.