要点

- pseudo-labeling
- 高学习率下网络不会记忆噪声
- 模型做法：step1: 先高学习率 fit easy pattern，再修改标签 step2: 不再更新标签，低学习率学习step1修改后的标签

方法

模型架构图：

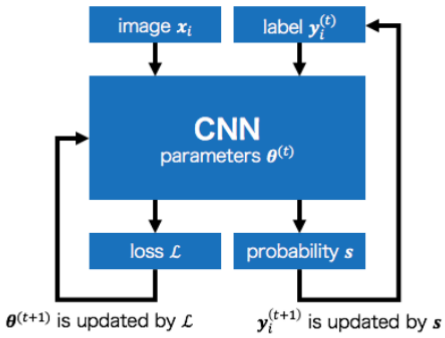

Figure 1. The concept of our joint optimization framework. Noisy labels are reassigned to the probability output by CNNs. Network parameters and labels are alternatively updated for each epoch.

Algorithm 1 Alternating Optimization

for $t \leftarrow 1$ to *num_epochs* **do**
  update $\boldsymbol{\theta}^{(t+1)}$ by SGD on $\mathcal{L}(\boldsymbol{\theta}^{(t)}, Y^{(t)}|X)$
  update $Y^{(t+1)}$ by Eq. (8) (hard-label)
    or Eq. (9) (soft-label)
end for

更新方式

算法针对Y和Θ分别对模型进行优化：

$$\mathcal{L}(\boldsymbol{\theta}, Y|X) = \mathcal{L}_c(\boldsymbol{\theta}, Y|X) + \alpha\mathcal{L}_p(\boldsymbol{\theta}|X) + \beta\mathcal{L}_e(\boldsymbol{\theta}|X), \quad (5)$$

- 对于网络，用SGD进行更新
- 对标签运用KL散度更新，简单来说就是将标签赋值为预测的向量(软标签直接相等，硬标签进行独热近似)

$$\mathcal{L}_c(\boldsymbol{\theta}, Y|X) = \frac{1}{n}\sum_{i=1}^{n} D_{KL}(\boldsymbol{y}_i||\boldsymbol{s}(\boldsymbol{\theta}, \boldsymbol{x}_i)), \qquad (6)$$

$$D_{KL}(\boldsymbol{y}_i||\boldsymbol{s}(\boldsymbol{\theta}, \boldsymbol{x}_i)) = \sum_{j=1}^{c} y_{ij} \log\left(\frac{y_{ij}}{s_j}(\boldsymbol{\theta}, \boldsymbol{x}_i)\right). \qquad (7)$$

正则化

$L_p: $

用来避免网络变成：全部标签为一个标签的状态。

损失函数设计：先引入一个先验Y概率分布，利用KL散度使一个batch下预测的分布与该先验分布尽可能相等

$$\mathcal{L}_p = \sum_{j=1}^{c} p_j \log \frac{p_j}{\bar{s}_j(\boldsymbol{\theta}, X)}$$

$$\bar{\boldsymbol{s}}(\boldsymbol{\theta}, X) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}(\boldsymbol{\theta}, \boldsymbol{x}_i) \approx \frac{1}{|\mathcal{B}|}\sum_{\boldsymbol{x}\in\mathcal{B}} \boldsymbol{s}(\boldsymbol{\theta}, \boldsymbol{x}) \qquad (11)$$

$L_e: $

用来避免soft label “弥散”成比如[0.2,0.2,0.2,0.2,0.2]的均匀模式

$$\mathcal{L}_e = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} s_j(\boldsymbol{\theta}, \boldsymbol{x}_i) \log s_j(\boldsymbol{\theta}, \boldsymbol{x}_i).$$

这是一个entropy term, 本身越混乱损失函数越大， 因此标签值倾向于收敛成一个独热码

实验

对“高学习率抑制噪声学习”结论的验证

实验条件：对称噪声、CE loss

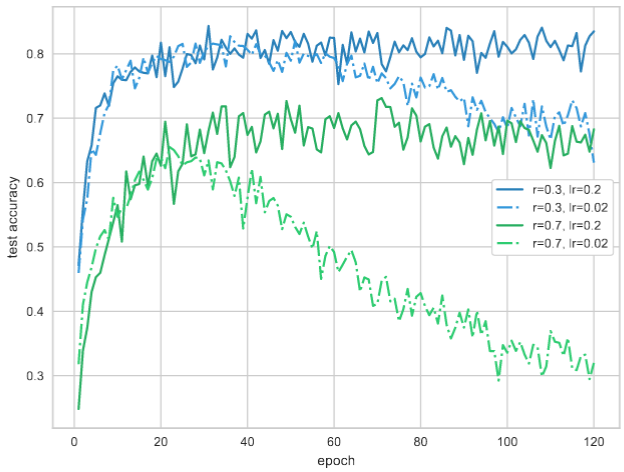

Figure 2. The test accuracy curve with different learning rates. The test accuracy gradually decreases when the learning rate is low (lr=0.02). Conversely, the test accuracy remains high at the end of training when the learning rate is high (lr=0.2).

不同于低学习率，高学习率的网络不会记忆噪声

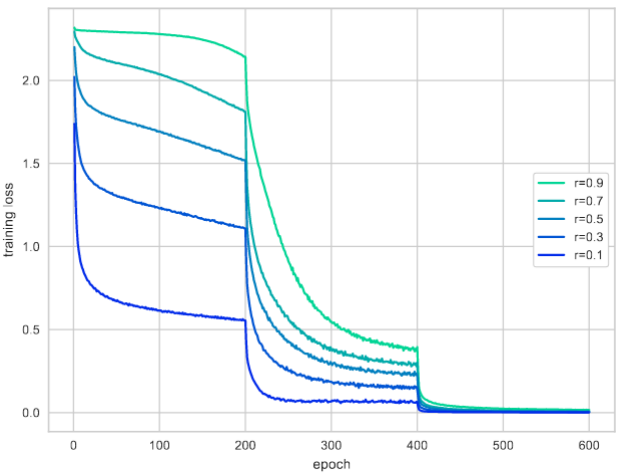

Figure 3. The training loss curve with different noise rates. At the end of training with a low learning rate, the value of the training loss is close to 0 even if the error rate is 0.9. In contrast, in the early phase of training with a high learning rate, an increase in the noise rate increases the training loss.

0~200： 学习率0.2

200~400：学习率0.02

400~600：学习率0.002

高学习率下，训练损失随着噪声率的升高而显著升高；*因此在高学习率下，通过优化标签来实现降低训练损失是可能的。*

软硬标签对标签复原率的贡献

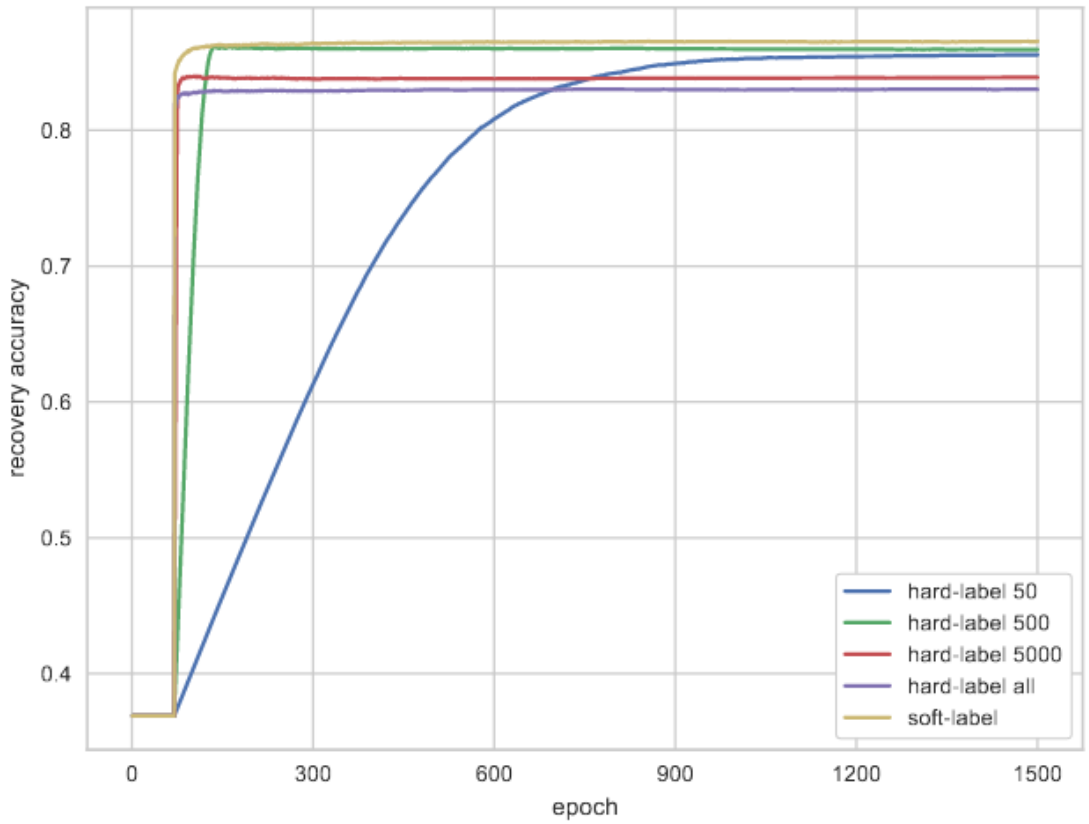

Figure 4. Comparison between the soft-label and the hard-label methods, showing the recovery accuracy. The soft-label method achieves faster convergence than the hard-label methods, and performs the best recovery accuracy.

recover accuracy表示复原的标签中正确复原的标签所占的比重

hard-label 50 表示每次update更新前五十个与预测标签最不同的真实标签

结论：soft标签效果最好收敛效果也最快，因为标签向量不仅包含了类别还包含了模型对该标签的置信度

t% t%e

Table 1. Test and recovery accuracy of different baselines on the CIFAR-10 dataset with symmetric noise. We report the average score of 5 trials.

#	method		Test Accuracy (%)						Recovery Accuracy (%)					
	noise rate (%)		0	10	30	50	70	90	0	10	30	50	70	90
1	Cross Entropy Loss	*best*	93.5	91.0	88.4	85.0	78.4	41.1	**100.0**	96.4	92.7	88.2	80.1	41.4
		last	93.4	87.0	72.2	55.3	36.6	20.4	**100.0**	91.1	74.6	57.6	39.6	21.7
2	Our Method	*best*	93.4	92.7	91.4	89.6	85.9	58.0	**100.0**	97.9	**95.1**	91.7	86.3	58.2
		last	**93.6**	**92.9**	**91.5**	**89.8**	**86.0**	**58.3**	99.9	**98.1**	**95.1**	**91.8**	**86.4**	**58.3**

Table 2. Test and recovery accuracy of different baselines on the CIFAR-10 dataset with asymmetric noise. We report the average score of 5 trials. #2, #3 are the results by our implementation.

#	method		Test Accuracy (%)					Recovery Accuracy (%)				
	noise rate (%)		10	20	30	40	50	10	20	30	40	50
1	Cross Entropy Loss	*best*	91.8	90.8	90.0	87.1	77.3	97.2	95.8	94.3	91.0	80.5
		last	89.8	85.4	81.0	75.7	70.5	95.0	90.2	85.3	80.2	75.2
2	Forward [14]	*best*	92.4	91.4	91.0	90.3	83.8	97.7	96.7	95.9	94.7	88.0
		last	91.7	89.7	88.0	86.4	80.9	97.9	95.8	93.6	91.5	85.5
3	CNN-CRF [19]	*best*	92.0	91.5	90.7	89.5	84.0	97.4	96.5	95.3	93.7	88.1
		last	90.3	86.6	83.6	79.7	76.4	95.1	90.5	86.4	82.1	78.7
4	Our Method	*best*	**93.2**	92.7	**92.4**	91.5	84.6	**98.3**	**97.2**	**96.3**	**95.2**	**88.3**
		last	**93.2**	**92.8**	**92.4**	**91.7**	**84.7**	98.1	97.1	**96.3**	**95.2**	88.1

除了our method的其他方法的recovery accuracy表示在训练集上预测的准确率

证明update标签的策略的有效性

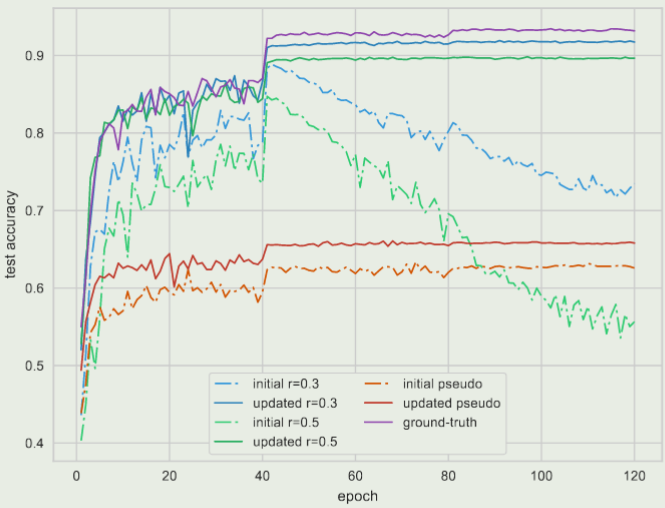

Figure 5. The test accuracy curve with different labels. The test accuracy on the pseudo labels is lower than that on the symmetric noise labels even if the number of inaccurate labels is lower. This trend remains if the labels are updated.

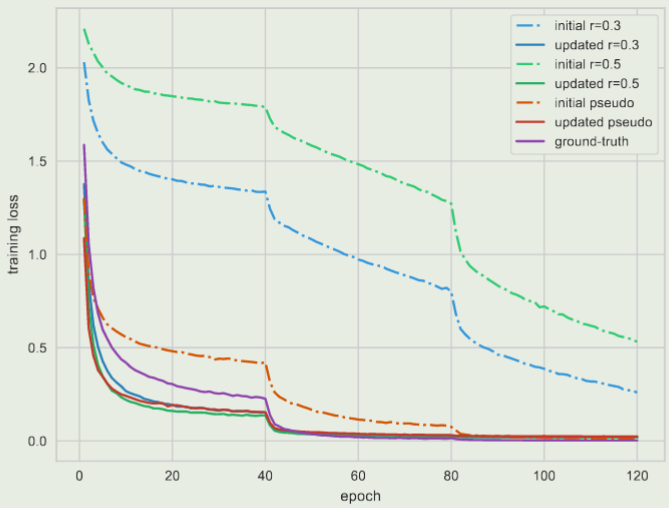

Figure 6. The training loss curve with different labels. The training loss on the pseudo labels is lower than that on the symmetric noise labels even if the number of inaccurate labels is higher. Each of the training losses on the labels updated from different noisy labels follows the training loss on the ground-truth labels. This implies that the updated labels are completely optimized for the network.

注:

可以看出来:

1.在update后的标签上训练时不会产生memorization effect的，说明真正的减少了很多wrong labels。

2.pseudo的training loss比symmetric的要小，但是测试效果确差。因为pseudo是经过网络优化的标签，但是并没有优化恰当(改变了标签分布)。

3.对噪声标签恢复的标签进行训练的测试精度比在真实标签上训练的测试精度要差，这表明优化后的标签不一定表示最优标签。