要点

- sample selection + learn how to teach + cross update.
- 运用$R(T)$来实现early stopping的软变种
- 运用两个网络cross update来解决confirmation bias 的问题

方法

Algorithm 1 Co-teaching Algorithm.

1: **Input** w_f and w_g, learning rate η, fixed τ, epoch T_k and $T_{\max}$, iteration $N_{\max}$;
for $T = 1, 2, \ldots, T_{\max}$ **do**
  2: **Shuffle** training set $\mathcal{D}$; //noisy dataset
  for $N = 1, \ldots, N_{\max}$ **do**
    3: **Fetch** mini-batch $\bar{\mathcal{D}}$ from $\mathcal{D}$;
    4: **Obtain** $\bar{\mathcal{D}}_f = \arg\min_{\mathcal{D}':|\mathcal{D}'|\geq R(T)|\bar{\mathcal{D}}|} \ell(f, \mathcal{D}')$; //sample $R(T)\%$ small-loss instances
    5: **Obtain** $\bar{\mathcal{D}}_g = \arg\min_{\mathcal{D}':|\mathcal{D}'|\geq R(T)|\bar{\mathcal{D}}|} \ell(g, \mathcal{D}')$; //sample $R(T)\%$ small-loss instances
    6: **Update** $w_f = w_f - \eta\nabla\ell(f, \bar{\mathcal{D}}_g)$; //update w_f by $\bar{\mathcal{D}}_g$;
    7: **Update** $w_g = w_g - \eta\nabla\ell(g, \bar{\mathcal{D}}_f)$; //update w_g by $\bar{\mathcal{D}}_f$;
  end
  8: **Update** $R(T) = 1 - \min\left\{\frac{T}{T_k}\tau, \tau\right\}$;
end
9: **Output** w_f **and** w_g.

R(T)

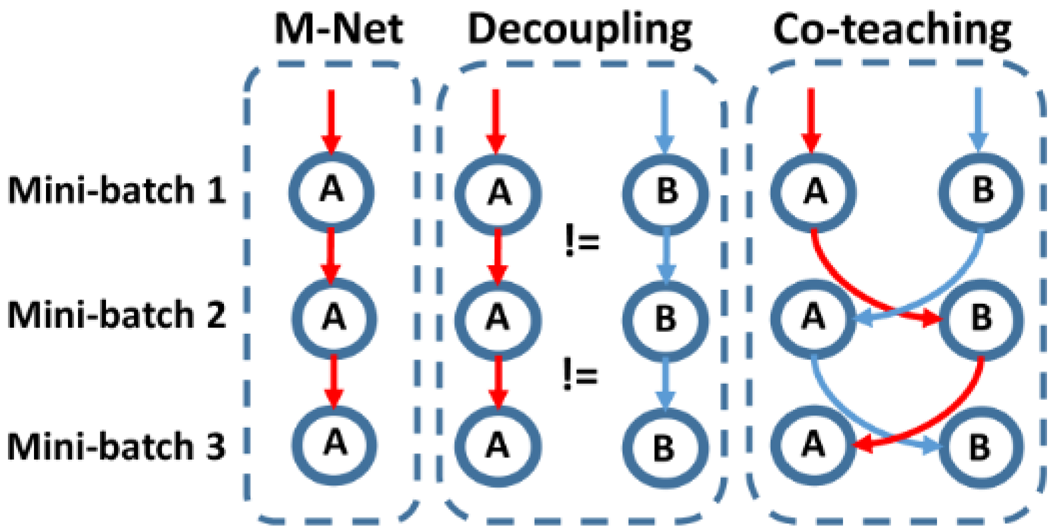
因为memorization effect, 网络会先记忆easy pattern后记忆noisy label。所以在早期小loss样本有更大可能是clean label。故R(T)随时间递减

cross update

cross update 的直观解释：

by the peer network if the selected instances are not fully clean. Take "peer-review" as a supportive example. When students check their own exam papers, it is hard for them to find any error or bug because they have some personal bias for the answers. Luckily, they can ask peer classmates to review their papers. Then, it becomes much easier for them to find their potential faults. To sum

可以看到error flow从两个网络之间流动，这样两个网络可以充分交流自己所学的信息，以减少confirmation bias



实验

t%10e

Table 4: Average test accuracy on *MNIST* over the last ten epochs.

Flipping-Rate	Standard	Bootstrap	S-model	F-correction	Decoupling	MentorNet	Co-teaching
Pair-45%	56.52% ±0.55%	57.23% ±0.73%	56.88% ±0.32%	0.24% ±0.03%	58.03% ±0.07%	80.88% ±4.45%	**87.63%** ±0.21%
Symmetry-50%	66.05% ±0.61%	67.55% ±0.53%	62.29% ±0.46%	79.61% ±1.96%	81.15% ±0.03%	90.05% ±0.30%	**91.32%** ±0.06%
Symmetry-20%	94.05% ±0.16%	94.40% ±0.26%	98.31% ±0.11%	**98.80%** ±0.12%	95.70% ±0.02%	96.70% ±0.22%	97.25% ±0.03%

Table 5: Average test accuracy on *CIFAR-10* over the last ten epochs.

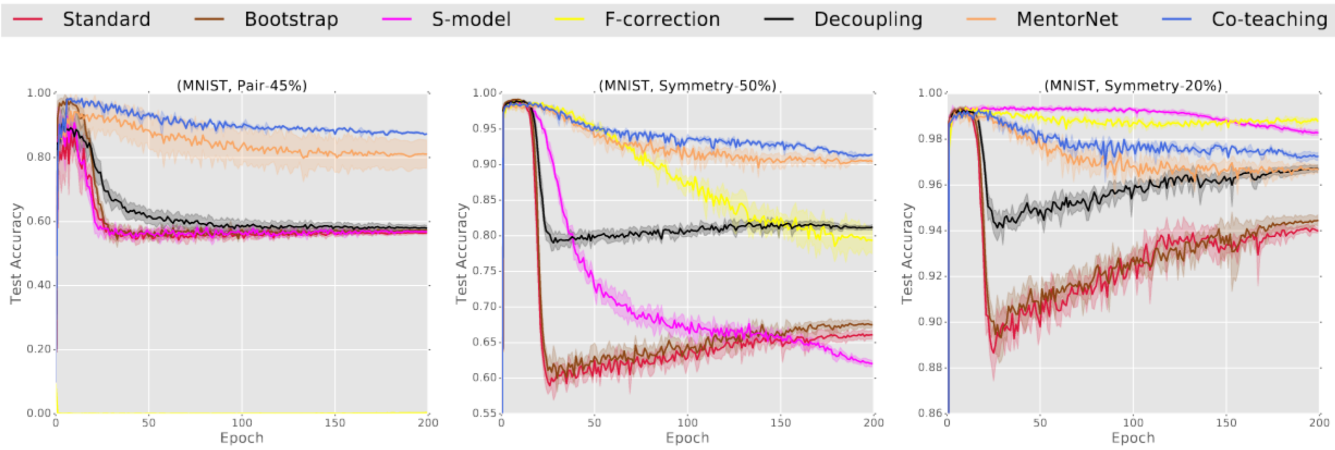
Flipping,Rate	Standard	Bootstrap	S-model	F-correction	Decoupling	MentorNet	Co-teaching
Pair-45%	49.50% ±0.42%	50.05% ±0.30%	48.21% ±0.55%	6.61% ±1.12%	48.80% ±0.04%	58.14% ±0.38%	**72.62%** ±0.15%
Symmetry-50%	48.87% ±0.52%	50.66% ±0.56%	46.15% ±0.76%	59.83% ±0.17%	51.49% ±0.08%	71.10% ±0.48%	**74.02%** ±0.04%
Symmetry-20%	76.25% ±0.28%	77.01% ±0.29%	76.84% ±0.66%	**84.55%** ±0.16%	80.44% ±0.05%	80.76% ±0.36%	82.32% ±0.07%

Table 6: Average test accuracy on *CIFAR-100* over the last ten epochs.

Flipping,Rate	Standard	Bootstrap	S-model	F-correction	Decoupling	MentorNet	Co-teaching
Pair-45%	31.99% ±0.64%	32.07% ±0.30%	21.79% ±0.86%	1.60% ±0.04%	26.05% ±0.03%	31.60% ±0.51%	**34.81%** ±0.07%
Symmetry-50%	25.21% ±0.64%	21.98% ±6.36%	18.93% ±0.39%	41.04% ±0.07%	25.80% ±0.04%	39.00% ±1.00%	**41.37%** ±0.08%
Symmetry-20%	47.55% ±0.47%	47.00% ±0.54%	41.51% ±0.60%	**61.87%** ±0.21%	44.52% ±0.04%	52.13% ±0.40%	54.23% ±0.08%

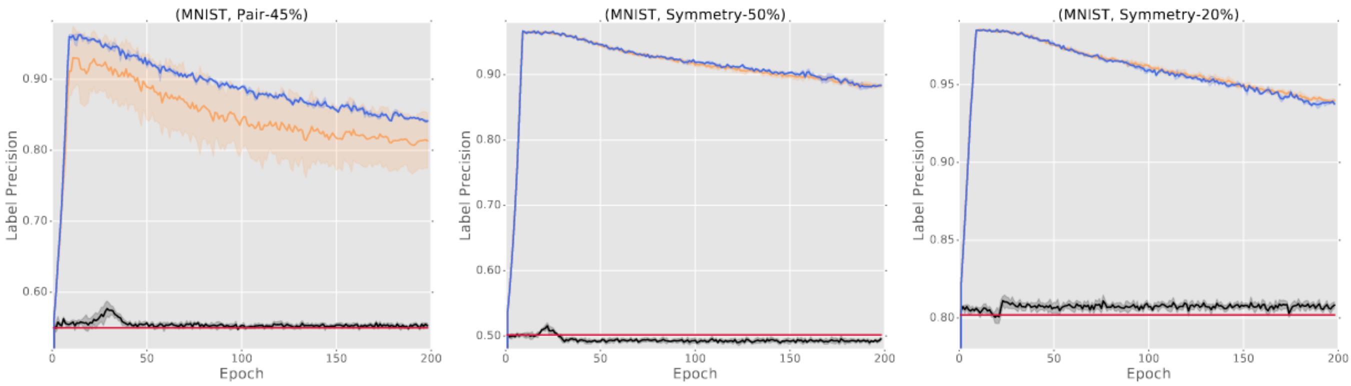
可以看出基于sample selection的模型在强噪声下效果比噪声建模的方法好得多

[t,l]%e



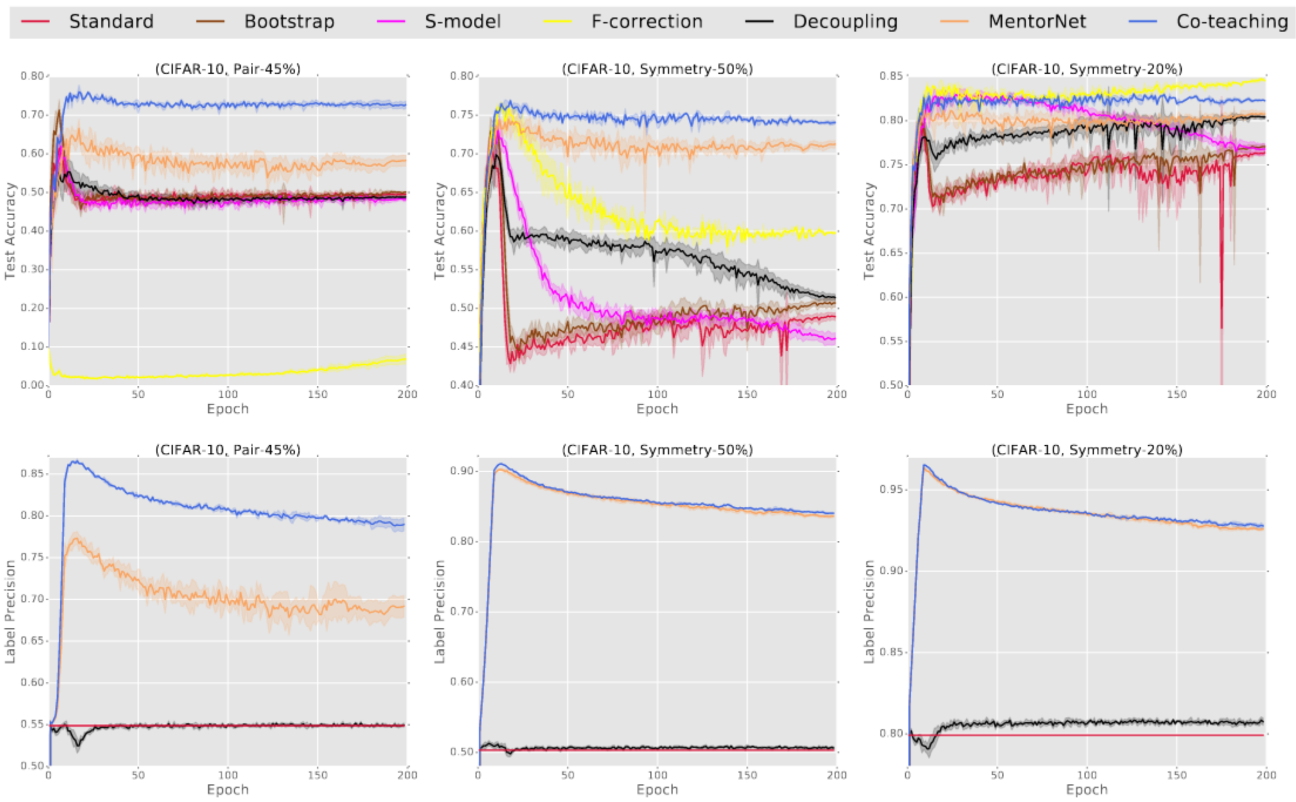
(a) Pair-45%. (b) Symmetry-50%. (c) Symmetry-20%.

Figure 3: Test accuracy vs. number of epochs on *MNIST* dataset.



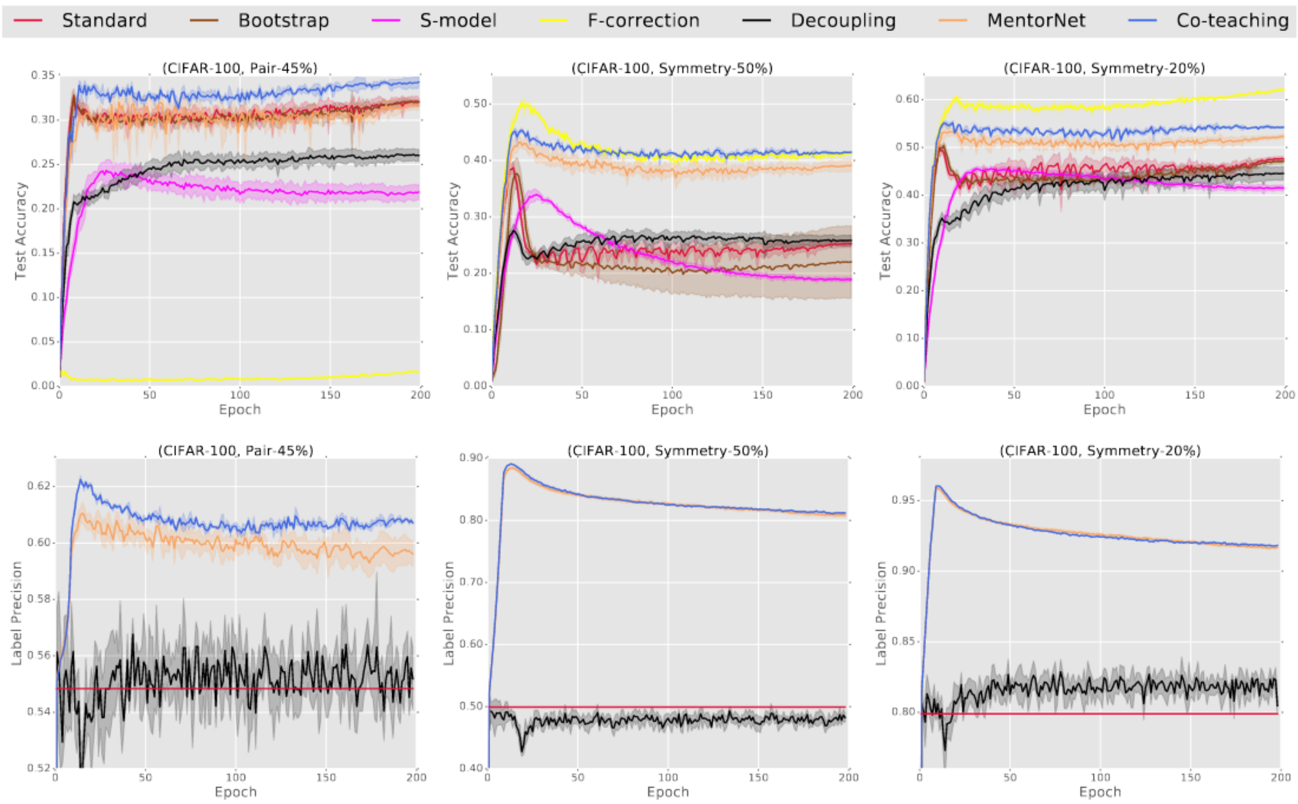
(a) Pair-45%. (b) Symmetry-50%. (c) Symmetry-20%.

Figure 4: Label precision vs. number of epochs on *MNIST* dataset.



(a) Pair-45%. (b) Symmetry-50%. (c) Symmetry-20%.

Figure 5: Results on *CIFAR-10* dataset. Top: test accuracy vs. number of epochs; bottom: label precision vs. number of epochs.



(a) Pair-45%. (b) Symmetry-50%. (c) Symmetry-20%.

Figure 6: Results on *CIFAR-100* dataset. Top: test accuracy vs. number of epochs; bottom: label precision vs. number of epochs.

在sample-selection类模型中:

Furthermore, if the specified method belongs to the "sample selection" category, *label precision* and *label recall* [112], [135] can be used as the metrics,

$$\text{Label Precision} = \frac{|\{(x_i, \tilde{y}_i) \in \mathcal{S}_t : \tilde{y}_i = y_i\}|}{|\mathcal{S}_t|},$$
$$\text{Label Recall} = \frac{|\{(x_i, \tilde{y}_i) \in \mathcal{S}_t : \tilde{y}_i = y_i\}|}{|\{(x_i, \tilde{y}_i) \in \mathcal{B}_t : \tilde{y}_i = y_i\}|}, \tag{23}$$

where $\mathcal{S}_t$ is the set of selected clean examples in a mini-batch $\mathcal{B}_t$. The two metrics are performance indicators for the examples selected from the mini-batch as true-labeled ones [112].

可以看出因为co-teaching是2个网络所以在pair noise下label precision要好于mentornet;

decoupling的label precision很低，因为这个方法并不是基于small loss选的样本