CroSel: Cross Selection of Confident Pseudo Labels for Partial-Label Learning

Supplementary Material

A. Discussion about the efficiency of CroSel.

Our method can be viewed as an *example-level* method without interactions between the examples. As a result, it does not suffer from a sharp drop in efficiency as n and k increase. Regarding the update operation, we only need to retain the output of the last layer of the model for each example and update the previous records in the memory bank (MB). This process has a time complexity of $\mathcal{O}(tnk)$. For the selection operation, we just find the maximum and calculate the mean of historical prediction stored in MB of each example, and the time complexity remains $\mathcal{O}(tnk)$.

B. Experimental details

B.1. Datasets

- **CIFAR-10:** It contains 60,000 32×32 RGB color pictures in a total of 10 categories. Including 50,000 for the training set and 10,000 for the test set.
- **CIFAR-100:** It has 100 classes, each containing 600 32×32 RGB color images. Each category has 500 training images and 100 test images. The 100 classes in CIFAR-100 are divided into 20 superclasses. Each image has a "fine" tag (the class to which it belongs) and a "rough" tag (the superclass to which it belongs).
- **SVHN:** It is derived from Google Street View door numbers, each image contains a set of Arabic numbers '0-9'. The training set contained 73,257 numbers, the test set 26,032 numbers, and 531,131 additional numbers. Each number is a 32×32 color picture.

B.2. Data augmentations

Data augmentation is widely used in weakly supervised learning algorithms. There are two types of data augmentations used in our algorithm: "weak" and "strong". For "weak" augmentation, it is just a standard flip-and-shift augmentation strategy consisting of Randomcrop and RandomHorizontalFlip. For "strong" augmentations, we use the RandAugment strategy for all, which randomly selects the type and magnitude of data augmentation with the same probability.

B.3. Compared methods

We reimplement CC, PRODEN, LWS, and CRDPLL using the same training scheme as CroSel. For PiCO and POP, we just follow their original training schemes and change the backbone to WRN-34-10. For a fair comparison, we add weak augmentation to those methods that are not equipped with data augmentation in their original scheme (CC, PRODEN, LWS). For other methods already equipped with augmentation (POP, PiCO, CRDPLL), we follow their settings illustrated in their paper.

B.4. Implementations

We set the batch size as 64 and total epochs as 200, using SGD as optimizer with a momentum of 0.9, and set the initial learning rate as 0.1, which is divided by 10 after 100 and 150 epochs respectively. For the hyper-parameters in our method, we set $t = 3$, $\alpha = 0.75$, $T = 0.5$ for all datasets, and $\lambda_{\text{cr}} = 1$ for CIFAR-100, $\lambda_{\text{cr}} = 4$ for others. For the selection threshold, we set $\gamma = 0.9$ for CIFAR-type datasets, and $\gamma = 0.85$ for SVHN.

B.5. Detailed results and more ablation experiment

Due to constraints on page space, we are unable to display all experimental results within the main text. Therefore, we focus on presenting the most crucial metrics or visualization outcomes. Specific experimental results and some more ablation experiments will be showcased in this section.

Discussion about the influence of parameter t and γ. As we mentioned before, t represents the length of historical prediction stored in MB, while γ represents the select threshold for the average prediction confidence of the model for the example prediction in the past t epochs. The two parameters determine the strictness of our selection criteria.

In the main text, we exclusively presented the final test accuracy. However, here we provide additional insights by showcasing the number of selection and selection accuracy in Tables 7 and 8. Regarding the parameter t, an increase in the storage length of the memory bank generally signifies a more stringent selection criterion. This results in a higher loss of selection numbers in exchange for a slight improvement in selecting accuracy. For the parameter γ, the model is more sensitive to some changes on this threshold. Lowering the threshold tends to introduce more noise, resulting in a decline in model performance, which subsequently affects the selection accuracy.

Detailed results for comparison between Dual model and Single model. In the main text, we only displayed the visual graph illustrating the number of selection. Here, we provide specific values for multiple metrics in Table 9. On one hand, the dual model substantially increases the number of selected pseudo-labeled samples without diminishing the selection accuracy, which demonstrates the effectiveness of the dual model. On the other hand, even without the utiliza-

Table 7. Detailed results for Parameter test on t.

Setting	t	Accuracy	S-ratio	S-acc
CIFAR-10 $q = 0.3$	$t = 2$	97.03%	99.07%	99.62%
	$t = 3$	97.50%	98.10%	99.55%
	$t = 4$	96.15%	89.23%	99.76%
CIFAR-100 $q = 0.1$	$t = 2$	82.74%	87.32%	98.50%
	$t = 3$	84.07%	93.61%	97.93%
	$t = 4$	83.56%	93.24%	98.23%

Table 8. Detailed results for Parameter test on γ.

Setting	γ	Accuracy	S-ratio	S-acc
CIFAR-10 $q = 0.3$	$\gamma = 0.8$	96.24%	95.29%	99.08%
	$\gamma = 0.9$	97.50%	98.10%	99.55%
	$\gamma = 0.95$	97.38%	93.10%	99.87%
CIFAR-100 $q = 0.1$	$\gamma = 0.8$	80.20%	82.11%	98.63%
	$\gamma = 0.9$	84.07%	93.61%	97.93%
	$\gamma = 0.95$	81.23%	67.32%	99.26%

tion of the dual model, we do not observe a sharp decrease in test accuracy, highlighting the stability of CroSel, whose performance improvement does not rely solely on the dual model.

Table 9. Detailed results of comparison between Dual model and Single model.

Setting	Model	Accuracy	S-ratio	S-acc
CIFAR-10 $q = 0.5$	Single model	96.51%	87.61%	99.72%
	Dual model	97.34%	96.25%	99.44%
CIFAR-100 $q = 0.1$	Single model	81.39%	85.39%	98.35%
	Dual model	84.07%	93.61%	97.93%

Detailed results for Parameter test on λ_{cr}. As mentioned earlier, λ_{d} weights the contribution of the consistency regularization term to the training loss. As described in Eq. (14), the parameter λ_{d} is directly influenced by the hyperparameter λ_{cr}. We tested λ_{cr} values of $\{1, 2, 4\}$ and $\lambda_{\text{cr}}(\text{fix})$ values of $\{0.5, 1, 2\}$.

In the main text, we visualized the evolution of various metrics as the training epoch progresses. The specific values at the end of training are provided in this section. Table 10 serves as an effective demonstration of our algorithm's robustness to the parameter λ_{cr}. Using dynamically varying λ_{cr} tends to result in higher selection accuracy compared to using a fixed value of λ_{cr}, thereby impacting the model performance. This observation aligns with our original intention of reducing the contribution of regularization terms in the final loss towards the end of training, thereby transitioning the model back to a simpler supervised learning scenario.

Table 10. Detailed results for Parameter test on λ_{cr}.

Setting	λ_{cr}	Accuracy	S-ratio	S-acc
CIFAR-10 $q = 0.5$	$\lambda_{\text{cr}} = 1$	95.94%	91.57%	99.51%
	$\lambda_{\text{cr}} = 2$	96.80%	97.32%	99.17%
	$\lambda_{\text{cr}} = 4$	97.33%	96.25%	99.44%
	$\lambda_{\text{cr}} = 1(\text{fixed})$	96.88%	87.88%	99.73%
	$\lambda_{\text{cr}} = 2(\text{fixed})$	96.95%	92.19%	99.68%
	$\lambda_{\text{cr}} = 0.5(\text{fixed})$	96.16%	95.21%	99.44%
CIFAR-100 $q = 0.1$	$\lambda_{\text{cr}} = 1$	84.07%	93.61%	97.93%
	$\lambda_{\text{cr}} = 2$	83.61%	94.15%	97.78%
	$\lambda_{\text{cr}} = 4$	83.88%	95.63%	97.59%
	$\lambda_{\text{cr}} = 1(\text{fixed})$	83.91%	99.35%	96.12%
	$\lambda_{\text{cr}} = 2(\text{fixed})$	84.03%	99.17%	96.15%
	$\lambda_{\text{cr}} = 0.5(\text{fixed})$	83.48%	97.02%	96.81%

Influence for the data augmentation on D_{sel}. Data augmentation plays a crucial role in weakly supervised learning. However, our selection criteria rely on historical prediction to select examples with high confidence. Consequently, we cannot guarantee that employing stronger data augmentation strategy will invariably yield superior results and selection effects. The randomness and variability inherent in data augmentation may introduce some adverse effects on the model's memorization capabilities, particularly as the strength of the augmentation strategy increases. Therefore, in this section, we delve into the impact of data augmentation on $\mathcal{D}_{\text{sel}}$ and its influence on label selection and overall training dynamics.

As shown in Table 11, weak augmentation is a more appropriate and effective choice. However, because of the impact of consistency regularization items, even if data augmentation is not used on $\mathcal{D}_{\text{sel}}$, there is no significant performance degradation. It is noting that strong data augmentation has an adverse effect on the selection of examples, especially in CIFAR-100, which may be related to the fact that the historical predictions stored in MB are produced by data that have not been augmented.

Table 11. Detailed results for Parameter test on $\mathcal{D}_{\text{sel}}$.

Setting	Data augmentation	Accuracy	S-ratio	S-acc
CIFAR-10 $q = 0.3$	None	96.91%	97.44%	99.60%
	Weak	97.50%	98.10%	99.55%
	Strong	97.22%	96.26%	99.75%
CIFAR-100 $q = 0.1$	None	83.74%	90.48%	98.32%
	Weak	84.07%	93.61%	97.93%
	Strong	81.01%	79.16%	98.98%