Poster Number: MLSP-P29.5

ESA: Expert-and-Samples-Aware Incremental Learning under Longtail Distribution

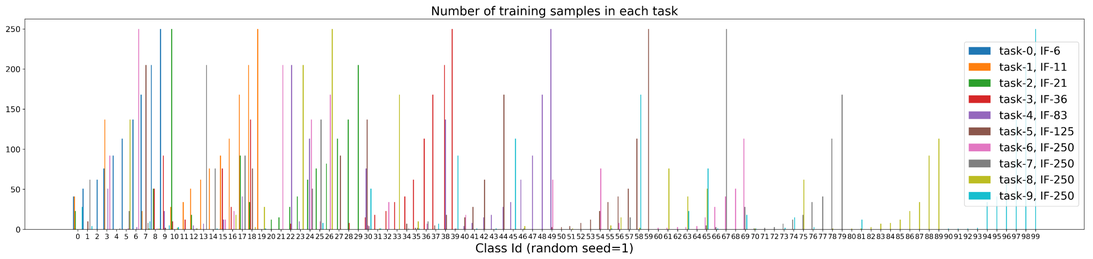
Jie Mei*, Jenq-Neng Hwang*
*Department of ECE, University of Washington, Seattle, WA, USA
https://ipl-uw.github.io/

Abstract

Most works in class incremental learning (CIL) assume disjoint sets of classes as tasks. Although a few works deal with overlapped sets of classes, they either assume a balanced data distribution or assume a mild imbalanced distribution. Instead, in this paper, we explore one of the understudied real-world CIL settings where (1) different tasks can share some classes but with new data samples, and (2) the training data of each task follows a long-tail distribution. We call this setting CIL-LT. We hypothesize that previously trained classification heads possess prototype knowledge of seen classes and thus could help learn the new model. Therefore, we propose a method with the multi-expert idea and a dynamic weighting technique to deal with the exacerbated forgetting introduced by the long-tail distribution. Experiments show that the proposed method effectively improves the accuracy in the CIL-LT setup on MNIST, CIFAR10, and CIFAR100. Code and data splits will be released.

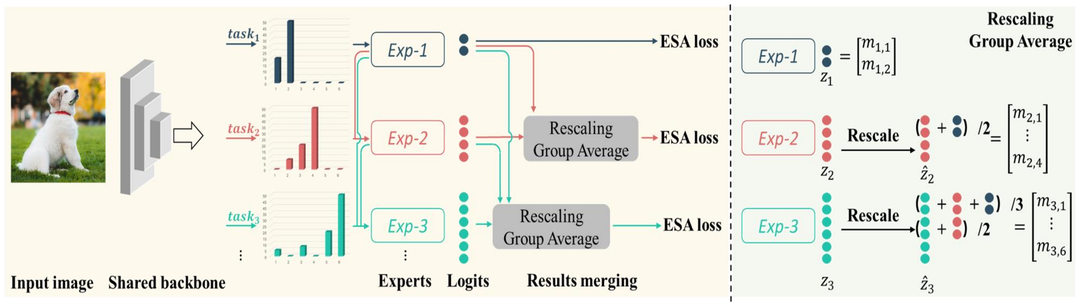
CIFAR100 Under CIL-LT

Each task includes 10 novel classes and 30% of seen classes that are randomly selected. Within each task, the training data follows a long-tail distribution. The maximum imbalance factor (IF) is 250, which is defined as the number of training samples in the largest class divided by that of the smallest. Thus, we call this split CIFAR100-Overlap30-IF250. Note that: We use 3 different seeds to generate 3 different class orders for evaluation purposes.



Pipeline

Most previous works reinitialize classification heads when new tasks come in while some work concatenate logits from non-overlapped heads to keep a unified classification head. However, we hypothesize that previously trained classification heads possess prototype knowledge of seen classes and will help the continual learning of the new model. In general, our proposed model dynamically creates an overlapped expert, i.e., a light-weight linear head, for each task as shown in the following figure. The remaining questions are (1) how to merge their decisions to get a final prediction for each sample, and (2) how to maintain their expertise as more novel classes come in and more seen classes wither away.



Multi-Expert Strategy

To tackle the first question, we make our experts output their knowledge, i.e., logits (before softmax), on all novel classes and classes seen so far as shown in the above Figure. Thus, overlapped classes among different tasks have corresponding outputs from different experts which gives experts the opportunity to communicate with each other via the proposed rescaling-group-average operation as shown in the right of the above figure.

Experts-and-Samples-Aware Loss

To tackle the second question, i.e., how to maintain experts' expertise as more novel classes come in and more seen classes wither away, we propose to apply an experts-and-samples-aware (ESA) loss to the merged logits, i.e., m_T, in each task.

Sampless-Aware Loss

$$L_{EA}^{T} = -\sum_{j} y \log\left(\sigma\left(\mathbf{m}_T\right)\right)$$

Samples-Aware Loss

$$L_{SA}^{T} = -\sum_{j_T^*} y \log\left(\sigma\left(\mathbf{m}_T\right)\right) \cdot w_{T,c_{j_T^*}},$$

$$w_{T,c_{j_T^*}} = log\left(\frac{N_{T,c_{j_T^*}} \cdot P_T}{M_{T,c_{j_T^*}}} + 1\right) + 1,$$

Experiments

Table 1: **Comparison** with two metrics (A{5, 10} and I{5, 10}: %) in MNIST-Overlap50-IF100, CIFAR10-Overlap50-IF100, and CIFAR100-Overlap30-IF250 under CIL-LT setup. K represents memory size.

Methods	MNIST10 (K=200)		CIFAR10 (K=200)		CIFAR100 (K=2,000)	
	A5(↑)	**I5(↓)**	**A5(↑)**	**I5(↓)**	**A10(↑)**	**I10(↓)**
Joint	88.49	-	91.40	-	62.16	-
EWC [2]	$39.23_{\pm 0.98}$	$49.26_{\pm 0.84}$	$63.16_{\pm 1.02}$	$28.24_{\pm 0.78}$	$27.67_{\pm 0.85}$	$34.49_{\pm 0.54}$
Rwalk [12]	$39.73_{\pm 0.88}$	$48.76_{\pm 0.69}$	$64.38_{\pm 0.87}$	$27.02_{\pm 0.98}$	$32.56_{\pm 0.99}$	$29.60_{\pm 0.77}$
iCaRL [1]	$59.78_{\pm 0.73}$	$28.71_{\pm 0.63}$	$52.16_{\pm 0.78}$	$39.24_{\pm 0.83}$	$33.74_{\pm 0.84}$	$28.42_{\pm 1.33}$
GDumb [13]	$41.71_{\pm 0.71}$	$46.78_{\pm 0.62}$	$37.22_{\pm 0.63}$	$55.18_{\pm 0.65}$	$28.33_{\pm 0.68}$	$33.83_{\pm 1.45}$
BiC [3]	$33.38_{\pm 0.77}$	$55.11_{\pm 0.59}$	$64.23_{\pm 0.66}$	$27.17_{\pm 0.76}$	$37.68_{\pm 0.69}$	$24.48_{\pm 0.99}$
DER++ [14]	$84.55_{\pm 0.56}$	$3.94_{\pm 0.55}$	$65.10_{\pm 0.54}$	$26.30_{\pm 0.89}$	$35.57_{\pm 0.56}$	$26.59_{\pm 1.52}$
RM [8]	$86.50_{\pm 0.43}$	$1.99_{\pm 0.49}$	$65.97_{\pm 0.33}$	$25.43_{\pm 0.56}$	$40.53_{\pm 0.49}$	$21.63_{\pm 0.76}$
CLS-ER [10]	$86.08_{\pm 0.40}$	$2.41_{\pm 0.37}$	$66.88_{\pm 0.35}$	$24.52_{\pm 0.62}$	$41.52_{\pm 0.61}$	$20.64_{\pm 0.51}$
ESA (ours)	**88.51**$_{\pm 0.31}$	**−0.02**$_{\pm 0.23}$	**71.32**$_{\pm 0.44}$	**20.08**$_{\pm 0.32}$	**44.21**$_{\pm 0.68}$	**17.95**$_{\pm 0.54}$

Table 3: **Comparison** with two metrics (A{10} and I{10}: %) in CIFAR100 for another two overlap ratios.

Methods	Overlap30		Overlap50	
	A10(↑)	**I10(↓)**	**A10(↑)**	**I10(↓)**
Joint	64.05	-	64.71	-
EWC [2]	$32.45_{\pm 0.78}$	$31.60_{\pm 0.96}$	$33.95_{\pm 0.63}$	$30.76_{\pm 0.98}$
Rwalk [12]	$36.22_{\pm 0.67}$	$27.83_{\pm 0.69}$	$38.05_{\pm 0.96}$	$26.66_{\pm 0.65}$
iCaRL [1]	$37.11_{\pm 0.57}$	$26.94_{\pm 0.91}$	$37.57_{\pm 1.22}$	$27.14_{\pm 1.63}$
GDumb [13]	$29.53_{\pm 0.96}$	$34.52_{\pm 0.87}$	$31.86_{\pm 1.53}$	$32.85_{\pm 0.75}$
BiC [3]	$42.08_{\pm 0.88}$	$21.97_{\pm 0.82}$	$43.51_{\pm 0.63}$	$21.20_{\pm 0.68}$
RM [8]	$45.52_{\pm 0.56}$	$18.53_{\pm 0.58}$	$48.01_{\pm 0.78}$	$16.70_{\pm 0.57}$
CLS-ER [10]	$45.86_{\pm 0.51}$	$18.19_{\pm 0.68}$	$47.57_{\pm 0.54}$	$17.14_{\pm 0.72}$
ESA (ours)	**47.22**$_{\pm 0.50}$	**16.83**$_{\pm 0.53}$	**50.06**$_{\pm 0.53}$	**14.65**$_{\pm 0.65}$