Continual-Zoo: Leveraging Zoo Models for Continual Classification of Medical Images

Supplementary Material

6. Datasets and Experimental Setup

To evaluate Continual-Zoo, we defined three medical benchmarks using publicly accessible datasets to facilitate reproducibility. For the **SKIN** domain, we utilized five skin lesion image datasets, each containing subsets of seven classes: melanocytic nevus (nv), melanoma (mel), basal cell carcinoma (bcc), dermatofibroma (df), benign keratosis (bkl), vascular lesion (vasc), and actinic keratosis (akiec). For instance, datasets like HAM and DMF encompass all seven classes, while UDA contains only two. We defined three continual learning (CL) scenarios for SKIN assessment:

1. Class-Incremental Learning (CIL): Here, a single dataset is divided into T tasks, each with non-overlapping classes. For example, CIL (HAM) splits the HAM dataset into three tasks, with two, two, and three classes in each task, respectively. Detailed information regarding the number of tasks and classes per task is provided in Fig. 5. Specifically, we created CIL (HAM), CIL (DMF) and CIL (D7P).

2. Domain-Incremental Learning (DIL): In this scenario, each dataset represents a separate task, with all tasks sharing the same set of classes. We create DIL (SKIN), comprising four tasks representing HAM, DMF, D7P, and MSK, respectively (see Fig. 5).

3. Domain- and Class-Incremental Learning (DCIL): DCIL (SKIN) is constructed with five tasks, representing UDA, MSK, D7P, DMF, and HAM, respectively. Each dataset represents a unique distribution and possesses a different set of classes, potentially with overlaps (Fig. 5).

Regarding the **BLOOD** dataset, it comprises 17,092 images categorized into eight classes: basophil (ba), eosinophil (eo), erythroblast (er), immature granulocytes (ig), lymphocyte (ly), monocyte (mo), neutrophil (ne), and platelet (pl). The **COLON** dataset consists of 107,180 images belonging to nine tissues: adipose (ad), background (bg), debris (de), lymphocytes (ly), mucus (mu), smooth muscle (sm), normal colon mucosa (nm), cancer-associated stroma (cs), and colorectal adenocarcinoma epithelium (ca).

For **BLOOD** and **COLON**, we solely evaluate them in the CIL setup, where each corresponding dataset is divided into $T = 4$ tasks with non-overlapping classes, similar to the SKIN scenario (Fig. 5).

For all the datasets, we use the official data split from source dataset (if provided) to avoid data leakage. If the source dataset has only a split of training and validation set, we use the official validation set as test set and split the official training set with a ratio of 9:1 into training-validation. For the dataset without an official split, we randomly select 70%, 10%, and 20% of the images for the training, validation, and test splits, where the random selection is stratified on class labels. For pre-processing, we center-crop and resize all images into 224×224 as network input, and balance the training sets using PyTorch sampler.

Table 6. Overview of pretrained models in Zoo-A to Zoo-E.

Zoo		Backbone	Architecture	Pretrained Scheme	Pretrained Data
A	1	ResNet-50 [20]	ResNet-50	Supervised	ImageNet
	2		ResNet-50		Abdominal CT
	3		MoCo [22]	Self-Supervised	ImageNet
	4		Mask R-CNN [21]	Supervised	COCO
	5		DeepLabV3 [10]		
	6		Keypoint R-CNN		
A*	1-6	ResNet-50	ResNet-50	Supervised	ImageNet
B	1	CNN	AlexNet [35]	Supervised	ImageNet
	2		DenseNet-121 [26]		
	3		DenseNet-169 [26]		
	4		ResNet-50		
	5		SE-ResNet-50 [23]		
	6		SqueezeNet-1.1 [27]		
C	1	ViT [35]	ViT-B/16 [35]	Supervised	ImageNet-21k
	2		ViT-L/16 [35]		
	3		DeiT-S/16 [53]		
	4		CoaT-Lite-M [62]		
	5		DINO-S/16 [8]	Self-Supervised	
	6		DINO-B/16 [8]		
D	1	CNN	ResNet-50	Supervised	ImageNet
	2		EfficientNetB3		
	3		DenseNet-101		
	4	ViT	ViT-B/16		ImageNet-21k
	5		ViT-L/16		
	6		DeiT-B/16		
E	1	ResNet-50	MoCo	Self-Supervised	ImageNet
	2		BYOL [18]		
	3		SimCLR [12]		
	4	ViT	DINO-S/16		ImageNet-21k
	5		DINO-B/16		
	6		DINO-B/8		

7. Zoo Details

Details about the zoo, including the number of pretrained models, backbone network, architecture, pretrained scheme, and pretrained data, are shown in Table 6.

8. Sequential Analysis on BLOOD and COLON

A sequential analysis of Continual-Zoo and other methods are reported in Fig. 6. Clearly, generative- and replay-based methods maintain a strong performance compared to regularization-based methods. The latter experiences a significant decline in performance, particularly in CIL (COLON), just after learning the first task.

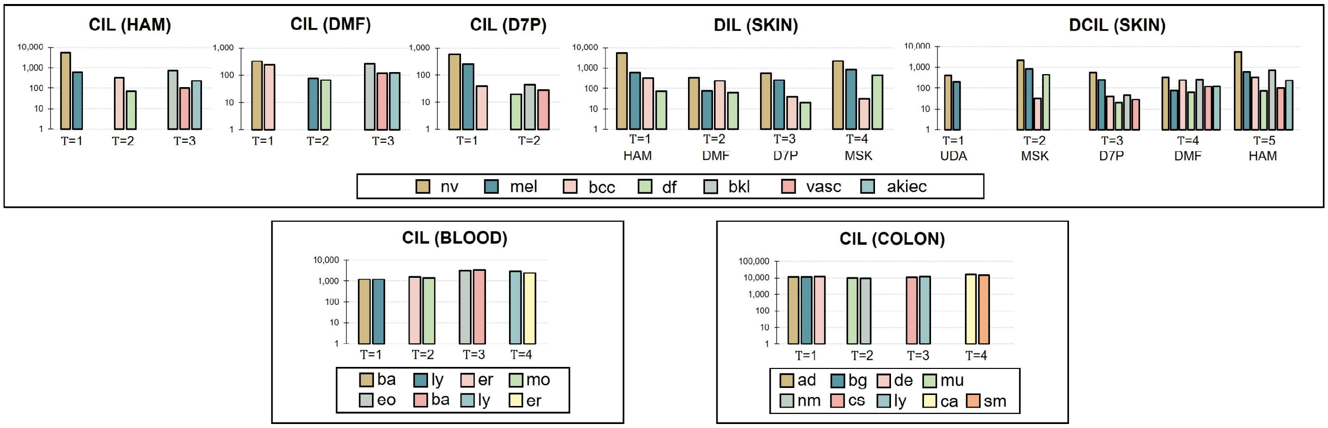

Figure 5. Class labels and distributions (in logarithmic scale, base 10) per task in each of the proposed evaluation benchmarks.

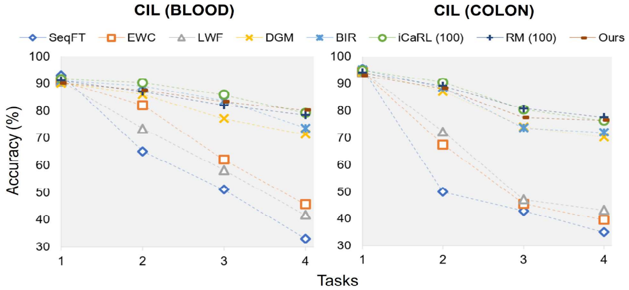

Figure 6. The accuracy of Continual-Zoo and other CL methods over the seen tasks after each training step in the continual sequence.