Dual-Teacher Class-Incremental Learning With Data-Free Generative Replay
— *Supplementary Materials* —

Yoojin Choi, Mostafa El-Khamy
SoC R&D, Samsung Semiconductor Inc.
San Diego, CA 92121, USA
{yoojin.c,mostafa.e}@samsung.com

Jungwon Lee
System LSI, Samsung Electronics
South Korea
jungwon2.lee@samsung.com

A. Experiments

A.1. Training procedure

At time 0 of CIL on CIFAR-100, the ResNet-32 model is trained with Nesterov's accelerated gradient (NAG) [22] of batch size 128 for 160 epochs by using the original CIFAR-100 data of randomly selected 50 base classes. The learning rate starts from 0.1 and is reduced to 0.01 and 0.001 at epoch 80 and 120, respectively. At time $i \geq 1$, to re-train the ResNet-32 model for CIL on additional CIFAR-100 classes, we use NAG of batch size 256 for 160 epochs, where each epoch consists of 50 batches. The learning rate starts from 0.1 and is reduced to 0.01 and 0.001 at epoch 80 and 120, respectively.

At time 0 of CIL on ImageNet-Subset and ImageNet-Full, the ResNet-18 models are trained with NAG of batch size 128 for 90 epochs by using the original ImageNet data of randomly selected 50 and 500 base classes, respectively. The learning rate starts from 0.1 and is reduced to 0.01 and 0.001 at epoch 30 and 60, respectively. At time $i \geq 1$, for CIL on ImageNet, we re-train the ResNet-18 models with NAG of batch size 100 for 90 epochs, where each epoch consists of 128 and 1280 batches for ImageNet-Subset and ImageNet-Full, respectively. The learning rate starts from 0.1 and is reduced to 0.01 and 0.001 at epoch 30 and 60, respectively.

At time $i \geq 1$, for CIL, the original (or exemplary) and synthetic samples for old and new classes are mixed in each batch by the following distribution, depending on their availability, where B denotes the batch size (see Table 2 for the notations of N_D, N_R, and N_g).

N_D	N_R	N_g	New classes		Old classes	
			Original (or exemplary)	Synthetic	Exemplary	Synthetic
>0	>0	0	$B/2$	0	$B/2$	0
		1	$B/2$	0	$B/4$	$B/4$
		2	$B/4$	$B/4$	$B/4$	$B/4$
>0	0	0	B	0	0	0
		1	$B/2$	0	0	$B/2$
		2	$B/4$	$B/4$	0	$B/2$

For the factor α_i in front of the less-forget (LF) loss in (10), we set

$$\alpha_i = \alpha_0 \sqrt{\frac{|C_0^{i-1}|}{|C_i|}},$$

with $\alpha_0 = 5$ for CIFAR-100 and $\alpha_0 = 10$ for ImageNet in the conventional CIL scenario, as suggested in [13], where $|C|$ denotes the size of set C. If we have no exemplars for old classes, we double α_0 to enforce a stronger less-forget constraint. In the data-limited CIL scenario, we search the best α_0 in $\{1, 5, 10\}$ for CIFAR-100 and in $\{2, 10, 20\}$ for ImageNet. We also use gradient clipping to avoid exploding gradients, as suggested in [1]—we clip the norm of gradients by 1.

For ImageNet-Full, we additionally use the margin ranking loss for inter-class separation and also perform class-balanced fine-tuning at the end of each incremental training, if there are reserved exemplars for old classes, in ours as well as when we

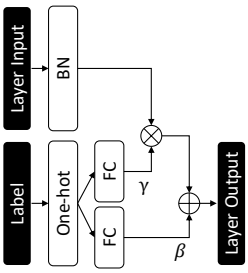
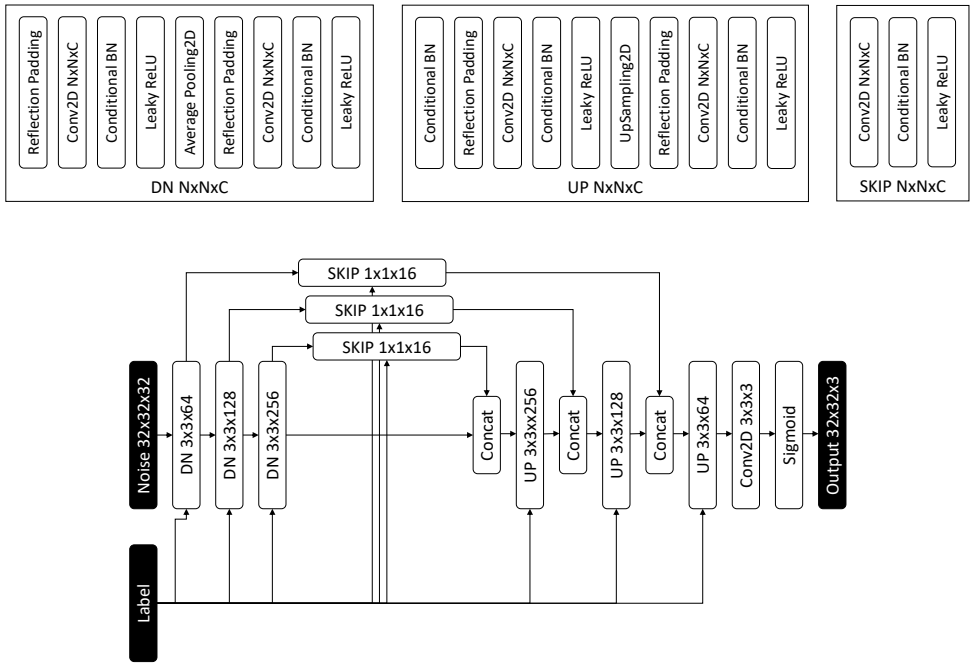


Figure 10: Conditional batch normalization. **BN** denotes a vanilla batch normalization layer without scaling and shifting after normalization. **One-hot** implies one-hot encoding. **FC** denotes a fully-connected layer. The symbols $\bigotimes$ and $\bigoplus$ stand for channel-wise scaling and channel-wise bias addition, respectively.

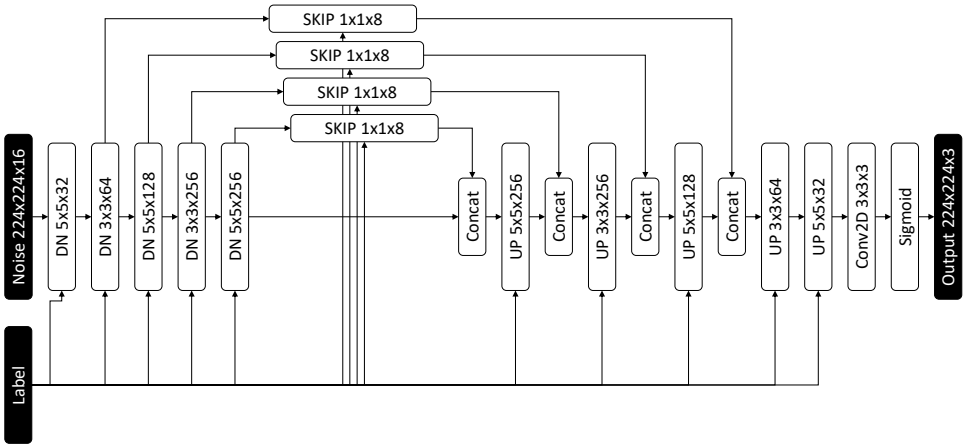
reproduce the baseline LUCIR [13]. Note that those two components were shown to be effective significantly for ImageNet-Full in [13]. For CIFAR-100 and ImageNet-Subset, we do not use those two additional components either in ours or when we reproduce the baseline LUCIR [13].

A.2. Generator architecture

In Figure 10, we illustrate conditional batch normalization (BN) used in our conditional generators for DF-GR. Observe that the channel-wise scaling factor γ and the channel-wise bias β depend on the input label (condition). In Figure 11, we depict the generator architectures used in our experiments for DF-GR in CIL on CIFAR-100 and ImageNet, respectively.



(a) Generator for DF-GR in CIL on CIFAR-100



(b) Generator for DF-GR in CIL on ImageNet-Subset and ImageNet-Full

Figure 11: Generator architectures for DF-GR in CIL on CIFAR-100 and ImageNet. **Conv2D NxNxC** denotes a convolutional layer of kernel size $N \times N$ with C output channels. **Reflection Padding** implies that we add reflection padding to make the input size (width and height) and the output size of the following convolution be the same. For **UpSampling2D**, we employ bilinear upsampling. **Conditional BN** can be found in Figure 10. **Concat** stands for concatenation. **Noise NxNxC** denotes the input random noise of shape $N \times N \times C$, sampled from the standard normal distribution.