Transfer without Forgetting
Supplementary Materials

Matteo Boschini[1], Lorenzo Bonicelli[1], Angelo Porrello[1], Giovanni Bellitto[2], Matteo Pennisi[2], Simone Palazzo[2], Concetto Spampinato[2], and Simone Calderara[1]

[1] AImageLab, University of Modena and Reggio Emilia, Italy
`firstname.lastname@unimore.it`
[2] PeRCeiVe Lab, University of Catania, Italy
`firstname.lastname@unict.it`

A Additional Details on the Model

In this section, we report some additional details on the inner workings of the model which were omitted in the main paper for the sake of brevity.

A.a Further details on $\mathbb{M}_{\text{Sp}}$

The spatial attention map $\mathbb{M}_{\text{Sp}}$ is computed on top of the activations of a given layer of the fixed sibling network $\widehat{h} \in \mathbb{R}^{b\times c\times h\times w}$, processed through a ResNet-inspired bottleneck structure [1,2]. In detail, we expand and detail Eq. 5 in the main paper:

$$\mathbb{M}_{\text{Sp}} \triangleq \mathrm{C}^{\mathrm{C}}_{1\times1} \circ \mathrm{ReLU} \circ \mathrm{BN} \circ \mathrm{C}^{\mathrm{B}}_{3\times3} \circ \mathrm{ReLU} \circ \mathrm{BN} \circ \mathrm{C}^{\mathrm{B}}_{3\times3} \circ \mathrm{ReLU} \circ \mathrm{BN} \circ \mathrm{C}^{\mathrm{A}}_{1\times1}, \quad (1)$$

where ReLU denotes a ReLU activation, BN indicates a Batch Normalization layer (conditioned on the task-identifier) and C indicates a Convolutional layer. More specifically, $\mathrm{C}^{\mathrm{A}}_{1\times1}$ is a 1×1 convolution, projecting from c channels to $c/4$; $\mathrm{C}^{\mathrm{B}}_{3\times3}$ is a 3×3 dilated convolution with dilation factor 2 and adequate padding to maintain the same spatial resolution as the input, with $c/4$ channels both as input and output; $\mathrm{C}^{\mathrm{C}}_{1\times1}$ is a 1×1 convolution projecting from $c/4$ channels to 1 channel. This results in $\mathbb{M}_{\text{Sp}}$ having shape $b\times1\times h\times w$.

A.b Scaling of $\mathbb{M}$

The second distillation term in Eq. 9 requires storing the binary attention maps $\mathbb{M}$ computed for each sample stored in the memory buffer. While this implies a memory overhead, we point out that this is limited by two factors:

- The binary nature of $\mathbb{M}$ means its elements can be saved using the smallest supported data-type (usually 1 byte due to hardware constraints);
- As $\mathbb{M}$ usually encodes low level features, it contains several redundancies that can be exploited by (a) using lossless compression algorithms, or (b) down-sampling its spatial dimensions before saving.

In TwF we save the feature maps $\mathbb{M}$ as bytes and apply down-scaling – with *nearest neighbor* rule – with a factor of 2 if the spatial dimensions are over 16×16. We use the same strategy to up-scale the maps before computing Eq. 9.

B Hyperparameters

For the experiments of Sec. 4, we employed a choice of hyperparameters validated by grid-search on a random split of 10% of the training set. In the following, we list the values resulting from this process, which can be used to replicate our result. For the sake of fairness, we initialize all models from the same pre-training weights and fix the allowance in terms of iterations and sample efficiency by excluding the number of epochs, lr decay schedule and batch size from the grid-search[3].

Split CIFAR-10 - Class-IL		
shared		Eps : 50 bs : 32 $\text{Eps}_{\text{pretr}}$: 200 lr_{decay} : no
JOINT	-	lr : 0.1
SGD	-	lr : 0.1
oEwC	-	lr : 0.1 λ : 10 γ : 1
LwF	-	lr : 0.1 α : 0.3 τ : 2 wd : 0.0001
ER	500	lr : 0.1
	5120	lr : 0.1
CO^2L	500	lr : 0.5 bs : 256 κ : 0.2 λ : 1 lr_{lin} : 1 $\text{lr}^{\text{lin}}_{\text{decay}}$: 0.2 κ^* : 0.01 τ : 0.07
	5120	lr : 0.5 bs : 256 κ : 0.2 λ : 1 lr_{lin} : 1 $\text{lr}^{\text{lin}}_{\text{decay}}$: 0.2 κ^* : 0.01 τ : 0.07
iCaRL	500	lr : 0.1 wd : 10^{-5}
	5120	lr : 0.03 wd : 10^{-5}
DER++	500	lr : 0.03 α : 0.2 β : 0.5
	5120	lr : 0.03 α : 0.1 β : 1
ER-ACE	500	lr : 0.03
	5120	lr : 0.03
TwF	500	lr : 0.03 α : 0.3 β : 0.9 λ : 0.1 λ_{FP} : 5×10^{-3} $\lambda^{\text{repl}}_{\text{FP}}$: 0.1
	5120	lr : 0.1 α : 0.3 β : 0.9 λ : 0.1 λ_{FP} : 5×10^{-3} $\lambda^{\text{repl}}_{\text{FP}}$: 0.3

[3] It must be noted that, to allow for its regular operation, CO^2L demands a larger batch size. All results for this method are influenced by this advantage.

Split CIFAR-10 - Task-IL		
shared		Eps : 50 bs : 32 $\text{Eps}_{\text{pretr}}$: 200 lr_{decay} : no
JOINT	-	lr : 0.1
SGD	-	lr : 0.1
oEwC	-	lr : 0.03 λ : 0.5 γ : 1
LwF	-	lr : 0.01 α : 0.3 τ : 2 wd : 0.0001
ER	500	lr : 0.1
	5120	lr : 0.1
CO^2L	500	lr : 0.5 bs : 256 κ : 0.2 λ : 1 lr_{lin} : 1 $\text{lr}^{\text{lin}}_{\text{decay}}$: 0.2 κ^* : 0.01 τ : 0.07
	5120	lr : 0.5 bs : 256 κ : 0.2 λ : 1 lr_{lin} : 1 $\text{lr}^{\text{lin}}_{\text{decay}}$: 0.2 κ^* : 0.01 τ : 0.07
iCaRL	500	lr : 0.1 wd : 10^{-5}
	5120	lr : 0.03 wd : 10^{-5}
DER++	500	lr : 0.03 α : 0.2 β : 0.5
	5120	lr : 0.03 α : 0.1 β : 1
ER-ACE	500	lr : 0.03
	5120	lr : 0.03
TwF	500	lr : 0.03 α : 0.3 β : 0.9 λ : 0.1 λ_{FP} : 5×10^{-3} $\lambda^{\text{repl}}_{\text{FP}}$: 0.1
	5120	lr : 0.1 α : 0.3 β : 0.9 λ : 0.1 λ_{FP} : 5×10^{-3} $\lambda^{\text{repl}}_{\text{FP}}$: 0.3

Split CIFAR-100 - Class-IL		
shared		Eps : 50 bs : 64 $\text{Eps}_{\text{pretr}}$: 200 lr_{decay} : 0.1 $\text{lr}^{\text{steps}}_{\text{decay}}$: $[35, 45]$
JOINT	-	lr : 0.1
SGD	-	lr : 0.1
oEwC	-	lr : 0.1 λ : 5 γ : 1
LwF	-	lr : 0.03 α : 0.3 τ : 2 wd : 0.0005
ER	500	lr : 0.01
	2000	lr : 0.01
CO^2L	500	lr : 0.1 bs : 256 κ : 0.2 λ : 1 lr_{lin} : 1 $\text{lr}^{\text{lin}}_{\text{decay}}$: 0.2 κ^* : 0.01 τ : 0.07
	2000	lr : 0.1 bs : 256 κ : 0.2 λ : 1 lr_{lin} : 1 $\text{lr}^{\text{lin}}_{\text{decay}}$: 0.2 κ^* : 0.01 τ : 0.07
iCaRL	500	lr : 1 wd : 10^{-5}
	2000	lr : 1 wd : 10^{-5}
DER++	500	lr : 0.1 α : 0.3 β : 0.3
	2000	lr : 0.1 α : 0.1 β : 0.5
ER-ACE	500	lr : 0.1
	2000	lr : 0.1
TwF	500	lr : 0.03 α : 0.3 β : 1.2 λ : 0.3 λ_{FP} : 0.03 $\lambda^{\text{repl}}_{\text{FP}}$: 1.5
	2000	lr : 0.1 α : 0.3 β : 1.2 λ : 0.3 λ_{FP} : 5×10^{-3} $\lambda^{\text{repl}}_{\text{FP}}$: 1.2

Split CIFAR-100 - Task-IL		
shared		$\text{Eps}: 50$ $\text{bs}: 64$ $\text{Eps}_{\text{pretr}}: 200$ $\text{lr}_{\text{decay}}: 0.1$ $\text{lr}^{\text{steps}}_{\text{decay}}: [35, 45]$
JOINT	-	$\text{lr}: 0.1$
SGD	-	$\text{lr}: 0.01$
oEwC	-	$\text{lr}: 0.01$ $\lambda: 0.5$ $\gamma: 0.7$
LwF	-	$\text{lr}: 0.03$ $\alpha: 0.3$ $\tau: 2$ $\text{wd}: 0.0005$
ER	500	$\text{lr}: 0.01$
	2000	$\text{lr}: 0.01$
CO^2L	500	$\text{lr}: 0.1$ $\text{bs}: 256$ $\kappa: 0.2$ $\lambda: 1$ $\text{lr}_{\text{lin}}: 1$ $\text{lr}^{\text{lin}}_{\text{decay}}: 0.2$ $\kappa^*: 0.01$ $\tau: 0.07$
	2000	$\text{lr}: 0.1$ $\text{bs}: 256$ $\kappa: 0.2$ $\lambda: 1$ $\text{lr}_{\text{lin}}: 1$ $\text{lr}^{\text{lin}}_{\text{decay}}: 0.2$ $\kappa^*: 0.01$ $\tau: 0.07$
iCaRL	500	$\text{lr}: 1$ $\text{wd}: 10^{-5}$
	2000	$\text{lr}: 1$ $\text{wd}: 10^{-5}$
DER++	500	$\text{lr}: 0.1$ $\alpha: 0.3$ $\beta: 1.2$
	2000	$\text{lr}: 0.1$ $\alpha: 0.1$ $\beta: 0.5$
ER-ACE	500	$\text{lr}: 0.1$
	2000	$\text{lr}: 0.1$
TwF	500	$\text{lr}: 0.03$ $\alpha: 0.3$ $\beta: 1.2$ $\lambda: 0.3$ $\lambda_{\text{FP}}: 0.03$ $\lambda^{\text{repl}}_{\text{FP}}: 1.5$
	2000	$\text{lr}: 0.1$ $\alpha: 0.3$ $\beta: 0.8$ $\lambda: 0.3$ $\lambda_{\text{FP}}: 0.03$ $\lambda^{\text{repl}}_{\text{FP}}: 0.3$

Split CUB-200 - Class-IL		
shared		$\text{Eps}: 50$ $\text{bs}: 64$ $\text{Eps}_{\text{pretr}}: 50$
JOINT	-	$\text{lr}: 0.1$
SGD	-	$\text{lr}: 0.1$
oEwC	-	$\text{lr}: 0.01$ $\lambda: 1$ $\gamma: 1$
LwF	-	$\text{lr}: 0.1$ $\alpha: 1$ $\tau: 2$ $\text{wd}: 0.0005$
ER	400	$\text{lr}: 0.03$
	1000	$\text{lr}: 0.1$
CO^2L	400	$\text{lr}: 0.1$ $\text{bs}: 256$ $\kappa: 0.2$ $\lambda: 1$ $\text{lr}_{\text{lin}}: 1$ $\text{lr}^{\text{lin}}_{\text{decay}}: 0.2$ $\kappa^*: 0.01$ $\tau: 0.07$
	1000	$\text{lr}: 0.1$ $\text{bs}: 256$ $\kappa: 0.2$ $\lambda: 1$ $\text{lr}_{\text{lin}}: 1$ $\text{lr}^{\text{lin}}_{\text{decay}}: 0.2$ $\kappa^*: 0.01$ $\tau: 0.07$
iCaRL	400	$\text{lr}: 0.1$ $\text{wd}: 10^{-5}$
	1000	$\text{lr}: 0.1$ $\text{wd}: 10^{-5}$
DER++	400	$\text{lr}: 0.1$ $\alpha: 1$ $\beta: 0.5$
	1000	$\text{lr}: 0.1$ $\alpha: 0.5$ $\beta: 0.5$
ER-ACE	400	$\text{lr}: 0.1$
	1000	$\text{lr}: 0.1$
TwF	400	$\text{lr}: 0.03$ $\alpha: 1$ $\beta: 1$ $\lambda: 0.3$ $\lambda_{\text{FP}}: 5 \times 10^{-4}$ $\lambda^{\text{repl}}_{\text{FP}}: 0.1$
	1000	$\text{lr}: 0.03$ $\alpha: 1$ $\beta: 1.2$ $\lambda: 0.3$ $\lambda_{\text{FP}}: 5 \times 10^{-4}$ $\lambda^{\text{repl}}_{\text{FP}}: 0.1$

Split CUB-200 - Task-IL		
shared		Eps : 50 bs : 64 $\text{Eps}_{\text{pretr}}$: 50
JOINT	-	lr : 0.1
SGD	-	lr : 0.1
oEwC	-	lr : 0.1 λ : 0.5 γ : 0.9
LwF	-	lr : 0.1 α : 1 τ : 2 wd : 0.0005
ER	400	lr : 0.1
	1000	lr : 0.1
CO^2L	400	lr : 0.1 bs : 256 κ : 0.2 λ : 1 lr_{lin} : 1 $\text{lr}^{\text{lin}}_{\text{decay}}$: 0.2 κ^* : 0.01 τ : 0.07
	1000	lr : 0.1 bs : 256 κ : 0.2 λ : 1 lr_{lin} : 1 $\text{lr}^{\text{lin}}_{\text{decay}}$: 0.2 κ^* : 0.01 τ : 0.07
iCaRL	400	lr : 0.1 wd : 10^{-5}
	1000	lr : 0.1 wd : 10^{-5}
DER++	400	lr : 0.1 α : 0.5 β : 0.5
	1000	lr : 0.1 α : 0.5 β : 0.5
ER-ACE	400	lr : 0.1
	1000	lr : 0.1
TwF	400	lr : 0.03 α : 0.3 β : 1 λ : 0.3 λ_{FP} : 5×10^{-4} $\lambda^{\text{repl}}_{\text{FP}}$: 0.1
	1000	lr : 0.03 α : 1 β : 1 λ : 0.3 λ_{FP} : 5×10^{-4} $\lambda^{\text{repl}}_{\text{FP}}$: 0.1

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2016)
2. Park, J., Woo, S., Lee, J.Y., Kweon, I.S.: Bam: Bottleneck attention module. In: British Machine Vision Conference (2018)