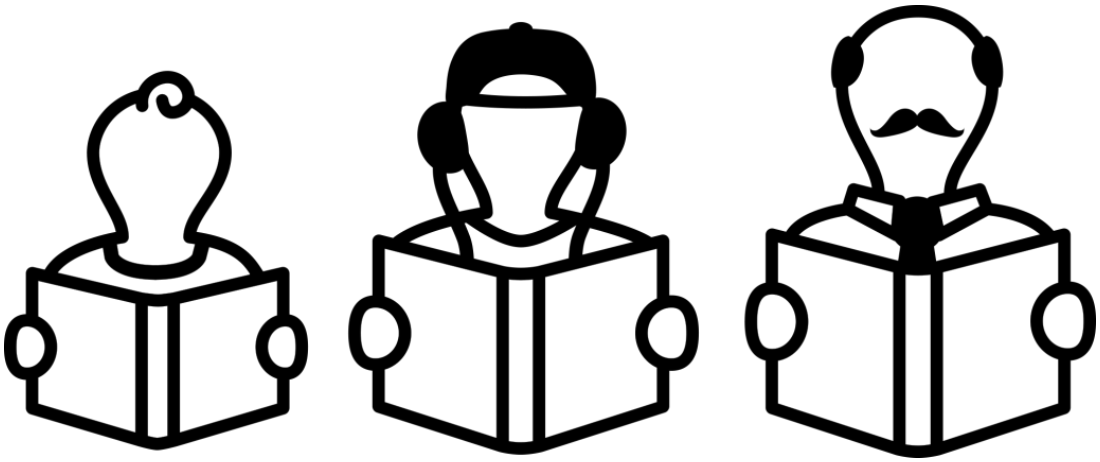
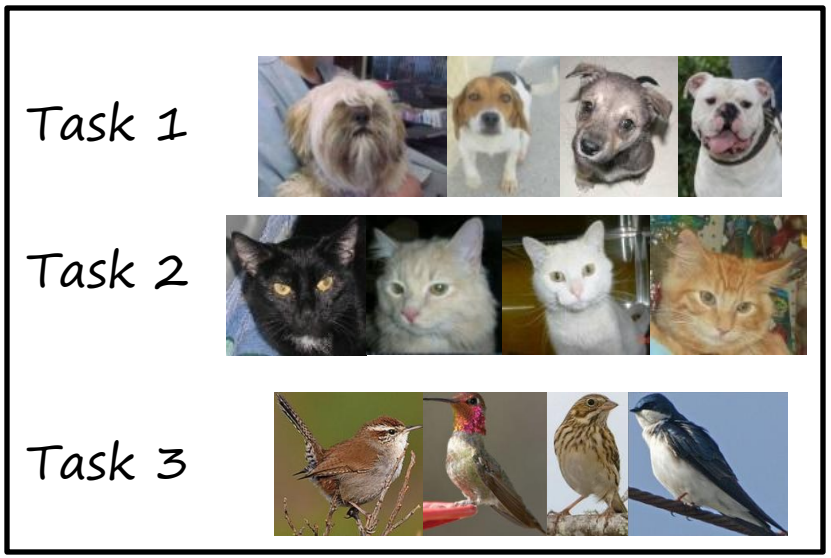
Rotate your networks: better weight consolidation and less catastrophic forgetting

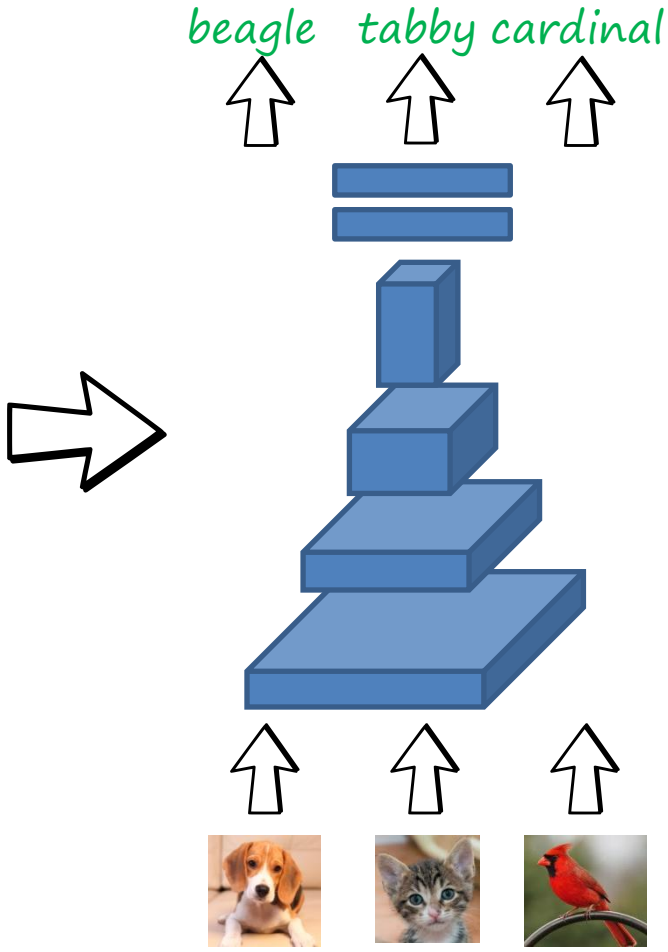
X. Liu, M. Masana, L. Herranz, J. van de Weijer, A. M. Lopez, A. D. Bagdanov
Computer Vision Center (Barcelona)
ICPR 2018

Lifelong learning

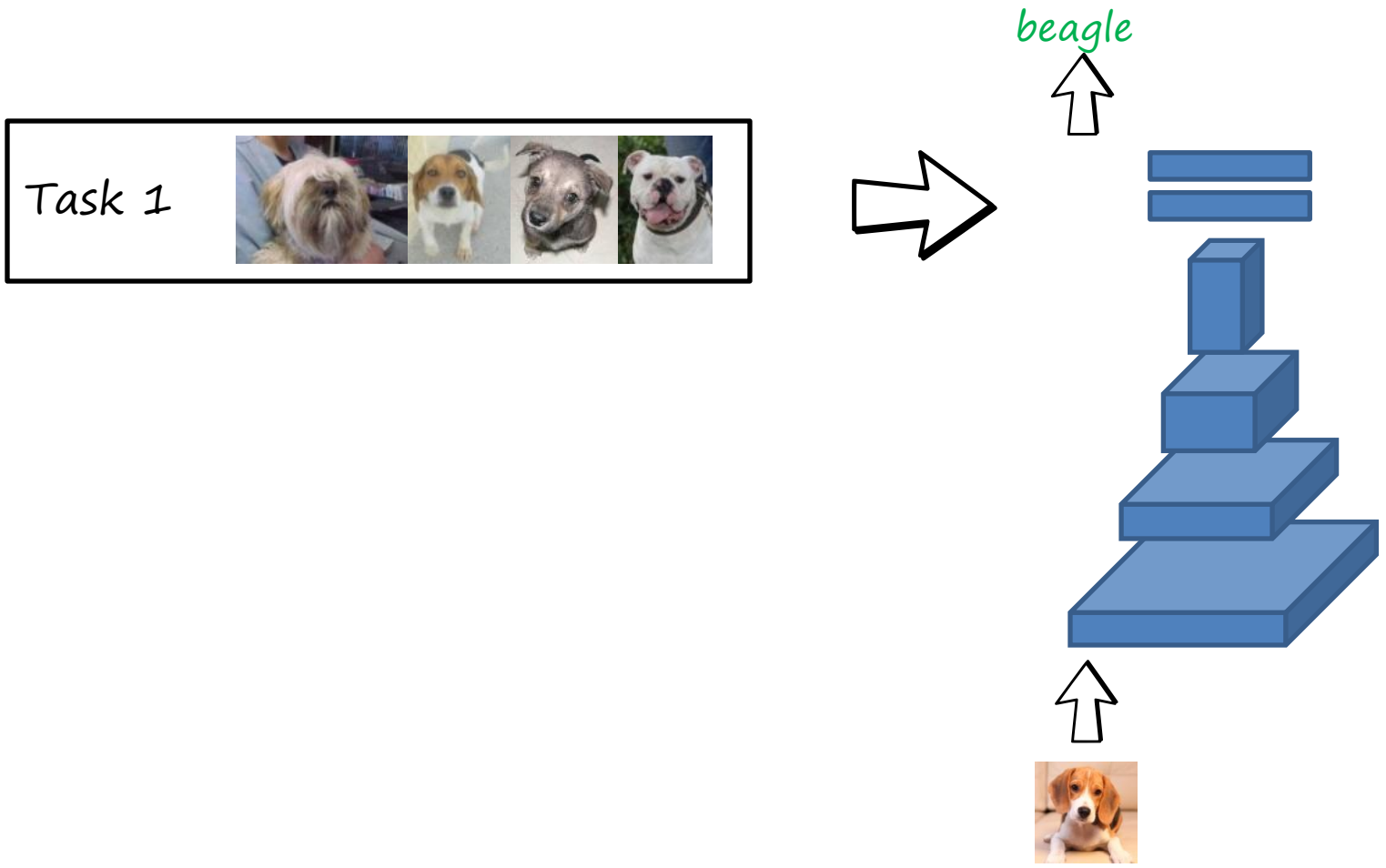


Joint learning

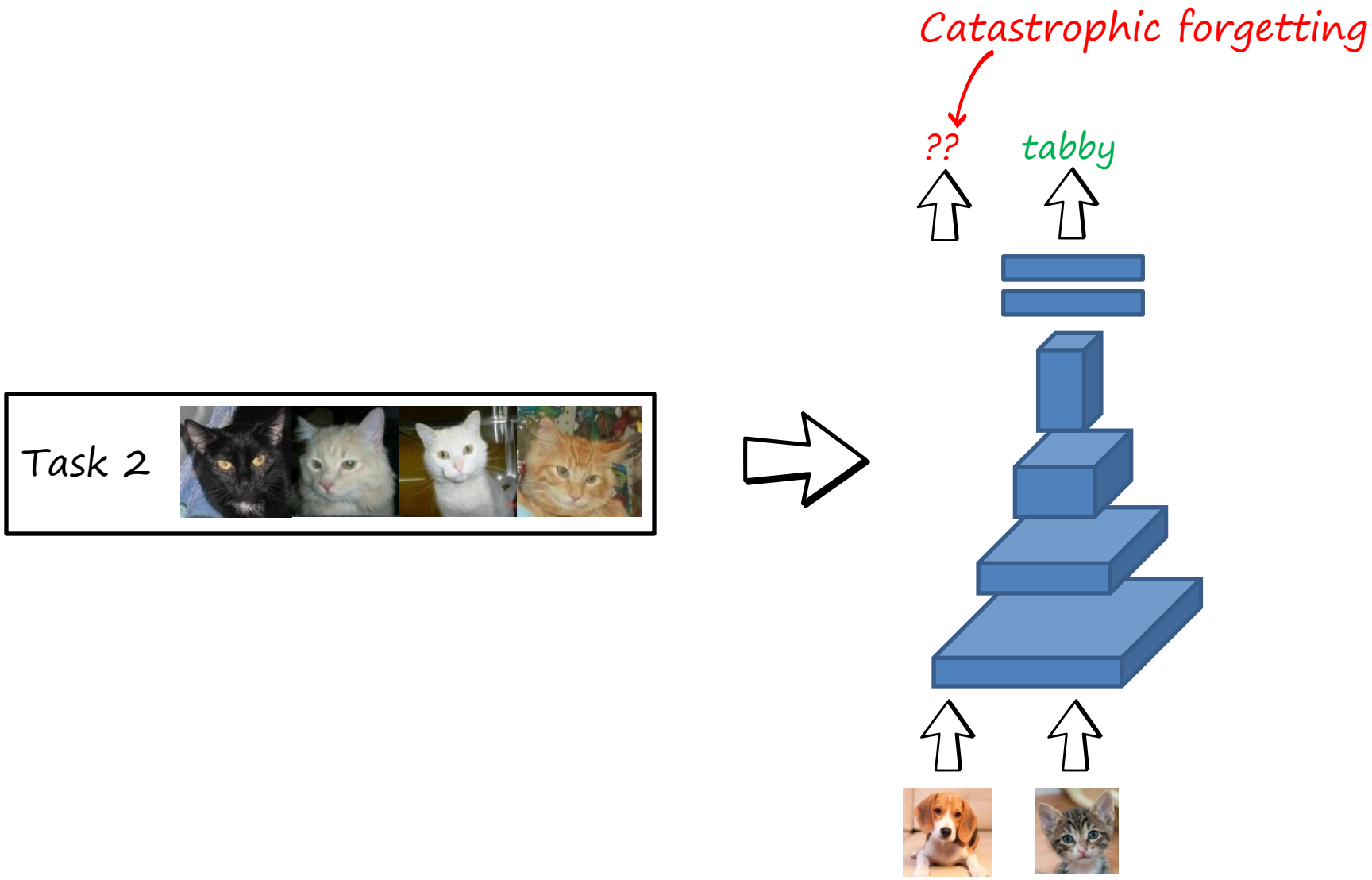




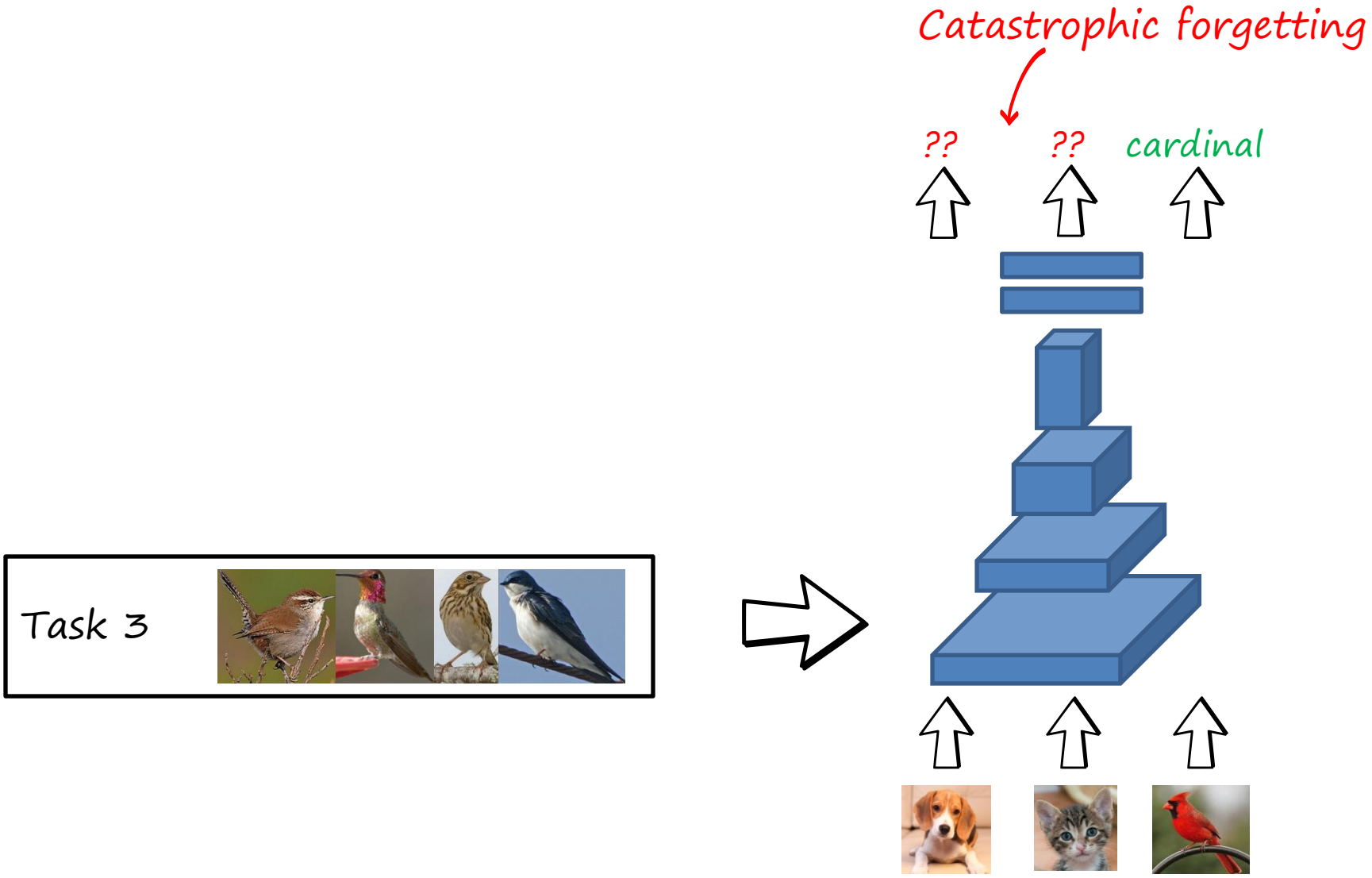
Sequential learning



Sequential learning



Sequential learning



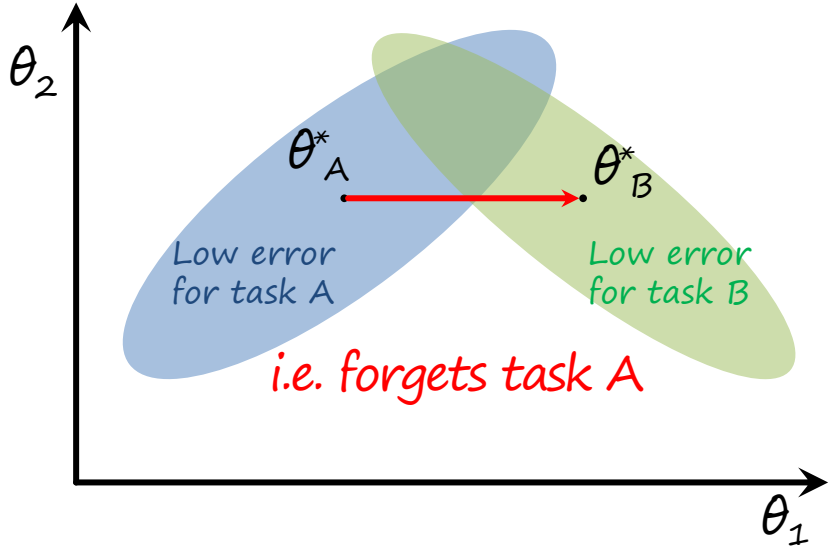
Preventing catastrophic forgetting

- Memories, replay and retrain
  - A few real samples (**exemplars**) from previous tasks
  - A generative **model** for previous data and sample
- Regularization
  - Weights
  - Activations
- Develop modularity
  - Enforce using different weights/neurons for different tasks

Elastic weight consolidation

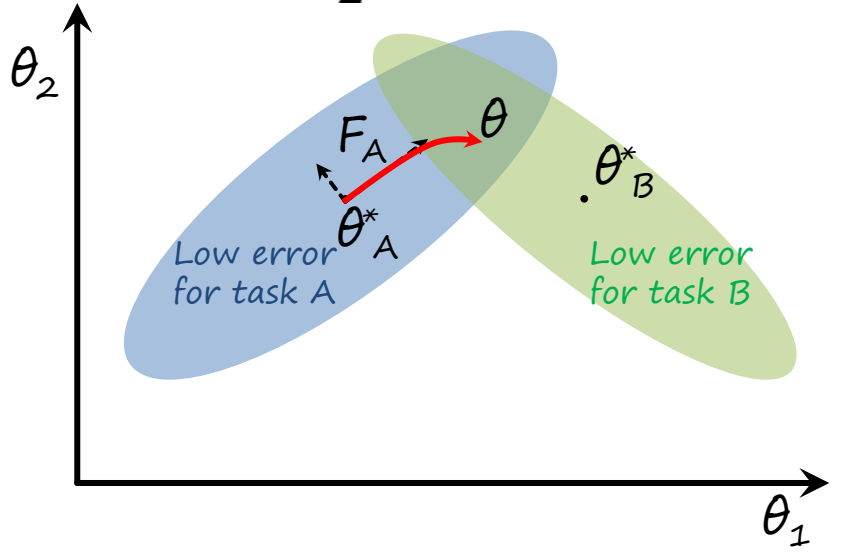
Fine tuning

$$\mathcal{L}(\theta) \approx \mathcal{L}_B(\theta)$$



Elastic weight consolidation (EWC)

$$\mathcal{L}(\theta) \approx \mathcal{L}_B(\theta) + \frac{\lambda}{2}\left(\theta - \theta^*_A\right)^\intercal F_A \left(\theta - \theta^*_A\right)$$



Problem: size of F_A is $(\#\theta)^2$

Kirkpatric et al., Overcoming catastrophic forgetting in neural networks, PNAS, 2017

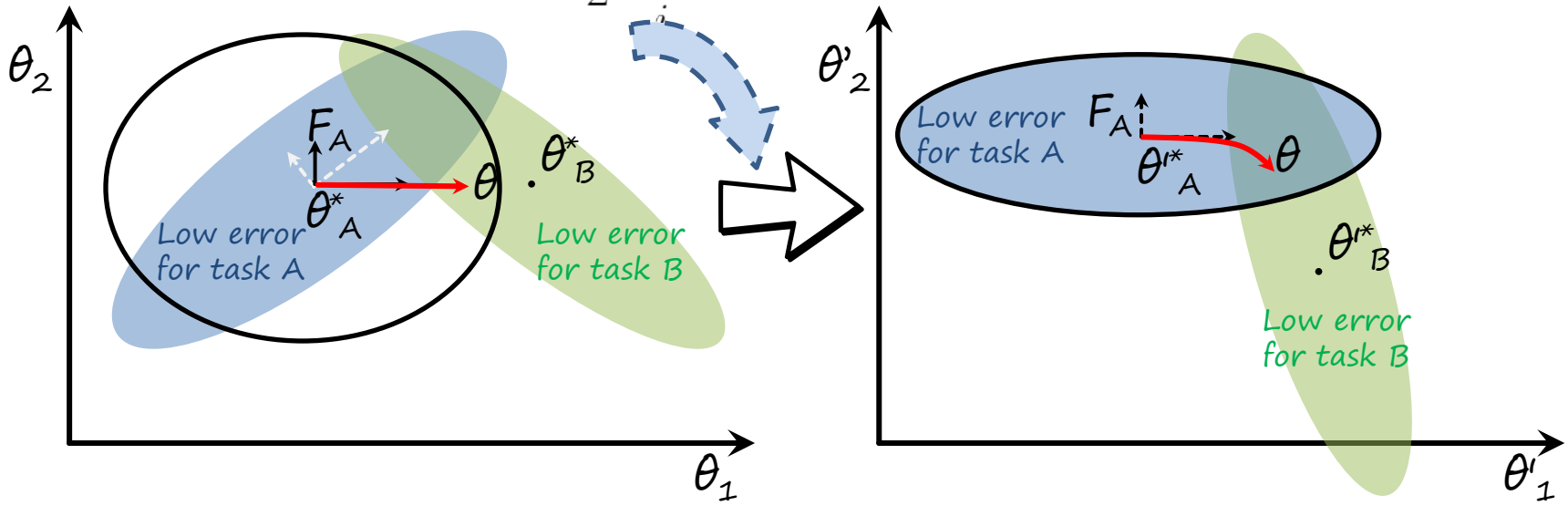
Rotated elastic weight consolidation

Size of the diagonal of F_A is #θ

Elastic weight consolidation in practice (with diagonal approx. of F_A)

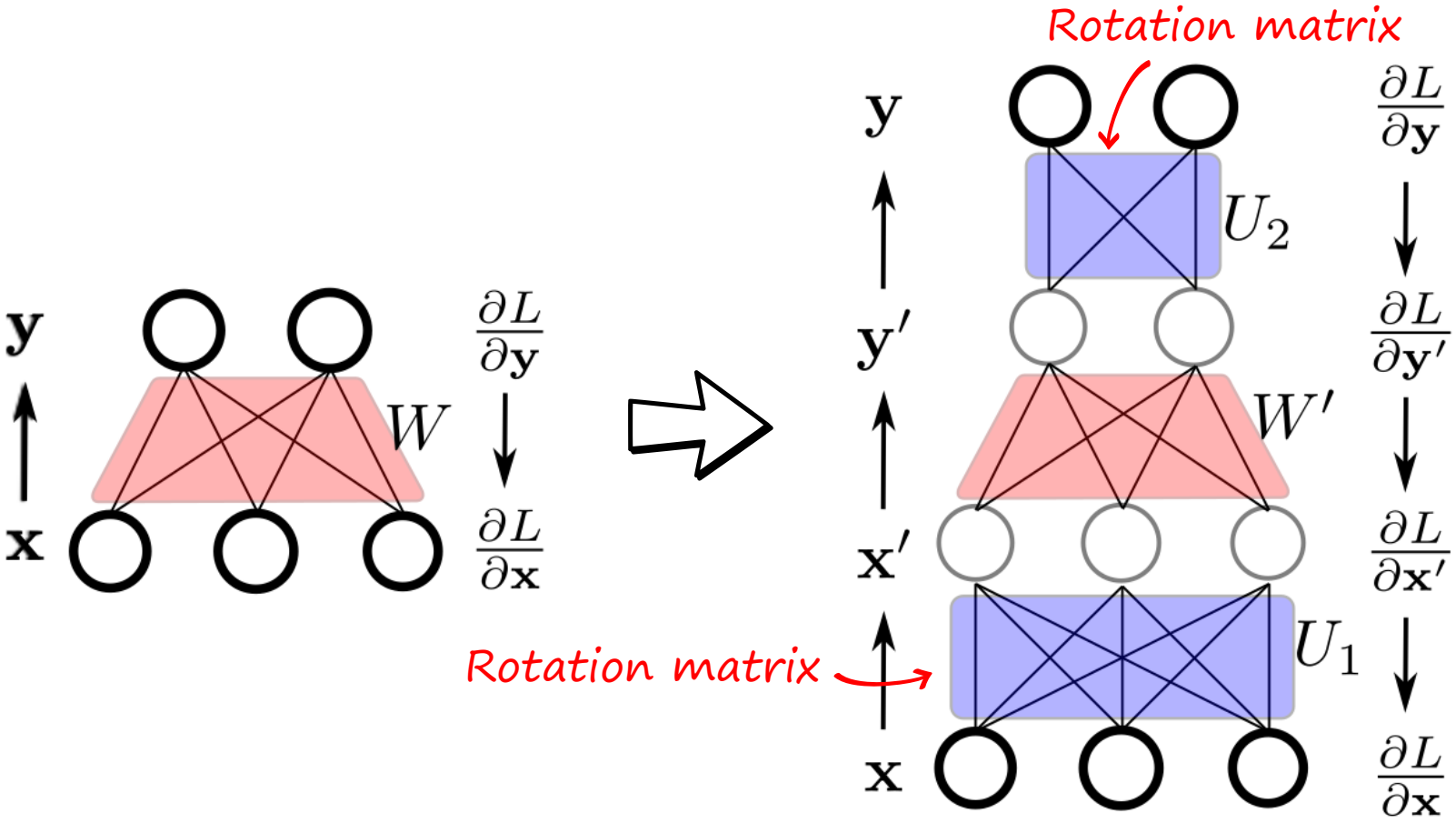
Rotated elastic weight consolidation (R-EWC)

$$\mathcal{L}(\theta) \approx \mathcal{L}_B(\theta) + \frac{\lambda}{2} \sum_i F_{A_i} \left(\theta - \theta^*_A\right)^2$$



Liu, Masana et al., Rotate your Networks: Better Weight Consolidation and Less Catastrophic Forgetting, ICPR 2018

Rotating fully connected layers



Liu , Masana et al., Rotate your Networks: Better Weight Consolidation and Less Catastrophic Forgetting, ICPR 2018

Computing the rotations

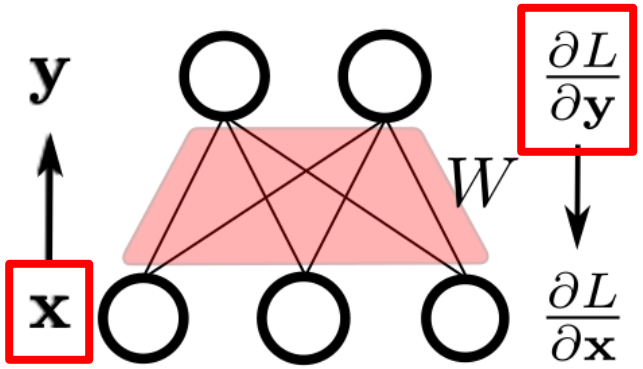
$$F_W = \mathbb{E}_{p\sim\pi}\left[\left(\frac{\partial L}{\partial \mathbf{y}}\right)\mathbf{x}\mathbf{x}^\intercal\left(\frac{\partial L}{\partial \mathbf{y}}\right)^\intercal\right]$$

$$F_W \approx \mathbb{E}_{\substack{x\sim\pi\\ y\sim p}}\left[\left(\frac{\partial L}{\partial \mathbf{y}}\right)\left(\frac{\partial L}{\partial \mathbf{y}}\right)^\intercal\right]\mathbb{E}_{x\sim\pi}\left[\mathbf{x}\mathbf{x}^\intercal\right]$$

$$\mathbb{E}_{x\sim\pi}\left[\mathbf{x}\mathbf{x}^\intercal\right] = U_1 S_1 V_1^\intercal$$

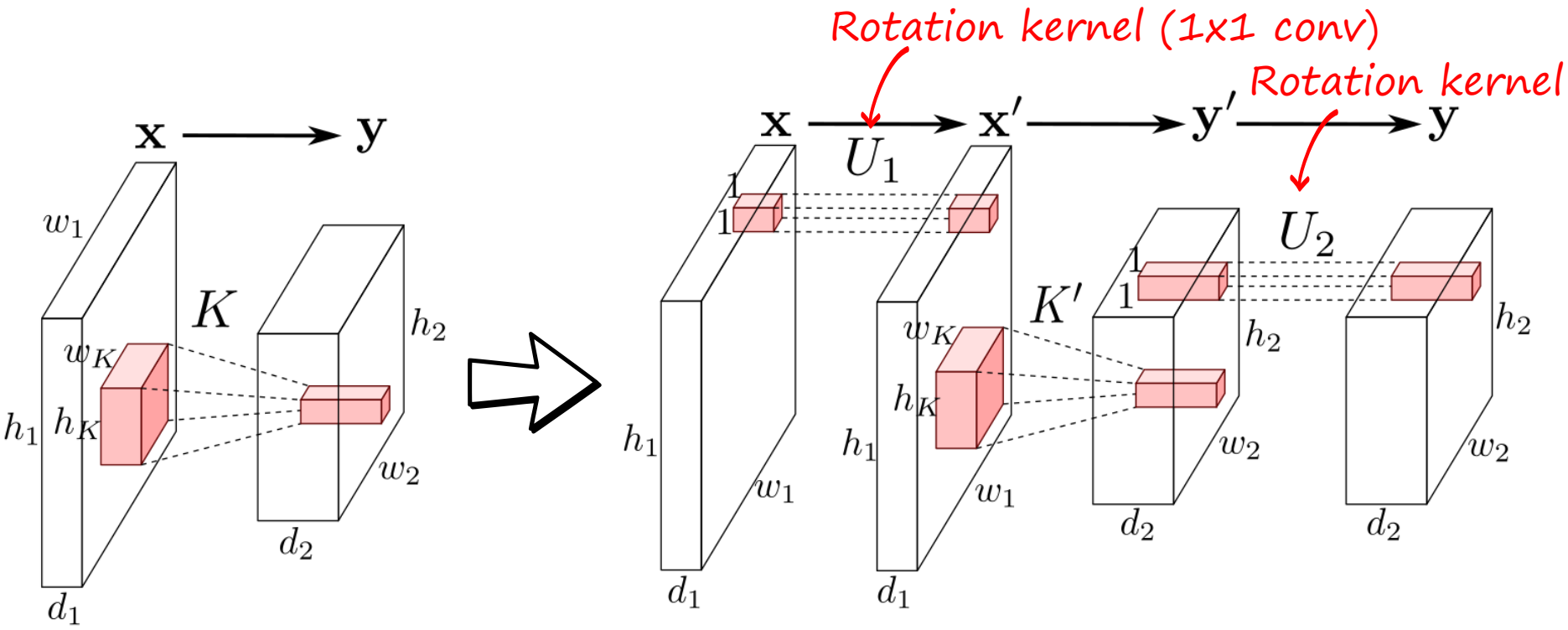
$$\mathbb{E}_{\substack{x\sim\pi\\ y\sim p}}\left[\left(\frac{\partial L}{\partial \mathbf{y}}\right)\left(\frac{\partial L}{\partial \mathbf{y}}\right)^\intercal\right] = U_2 S_2 V_2^\intercal$$

Using SVD



Liu , Masana et al., Rotate your Networks: Better Weight Consolidation and Less Catastrophic Forgetting, ICPR 2018

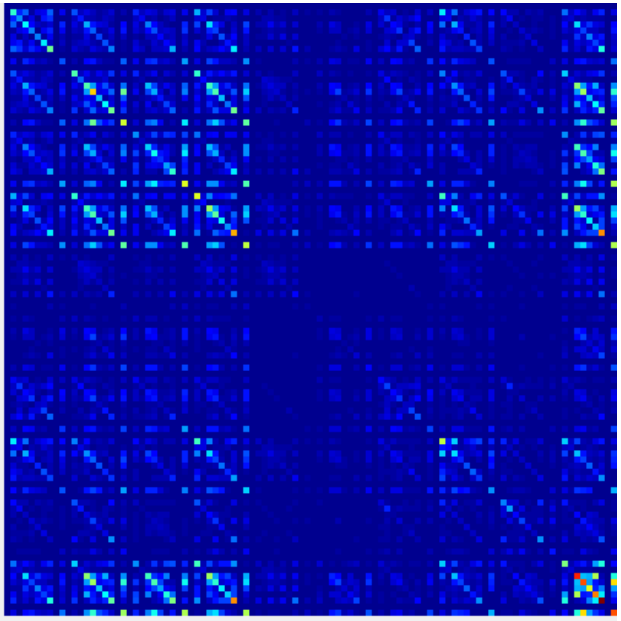
Rotating convolutional layers



Liu , Masana et al., Rotate your Networks: Better Weight Consolidation and Less Catastrophic Forgetting, ICPR 2018

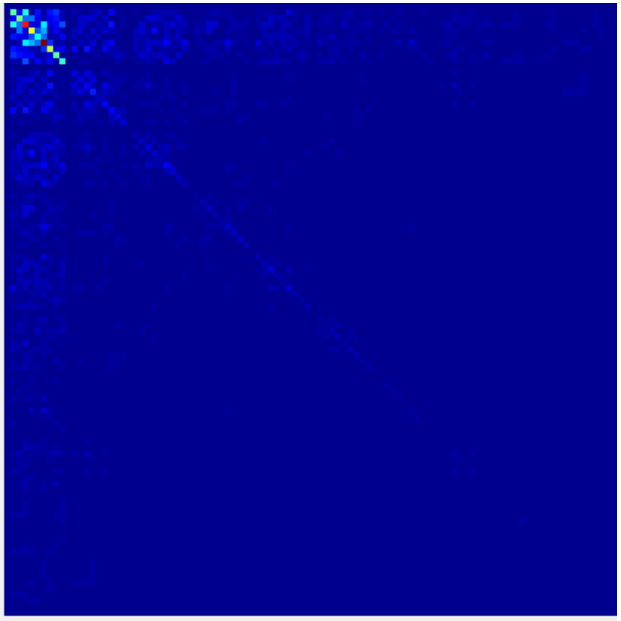
Fisher Information matrix

No rotation (i.e. EWC)



Energy in the diagonal: 40%

After rotation (i.e. R-EWC)



Energy in the diagonal: 74%

Liu , Masana et al., Rotate your Networks: Better Weight Consolidation and Less Catastrophic Forgetting, ICPR 2018

Experimental results

- MNIST dataset. Two tasks: 0-4 and 5-9

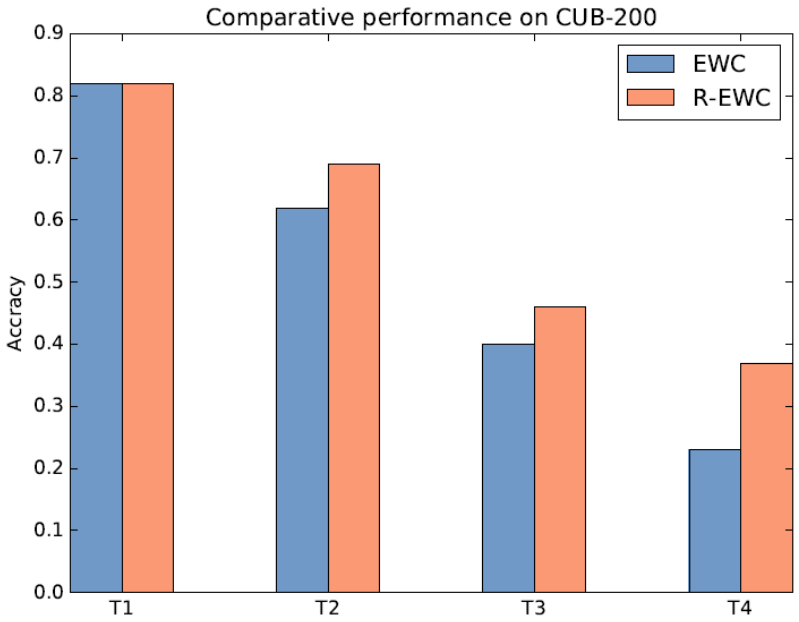
	$\lambda = 1$		$\lambda = 10$		$\lambda = 100$		$\lambda = 1000$		$\lambda = 10000$	
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
FT	6.1	97.6	6.1	97.6	6.1	97.6	6.1	97.6	6.1	97.6
EWC [5]	66.8	90.9	75.3	95.6	**85.8**	**92.8**	78.4	93.7	81.0	88.8
R-EWC - conv only	62.7	89.2	67.5	96.1	80.4	91.4	**84.7**	**93.1**	75.5	93.7
R-EWC - fc only	78.9	95.3	79.0	95.8	87.4	93.5	93.0	82.3	**94.3**	**88.0**
R-EWC - all	77.2	96.7	**91.7**	**91.2**	86.9	95.9	96.3	81.1	92.1	86.0
R-EWC - all no last	71.5	91.8	84.9	97.0	**91.6**	**94.5**	94.6	88.4	97.9	79.4

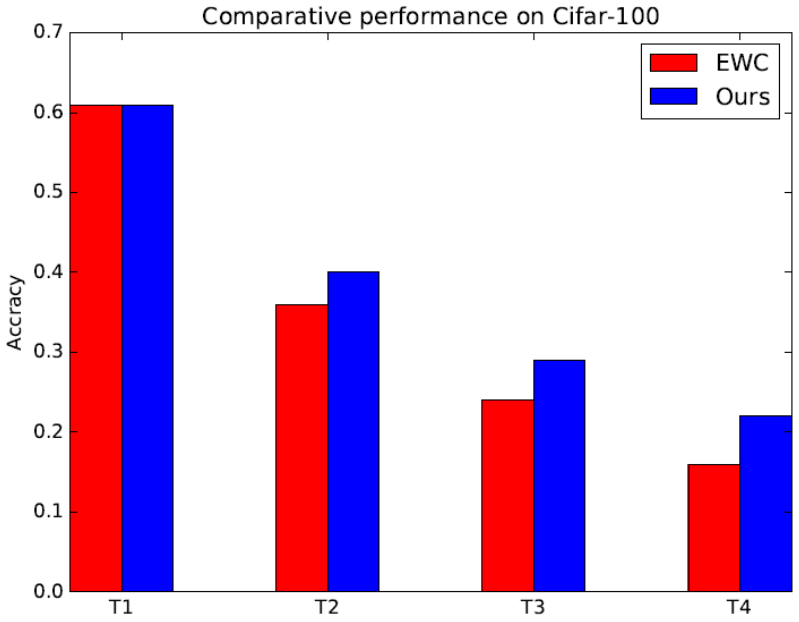
- Several datasets

	EWC [5] (T1 / T2)	R-EWC (T1 / T2)
MNIST	89.3 (85.8 / 92.8)	**93.1** (91.6 / 94.5)
CIFAR-100	37.5 (23.5 / 51.5)	**42.5** (30.2 / 54.7)
CUB-200 Birds	45.3 (42.3 / 48.6)	**48.4** (53.3 / 45.2)
Stanford-40 Actions	50.4 (44.3 / 58.4)	**52.5** (52.3 / 52.6)

Liu , Masana et al., Rotate your Networks: Better Weight Consolidation and Less Catastrophic Forgetting, ICPR 2018

Results on 4 tasks





CUB-200 Birds

CIFAR-100

Liu , Masana et al., Rotate your Networks: Better Weight Consolidation and Less Catastrophic Forgetting, ICPR 2018

THANK YOU!