南京航空航天大學

Nanjing University of Aeronautics and Astronautics

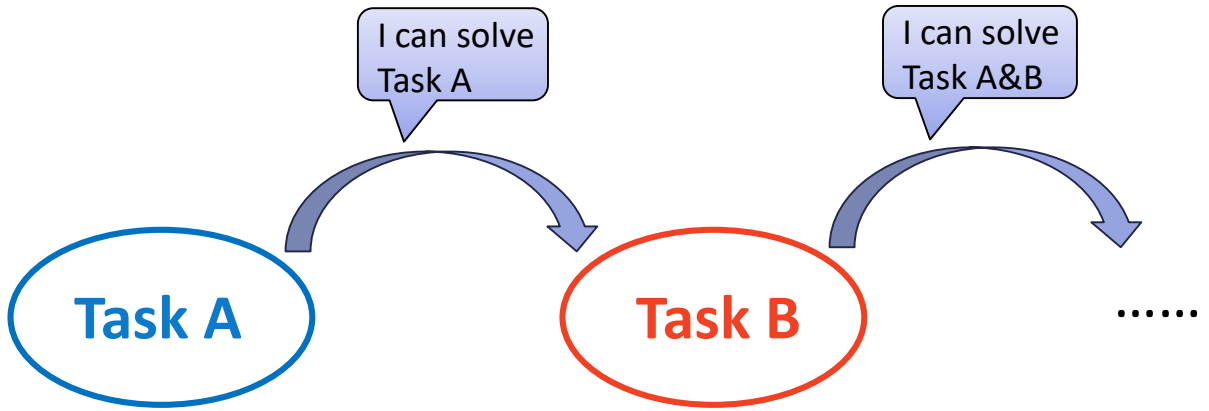
Two Papers about Continual Learning

2022-07-18

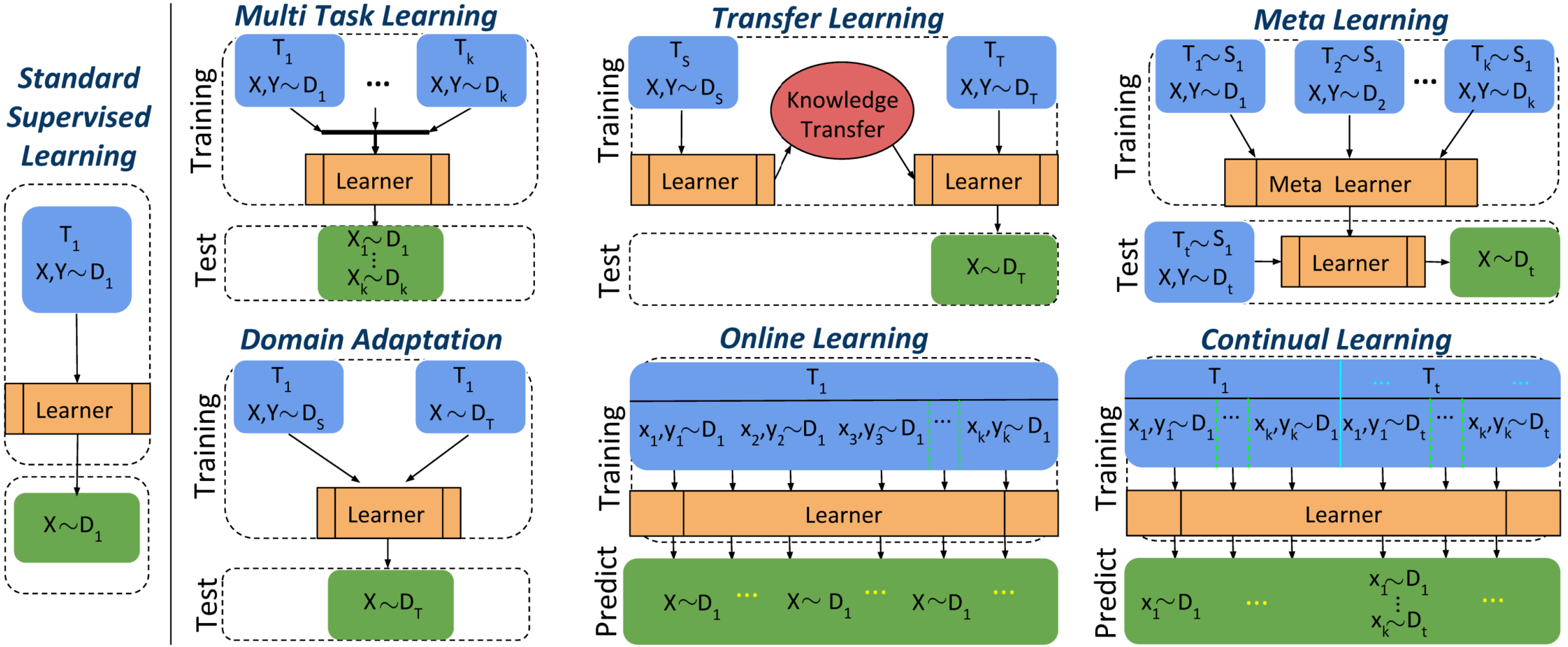
Continual Learning

☐ Need to be solved：

- **Catastrophic forgetting**
- **Stability-plasticity dilemma**

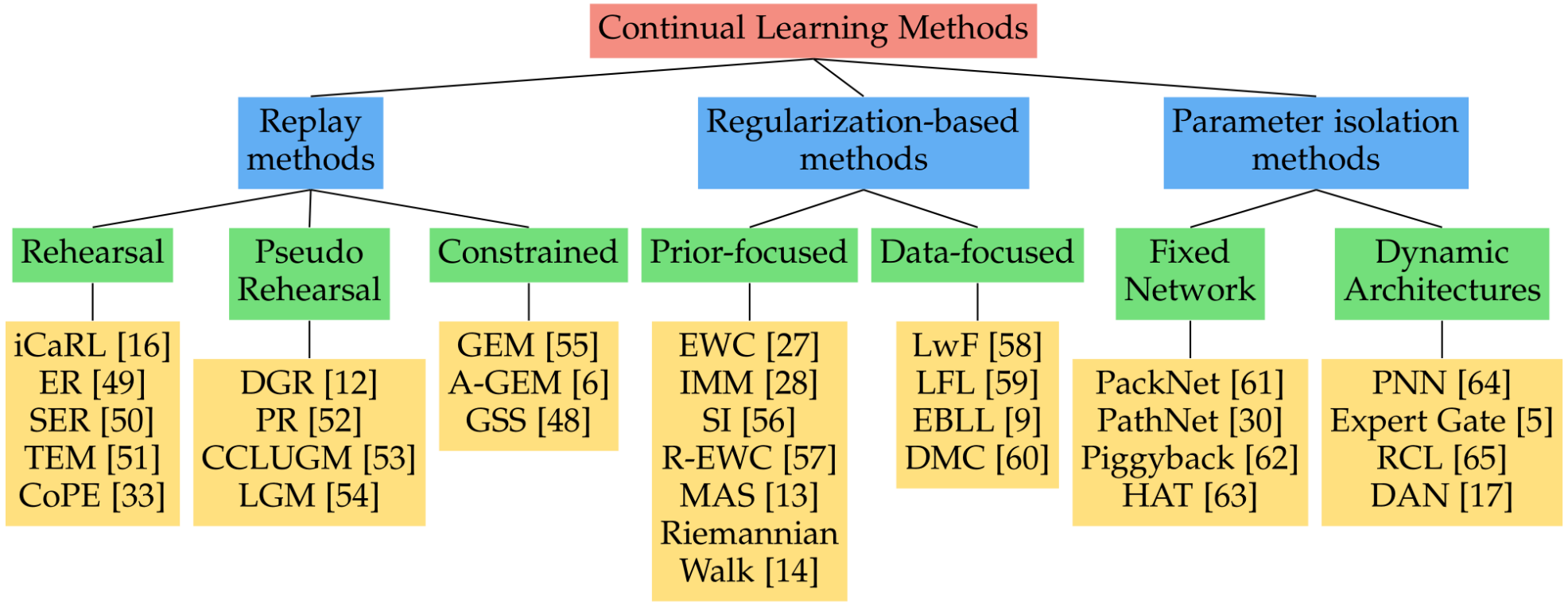


Different Learning Paradigms

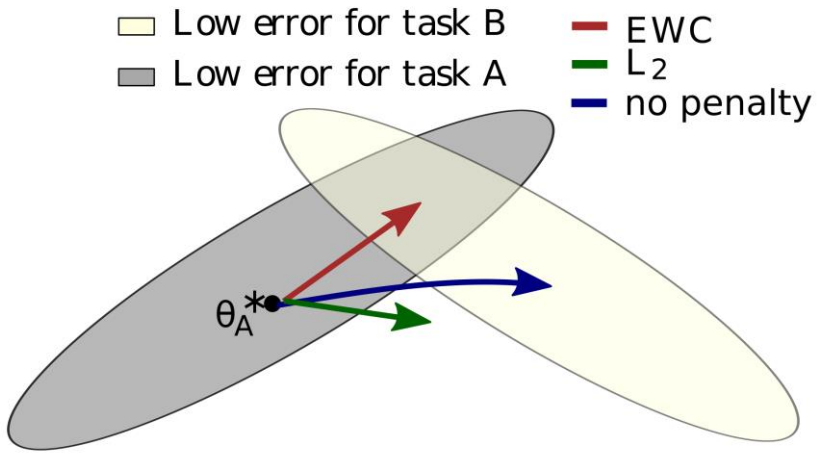


- Continual Learning
- Life long Learning (LLL)
- Incremental Learning

Continual Learning



Elastic weight consolidation(EWC)



$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}|\theta) + \log p(\theta) - \log p(\mathcal{D}) \tag{1}$$

$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}_B|\theta) + \log p(\theta|\mathcal{D}_A) - \log p(\mathcal{D}_B) \tag{2}$$

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2 \tag{3}$$

南京航空航天大學

Nanjing University of Aeronautics and Astronautics

Continual Learning Through Synaptic Intelligence

Friedemann Zenke [*1] **Ben Poole** [*1] **Surya Ganguli** [1]

ICML 2017

Synaptic Intelligence (SI)

$$\boldsymbol{g} = \frac{\partial L}{\partial \boldsymbol{\theta}}$$

$$L(\boldsymbol{\theta}(t) + \boldsymbol{\delta}(t)) - L(\boldsymbol{\theta}(t)) \approx \sum_k g_k(t)\delta_k(t)\ , \qquad (1)$$

$$\int_C \boldsymbol{g}(\boldsymbol{\theta}(t))d\boldsymbol{\theta} = \int_{t_0}^{t_1} \boldsymbol{g}(\boldsymbol{\theta}(t)) \cdot \boldsymbol{\theta}'(t)dt. \qquad (2)$$

$$\begin{aligned} \int_{t^{\mu-1}}^{t^{\mu}} \boldsymbol{g}(\boldsymbol{\theta}(t)) \cdot \boldsymbol{\theta}'(t)dt &= \sum_k \int_{t^{\mu-1}}^{t^{\mu}} g_k(\theta(t))\theta_k'(t)dt \\ &\equiv -\sum_k \omega_k^{\mu}. \end{aligned} \qquad (3)$$

$$\tilde{L}_\mu = L_\mu + \underbrace{c\sum_k \Omega_k^{\mu} \left(\tilde{\theta}_k - \theta_k\right)^2}_{\text{surrogate loss}} \qquad (4)$$

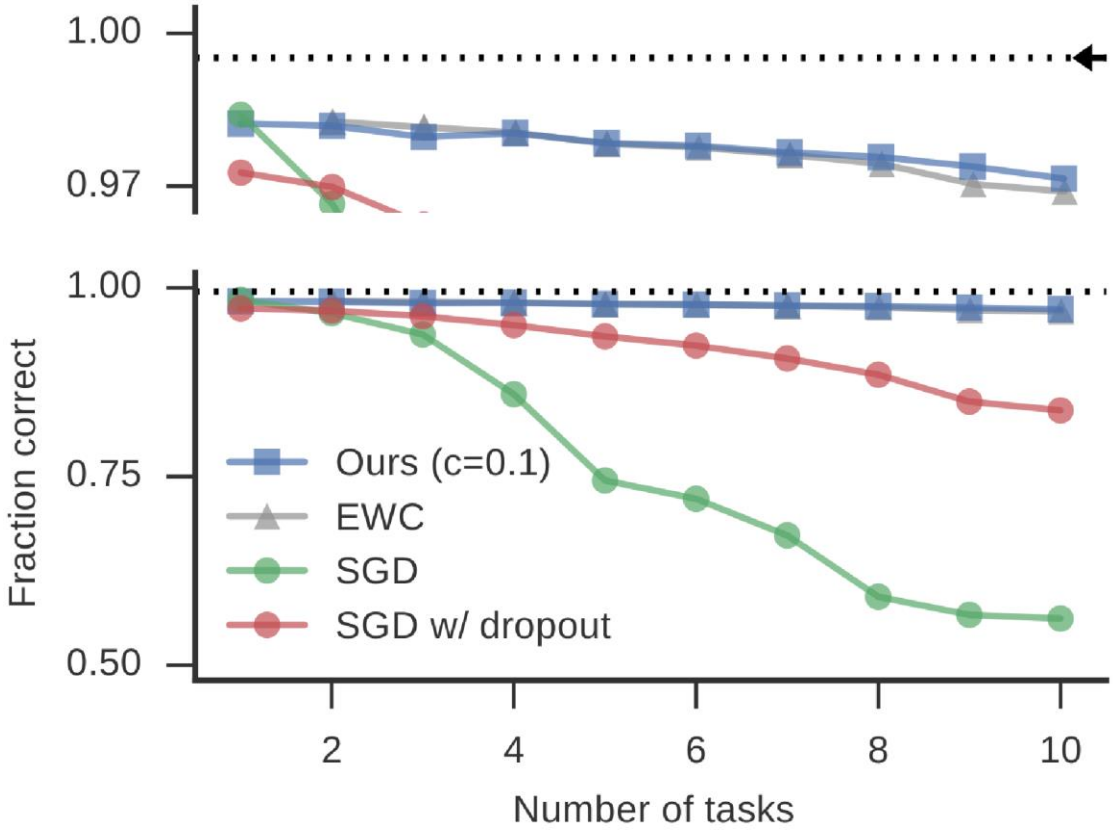
Importance measures

$$\Omega_k^{\mu} = \sum_{\nu<\mu} \frac{\omega_k^{\nu}}{(\Delta_k^{\nu})^2 + \xi}\ . \qquad (5)$$

$$\Delta_k^{\nu} \equiv \theta_k(t^{\nu}) - \theta_k(t^{\nu-1})$$

- *A **parameter importance** is proportional to its contribution to the loss decrease over time*

Experiments



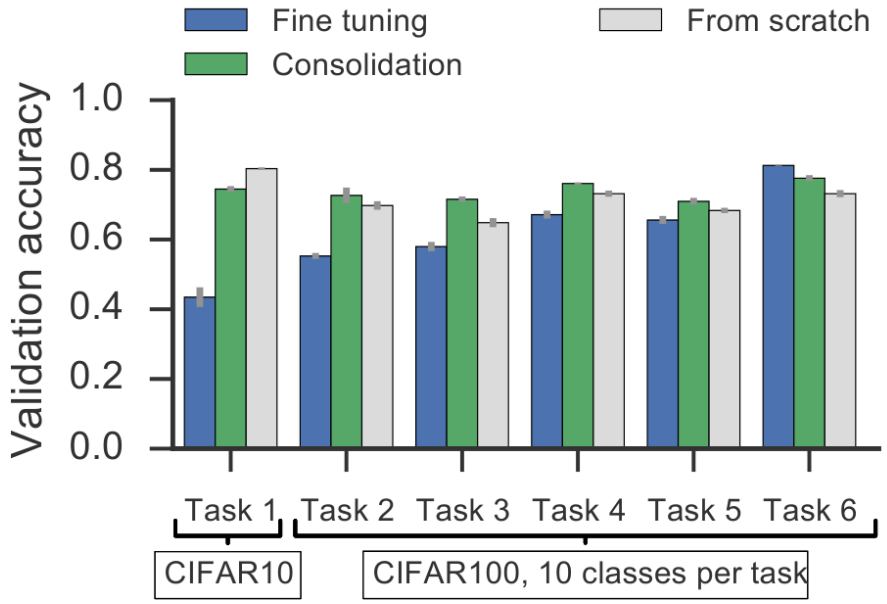


Figure 6. Validation accuracy on the split CIFAR-10/100 benchmark. Blue: Validation error, without consolidation ($c = 0$). Green: Validation error, with consolidation ($c = 0.1$). Gray: Network without consolidation trained from scratch on the single task only. Chance-level in this benchmark is 0.1. Error bars correspond to SD (n=5).

南京航空航天大學

Nanjing University of Aeronautics and Astronautics

Memory Aware Synapses: Learning what (not) to forget

Rahaf Aljundi[1], Francesca Babiloni[1], Mohamed Elhoseiny[2],
Marcus Rohrbach[2], and Tinne Tuytelaars[1]

[1] KU Leuven, ESAT-PSI, iMEC, Belgium
[2] Facebook AI Research

ECCV 2018

Memory Aware Synapses

Fig. 1: Our continuous learning setup. As common in the LLL literature, tasks are learned in sequence, one after the other. If, in between learning tasks, the agent is active and performs the learned tasks, we can use these unlabeled samples to update importance weights for the model parameters. Data that appears frequently, will have a bigger contribution. This way, the agent learns what is important and should not be forgotten.

Memory Aware Synapses (MAS)

Small perturbation

$$F(x_k;\theta+\delta)-F(x_k;\theta)\approx\sum_{i,j}g_{ij}(x_k)\delta_{ij} \tag{1}$$

$$g_{ij}(x_k)=\frac{\partial(F(x_k;\theta))}{\partial\theta_{ij}}$$

Importance weight: $\Omega_{ij}=\frac{1}{N}\sum_{k=1}^{N}\|\,g_{ij}(x_k)\,\|$ (2)

$$L(\theta)=L_n(\theta)+\lambda\sum_{i,j}\Omega_{ij}(\theta_{ij}-\theta^*_{ij})^2 \tag{3}$$

Comparison

Method	Type	Constant Memory	Problem agnostic	On Pre-trained	Unlabeled data	Adaptive
LwF [17]	data	✓	X	✓	n/a	X
EBLL [28]	data	X	X	X	n/a	X
EWC [12]	model	✓	✓	✓	X	X
IMM [16]	model	X	✓	✓	X	X
SI [39]	model	✓	✓	X	X	X
MAS (our)	model	✓	✓	✓	✓	✓

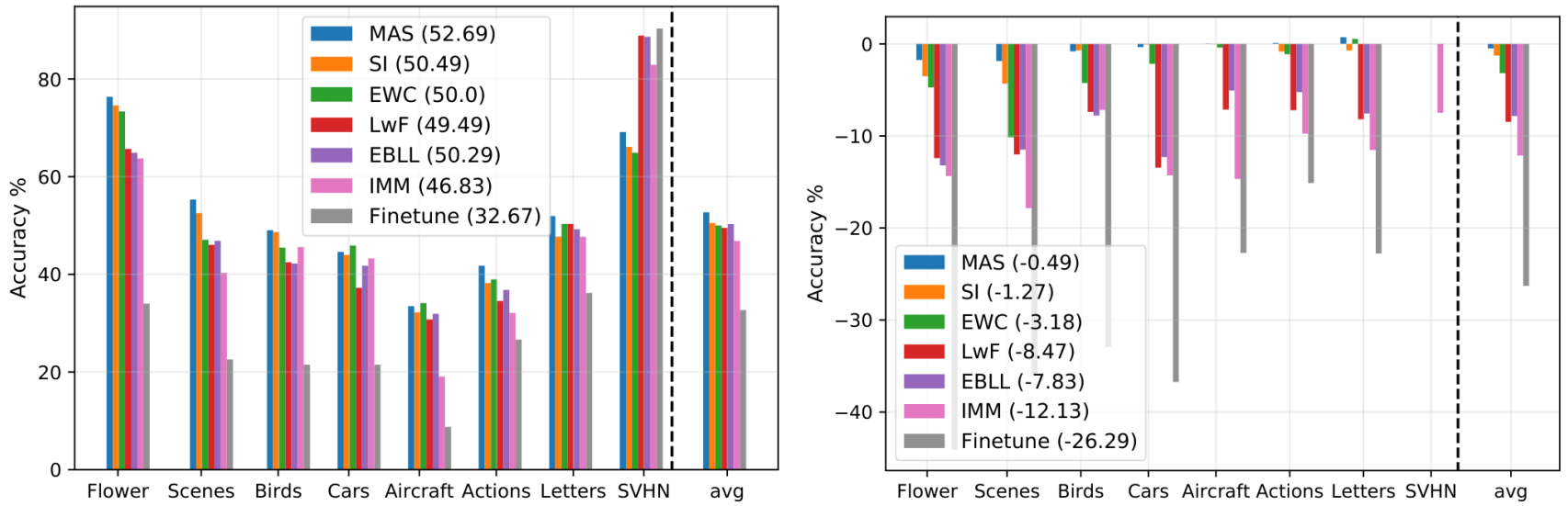
Table 1: LLL desired characteristics and the compliance of methods, that treat forgetting without storing the data, to these characteristics.

Experiments

Method	Birds → Scenes		Scenes → Birds		Flower → Birds		Flower → Scenes	
FineTune	45.20 (-8.0)	**57.8**	49.7 (-9.3)	**52.8**	64.87 (-13.2)	**53.8**	70.17 (-7.9)	57.31
LwF [17]	51.65 (-2.0)	55.59	55.89 (-3.1)	49.46	73.97 (-4.1)	53.64	76.20 (-1.9)	58.05
EBLL [28]	52.79 (-0.8)	55.67	56.34 (-2.7)	49.41	75.45 (-2.6)	50.51	76.20 (-1.9)	**58.35**
IMM [16]	51.51 (-2.1)	52.62	54.76 (-4.2)	52.20	75.68 (-2.4)	48.32	76.28 (-1.8)	55.64
EWC [12]	52.19 (-1.4)	55.74	**58.28 (-0.8)**	49.65	76.46 (-1.6)	50.7	77.0 (-1.1)	57.53
SI [39]	52.64 (-1.0)	55.89	57.46 (-1.5)	49.70	75.19 (-2.9)	51.20	76.61 (-1.5)	57.53
MAS (ours)	**53.24 (-0.4)**	55.0	57.61 (-1.4)	49.62	**77.33 (-0.7)**	50.39	**77.24 (-0.8)**	57.38

Table 2: Classification accuracy (%), drop in first task (%) for various sequences of 2 tasks using the object recognition setup.

Experiments



(a) (b)

Fig. 5: 5a performance on each task, in accuracy, at the end of 8 tasks object recognition sequence. 5b drop in each task relative to the performance achieved after training each task.

significant difference on forgetting over 3 random trials where we get a mean, over 6 numbers, of $0.51\% \pm 0.18$ for the drop on the first task in the vector output case compared to $0.50\% \pm 0.19$ for the ℓ_2^2 norm case. No significant difference is observed on the second task either. As such, using ℓ_2^2 is n times faster (where n is the length of the output vector) without loss in performance.

THANKS