Don't forget, there is more than forgetting: new metrics for Continual Learning

Natalia Díaz-Rodríguez[1*], Vincenzo Lomonaco[2*], Davide Maltoni[2] and David Filliat[1],

(1) U2IS, ENSTA ParisTech, Inria FLOWERS team, Palaiseau, France. https://flowers.inria.fr/ {natalia.diaz, david.filliat}@ensta-paristech.fr.
(2) University of Bologna, Italy. {vincenzo.lomonaco, davide.maltoni}@unibo.it http://continualai.org/
(*) Both authors contributed equally to this work.

Motivation

The lack of consensus in evaluating continual learning (CL) algorithms and the almost exclusive focus on catastrophic forgetting motivate us to propose a more comprehensive set of implementation independent metrics accounting for factors we believe have practical implications worth considering with respect to "static" machine learning settings.

New metrics for Continual Learning (CL)

A CL algorithm A^{CL} is an algorithm with the following signature:

$$\forall D_i \in \mathcal{D}, \quad A_i^{CL} : \ < h_{i-1}, Tr_i, M_{i-1}, t > \rightarrow < h_i, M_i > \tag{1}$$

Where h_i is the model, Tr_i is the training set of examples drawn from the respective D_i distribution, M_i is an external memory where we can store previous training examples and t is a task label. For simplicity, we can assume N as the number of tasks, one for each Tr_i.

- **Accuracy (A)**: Given matrix $R \in \mathbb{R}^{N \times N}$, which contains in each entry $R_{i,j}$ the test classification accuracy of the model on task t_j after observing the last sample from task t_i [3], Accuracy metric is:

$$A = \frac{\sum_{i \geq j}^{N} R_{i,j}}{\frac{N(N+1)}{2}} \tag{2}$$

Table 1: Accuracy matrix R: elements accounted to compute A (white & cyan), BWT (cyan), and FWT (gray) Tr_i = training, Te_i= test tasks.

R	Te_1	Te_2	Te_3
Tr_1	R^*	R_{ij}	R_{ij}
Tr_2	R_{ij}	R^*	R_{ij}
Tr_3	R_{ij}	R_{ij}	R^*

Backward Transfer (BWT) measures the influence that learning a task has on the performance on previous tasks [3].

$$BWT = \frac{\sum_{i=2}^{N} \sum_{j=1}^{i-1} (R_{i,j} - R_{j,j})}{\frac{N(N-1)}{2}} \tag{3}$$

- **Remembering (REM)**:

$$REM = 1 - |min(BWT, 0)| \tag{4}$$

is the originally negative BWT while the originally positive BWT, i.e., improvement over time, is:

- **Positive Backward Transfer (BWT^+)**:

$$BWT^+ = max(BWT, 0) \tag{5}$$

- **Forward Transfer (FWT)**: measures the influence that learning a task has on the performance of future tasks [3].

$$FWT = \frac{\sum_{i<j}^{N} R_{i,j}}{\frac{N(N-1)}{2}} \tag{6}$$

- **Model size efficiency (MS)**: The memory size of model h_i quantified in terms of parameters θ at each task i, $Mem(\theta_i)$, should not grow too rapidly with respect to the size of the model that learned the first task, $Mem(\theta_1)$.

$$MS = min(1, \frac{\sum_{i=1}^{N} \frac{Mem(\theta_1)}{Mem(\theta_i)}}{N}) \tag{7}$$

- **Samples storage size efficiency (SSS)**: The memory occupation in bits by the samples storage memory M, $Mem(M)$, should be bounded by the memory occupation of the total nr of examples encountered at the end of the last task (D is the lifetime dataset associated to all distributions $\mathcal{D}$):

$$SSS = 1 - min(1, \frac{\sum_{i=1}^{N} \frac{Mem(M_i)}{Mem(D)}}{N}) \tag{8}$$

- **Computational efficiency (CE)**: it is bounded by the nr of operations for training set Tr_i. $Ops(Tr_i)$ is the number of (mul-adds) operations needed to learn Tr_i and $Ops \uparrow\downarrow(Tr_i)$ is the nr of operations required to do one forward and one backward (backprop) pass on Tr_i).

$$CE = min(1, \frac{\sum_{i=1}^{N} \frac{Ops\uparrow\downarrow(Tr_i) \cdot \varepsilon}{1 + Ops(Tr_i)}}{N}) \tag{9}$$

CL_{score} and $CL_{stability}$ aggregating metrics

- CL_{score}: if $c_i \in [0, 1]$) is the avg. of r runs assigned a weight $w_i \in [0, 1]$ s.t. $\sum_i^{\mathcal{C}} w_i = 1$:

$$CL_{score} = \sum_{i=1}^{\#\mathcal{C}} w_i c_i \tag{10}$$

- $CL_{stability}$: the average of the std. devs. from all previous criteria c_i:

$$CL_{stability} = 1 - \sum_{i=1}^{\#\mathcal{C}} w_i stddev(c_i) \tag{11}$$

Experiments and Conclusion

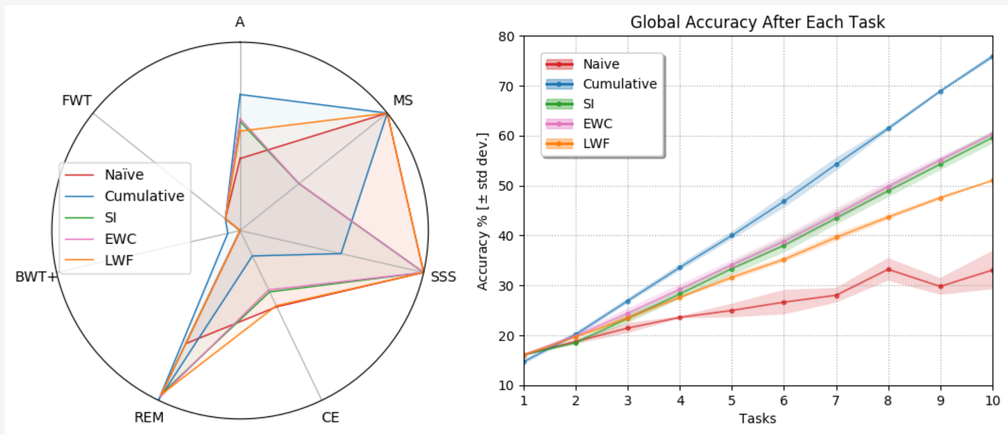

Figure 1: a) Spider chart: CL metrics per strategy (larger area is better). b) Accuracy per CL strategy computed over the fixed test set as proposed in [4].

We evaluate the CL metrics on cumulative and naïve baseline strategies as in [4], Elastic Weight Consolidation (EWC) [1], Synaptic Intelligence (SI) [5] and Learning without Forgetting (LwF) [2] on iCIFAR-100 dataset.

Table 2: CL metrics and CL_{score} for each CL strategy evaluated (higher is better).

Str.	A	REM	BWT^+	FWT	MS	SSS	CE	CL_{sco}	CL_{stab}
Naï	0.3825	0.6664	0.00	0.10	**1.00**	**1.00**	**0.4492**	0.5140	**0.9986**
Cum	**0.7225**	**1.00**	**0.0673**	0.10	**1.00**	0.55	0.1496	0.5128	0.9979
EWC	0.5940	0.9821	0.00	0.10	0.40	**1.00**	0.3495	0.4894	0.9972
LWF	0.5278	0.9667	0.00	0.10	**1.00**	**1.00**	0.4429	**0.5768**	**0.9986**
SI	0.5795	0.9620	0.00	0.10	0.40	**1.00**	0.3613	0.4861	0.9970

- **Experiments**: 3 weight configs W=[w_A, w_{MS}, w_{SSS}, w_{CE}, w_{BWT^+}, w_{REM}, w_{FWT}] (here $w_i = \frac{1}{\#\mathcal{C}}$, CNN in [5, 4]).
- **Future work**: provide insights that assess the importance of different metric schemes and their entanglement, and how to use these metrics wisely to assist choosing among algorithms.

References

[1] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell.
Overcoming catastrophic forgetting in neural networks.
Proceedings of the National Academy of Sciences, 114(13):3521–3526, 2017.

[2] Z. Li and D. Hoiem.
Learning without forgetting.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(12):2935–2947, Dec 2018.

[3] D. Lopez-Paz et al.
Gradient episodic memory for continual learning.
In *Advances in Neural Information Processing Systems*, pages 6467–6476, 2017.

[4] D. Maltoni and V. Lomonaco.
Continuous learning in single-incremental-task scenarios.
arXiv preprint arXiv:1806.08568, 2018.

[5] F. Zenke, B. Poole, and S. Ganguli.
Continual learning through synaptic intelligence.
In *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 3987–3995. PMLR, 2017.