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) ENSTA ParisTech, Inria FLOWERS team, France. https://flowers.inria.fr/
http://asr.ensta-paristech.fr/ {natalia.diaz,david.filliat}@ensta-paristech.fr.
(2) University of Bologna, Italy. {vincenzo.lomonaco,davide.maltoni}@unibo.it (*) Equal contribution.

2nd Continual Learning Workshop, NeurIPS. 7th Dec. 2018, Montreal, Canada
https://continualAI.org

Outline

- Continual Learning
- Motivation
- Continual Learning Framework
- New Metrics for Continual Learning
- Experiments
- Future Work (WIP)

Continual Learning
(CL)

Continual Learning Algorithms:

- learn from a stream of data/tasks
- continuously and adaptively thought time
- enable the incremental development of ever more complex knowledge and skills.

Motivation

Motivation:

- The lack of consensus in evaluating CL algorithms
- Almost exclusive focus on catastrophic forgetting[1]

We propose: Comprehensive, implementation independent metrics accounting for factors we believe have practical implications worth considering w.r.t.:

- "Non-static" ML settings
- Deployment of real AI systems that learn continually

[1] [McCloskey and Cohen, 1989, French, 1999]

Catastrophic forgetting[3]

- *The well-known phenomenon of a neural network experiencing a rapid overriding of previously learned knowledge when trained sequentially on new data.*
- An important objective quantified for assessing the quality of CL approaches[2].

[2] [Serrà et al., 2018, Lopez-Paz and Ranzato, 2017, Hayes et al., 2018, Farquhar and Gal, 2018]
[3] [McCloskey and Cohen, 1989, French, 1999]

Continual Learning Framework

Continual Learning Framework

In Continual Learning,

- $\mathcal{D} = \{D_1, \ldots, D_N\}$: a potentially infinite sequence of unknown distributions over $X \times Y$ encountered over time
- X and Y input and output r.v.
- h^*: general target function (i.e. our ideal prediction model)
- Task T: defined by a unique task label t and its target function $g^*_{\hat{t}}(x) \equiv h^*(x, t = \hat{t})$ (i.e., the objective of its learning).

A CL algorithm A^{CL} has the signature:

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

where

- h_i: the model
- Tr_i: training set of examples drawn from distribution D_i
- M_i: external memory (can store previous training examples)
- N: nr of tasks (one per Tr_i).

Continual Learning Metrics

Accuracy (A)[5]

Originally assessed the model performance at the end of the last task[4]; we extend A to account for performance at *every timestep in time*:

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

where $R_{i,j}$ in Accuracy matrix $R \in \mathbb{R}^{N \times N}$ is test classification accuracy on task t_j after observing the last sample from task t_i.

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^*

[4] [Lopez-Paz and Ranzato, 2017]

[5] Accuracy matrix R: elements accounted to compute A (white & cyan), BWT (cyan), and FWT (gray). $R^* = R_{ii}$, Tr_i = training, Te_i= test tasks.

Backward Transfer (BWT)[7]

BWT measures the influence that learning a task has on the performance on previous tasks[6].

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

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^*

[6] [Lopez-Paz and Ranzato, 2017]
[7] Accuracy matrix R: elements accounted to compute A (white & cyan), BWT (cyan), and FWT (gray). $R^* = R_{ii}$, Tr_i = training, Te_i= test tasks.

Backward Transfer (BWT^+) and Remembering (REM)

BWT is broken into two different clipped terms: (originally negative BWT, forgetting), **Remembering**:

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

and (originally positive BWT) improvement over time: **Positive Backward Transfer (BWT^+)**:

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

Forward Transfer (FWT)[9]

Measures the influence that learning a task has on the performance of future tasks[8]:

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

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^*

FWT can occur when the model is able to perform *zero-shot* learning.

[8] [Lopez-Paz and Ranzato, 2017]

[9] Accuracy matrix R: elements accounted to compute A (white & cyan), BWT (cyan), and FWT (gray). $R^* = R_{ii}$, Tr_i = training, Te_i = test tasks.

Model Size (MS) efficiency

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 w.r.t. the size of the model that learned the first task, $Mem(\theta_1)$:

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

Samples Storage Size (SSS) efficiency

The memory occupation in bits by the samples storage memory M, $Mem(M)$, should be bounded by the occupation of the total nr of examples encountered at the end of last task:

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

where D is the lifetime dataset associated to all distributions $\mathcal{D}$.

Computational Efficiency (CE)

CE is bounded by the nr of operations for training set Tr_i:

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

where

- $Ops(Tr_i)$: nr (mul-adds) operations needed to learn Tr_i
- $Ops\uparrow\downarrow(Tr_i)$: operations required to do one forward and one backward (backprop) pass on Tr_i
- ε: a scaling factor[10]

[10] Associated to the nr of epochs needed to learn Tr_i: when $Ops\uparrow\downarrow(Tr_i)$ is negligible w.r.t. $Ops(Tr_i)$, a $\varepsilon > 1$ makes CE more interpretable (here $\varepsilon = 10$).

CL_{score}

We fuse[11] these metrics into a single score:

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

where

- $c_i \in [0,1]$: avg. of r runs of c_i, assigned a weight $w_i \in [0,1]$ s.t. $\sum_i^{\mathcal{C}} w_i = 1$
- As each c_i, the final CL_{score}:
  - $\in [0,1]$
  - is to be maximized
  - can rank CL strategies

[11] Drawing inspiration from the standard Multi-Attribute Value Theory (MAVT)[Ishizaka and Nemery, 2013, Keeney and Raiffa, 1993]

The avg. of the std. deviations from all previous criteria c_i:

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

- $c_i \in [0, 1]$: avg. of r runs assigned a weight $w_i \in [0, 1]$ s.t. $\sum_i^{\mathcal{C}} w_i = 1$
- σ_{c_i}: std. deviation of criterion c_i

Experiments

Experiments: Dataset and Baselines

Dataset: iCIFAR-100: each of the 10 tasks: a training batch of 10 disjoint classes

Baselines:

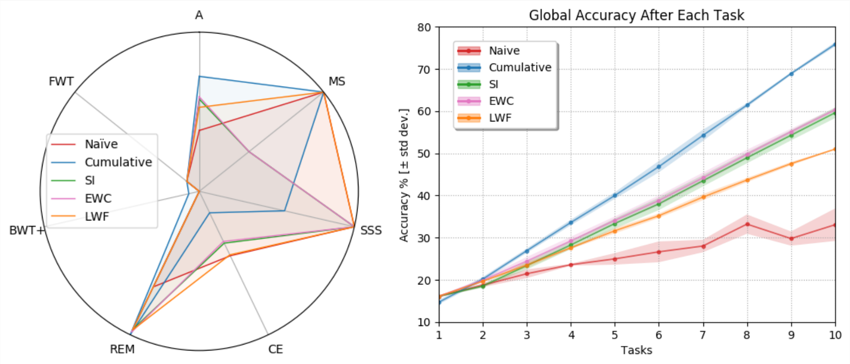
- Naïve strategy (Lower bound): starts at Tr_1 and learns continuously the coming training sets $Tr_2, ..., Tr_N$ simply tuning the model across batches[12].
- Cumulative strategy (Upper bound): starts from scratch every time, learning from the accumulation of $Tr_1, ..., Tr_{i-1}, Tr_i$ retrained with the patterns from the current batch and all previous batches[13].

CL strategies[14]:

- Elastic Weight Consolidation (EWC)
- Synaptic Intelligence (SI)
- Learning without Forgetting (LwF)

[12] Without any specific mechanism to control forgetting, except early stopping.
[13] Only in this approach we assume all previous data can be stored and reused.
[14] EWC [Kirkpatrick et al., 2016], SI [Zenke et al., 2017], LwF [Li and Hoiem, 2016].

Experiments: Accuracy per CL strategy computed over the fixed test set



- The larger the area under the CL algorithm curve,
  → the highest (more optimal) CL_{score} is.
- The farther away from the cumulative (blue) surface,
  → the larger room for improvement

Experiments[15]

Strategy/CL Metric	CL_{score}			$CL_{stability}$		
	W_1	W_2	W_3	W_1	W_2	W_3
Naïve	0.5140	0.5529	0.5312	**0.9986**	0.9969	**0.9973**
Cumulative	0.5128	0.6223	0.5373	0.9979	0.9976	0.9964
EWC	0.4894	0.6449	0.5816	0.9972	0.9976	0.9940
LWF	**0.5768**	**0.6554**	**0.6030**	**0.9986**	**0.9990**	0.9972
SI	0.4861	0.6372	0.5772	0.9970	0.9945	0.9927

Three weight configurations $W = [w_A, w_{MS}, w_{SSS}, w_{CE}, w_{BWT^+}, w_{REM}, w_{FWT}]$:

- W_1: $w_i = \frac{1}{\#\mathcal{C}}$
- $W_2 = [0.4,\ 0.1,\ 0.1,\ 0.1,\ 0.2,\ 0.05,\ 0.05]$
- $W_3 = [0.4,\ 0.05,\ 0.2,\ 0.2,\ 0.05,\ 0.05,\ 0.05]$

[15] Same CNN model as in [Zenke et al., 2017, Maltoni and Lomonaco, 2018] (4 conv. + 2 FC layers)

Experiments[16]

CL metrics for each CL strategy (higher is better)

Str.	A	REM	BWT^+	FWT	MS	SSS	CE	CL_{score}	$CL_{stability}$
Naï	0.3825	0.6664	0.0000	0.1000	**1.0000**	**1.0000**	**0.4492**	0.5140	**0.9986**
Cum	**0.7225**	**1.0000**	**0.0673**	0.1000	**1.0000**	0.5500	0.1496	0.5128	0.9979
EWC	0.5940	0.9821	0.0000	0.1000	0.4000	**1.0000**	0.3495	0.4894	0.9972
LWF	0.5278	0.9667	0.0000	0.1000	**1.0000**	**1.0000**	0.4429	**0.5768**	**0.9986**
SI	0.5795	0.9620	0.0000	0.1000	0.4000	**1.0000**	0.3613	0.4861	0.9970

[16] Using $W_{\mathbf{1}} : w_i = \frac{\mathbf{1}}{\#\mathcal{C}}$

Future Work

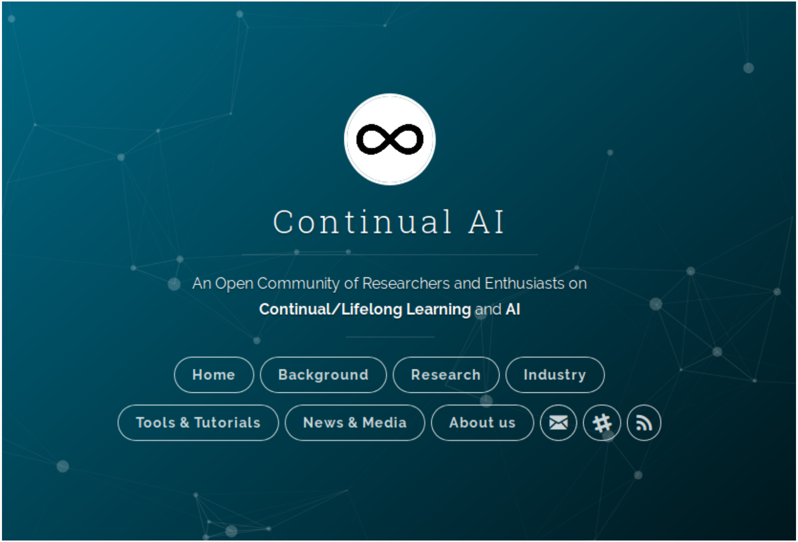
- Provide more insights to assess:
  - importance of different metric schemes
  - their entanglement
- How to use metrics wisely to assist choosing among algorithms?
- Evolve & extend metrics beyond classification
- More datasets[17], tasks, ...
- Adoption (!)

[17] CORe50 CL Dataset `https://vlomonaco.github.io/core50/` [Lomonaco and Maltoni, 2017]

Thank you!

Join! https://www.continualai.org/

Slack channel: https://continualai.herokuapp.com/



References

Sebastian Farquhar and Yarin Gal. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*, 2018.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. ISSN 13646613. doi: 10.1016/S1364-6613(99)01294-2.

Tyler L Hayes, Nathan D Cahill, and Christopher Kanan. New Metrics and Experimental Paradigms for Continual Learning. pages 1–4, 2018. doi: 10.1109/CVPRW.2018.00273.

Alessio Ishizaka and Philippe Nemery. *Multi-criteria decision analysis: methods and software*. John Wiley & Sons, 2013.

RL Keeney and H Raiffa. Decision with multiple objectives, preferences and value tradeoffs. 1993.

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. *ArXiv e-prints*, December 2016.

Z. Li and D. Hoiem. Learning without Forgetting. *ArXiv e-prints*, June 2016.

Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg, editors, *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 17–26. PMLR, 13–15 Nov 2017. URL http://proceedings.mlr.press/v78/lomonaco17a.html.

D. Lopez-Paz and M. Ranzato. Gradient Episodic Memory for Continual Learning. *ArXiv e-prints*, June 2017.

Davide Maltoni and Vincenzo Lomonaco. Continuous learning in single-incremental-task scenarios. *arXiv preprint arXiv:1806.08568*, 2018.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

Joan Serrà, Dídac Surís, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. *CoRR*, abs/1801.01423, 2018. URL `http://arxiv.org/abs/1801.01423`.

Friedeman Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3987–3995, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v70/zenke17a.html`.