Efficient Continual Learning with Modular Networks and Task-Driven Priors

ICLR 2021

Tom Veniat
Ludovic Denoyer
Marc'Aurelio Ranzato

Machine Learning & Deep Learning for Information Access

facebook AI Research

Different levels of diversity

- Classes (incremental Cifar-100 [Krizhevsky 2009], ...)

Different levels of diversity

- Classes (incremental Cifar-100 [Krizhevsky 2009], ...)
- Domain (VDD [Rebuffi, Bilen, and Vedaldi 2017] ...)

Different levels of diversity

- Classes (incremental Cifar-100 [Krizhevsky 2009], ...)
- Domain (VDD [Rebuffi, Bilen, and Vedaldi 2017] ...)
- Modalities (Text, Vison, Audio, ...)

Different levels of diversity

- Classes (incremental Cifar-100 [Krizhevsky 2009], ...)
- Domain (VDD [Rebuffi, Bilen, and Vedaldi 2017] ...)
- Modalities (Text, Vison, Audio, ...)
- **Amount of data**

The problem

- Some tasks will require transferring knowledge from past experiences to be solved.
- Focus on maximizing transfer instead of minimizing forgetting.

The problem

- Some tasks will require transferring knowledge from past experiences to be solved.
- Focus on maximizing transfer instead of minimizing forgetting.

Contribution

- Identify a set of desired properties of a CL system to deal with this diversity.
- Define metrics and streams for assessing the performance of an algorithm in each of these dimensions : CTrL
- Propose a new model able to perform well in most of the identified scenarios : MNTDP

Direct Transfer

$$\mathcal{S}^{-} = (t_1^+, t_2, t_3, t_4, t_5, t_1^-)$$

SCIENCES SORBONNE UNIVERSITÉ

Direct Transfer

$$\mathcal{S}^{-} = (t_1^{+}, t_2, t_3, t_4, t_5, t_1^{-})$$

Knowledge Update

$$\mathcal{S}^{+} = (t_1^{-}, t_2, t_3, t_4, t_5, t_1^{+})$$

SCIENCES SORBONNE UNIVERSITÉ

Direct Transfer

$$\mathcal{S}^{-} = (t_1^+, t_2, t_3, t_4, t_5, t_1^-)$$

Knowledge Update

$$\mathcal{S}^{+} = (t_1^-, t_2, t_3, t_4, t_5, t_1^+)$$

Transfer to similar Input/Output Distributions

$$\mathcal{S}^{\text{in}} = (t_1, t_2, t_3, t_4, t_5, t_1')$$

$$\mathcal{S}^{\text{out}} = (t_1, t_2, t_3, t_4, t_5, t_1'')$$

Plasticity

$$\mathcal{S}^{\mathsf{pl}} = (t_1, t_2, t_3, t_4, t_5)$$

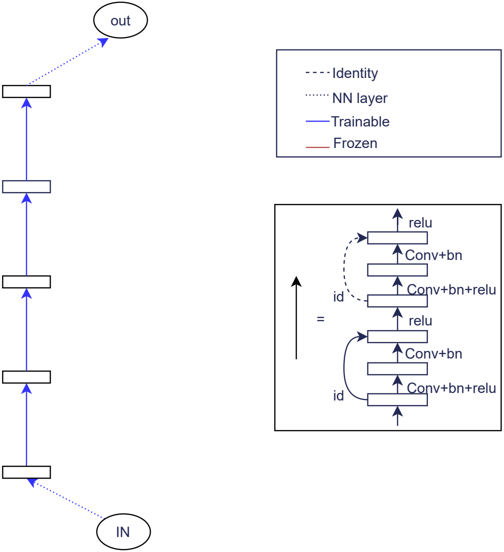
Plasticity

$$\mathcal{S}^{\mathsf{pl}} = (t_1, t_2, t_3, t_4, t_5)$$

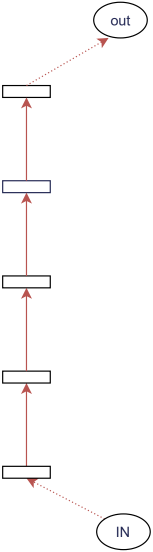
Scalability

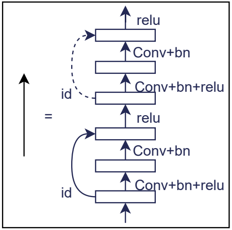
- $\mathcal{S}^{\mathsf{long}}$ composed of 100 tasks.
- Contains the 5 other scenarios

Modular Network

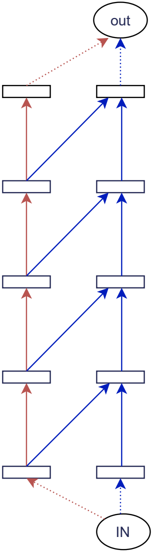


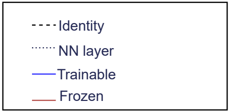
Modular Network

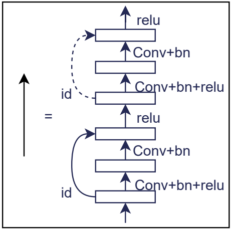




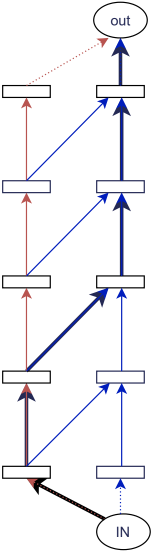
Modular Network

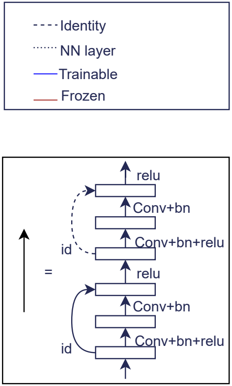




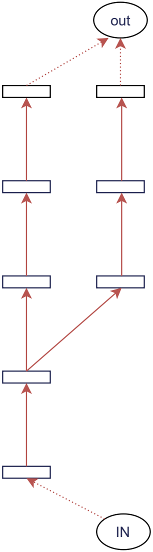


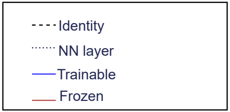
Modular Network

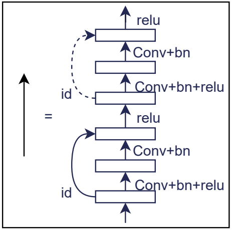




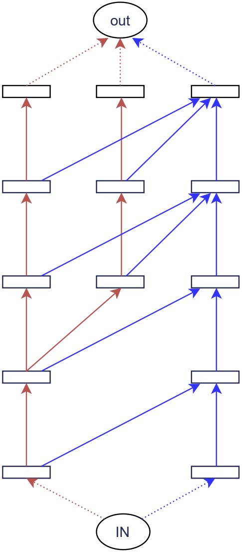
Modular Network

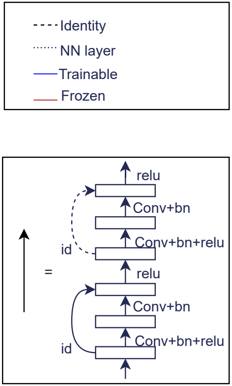




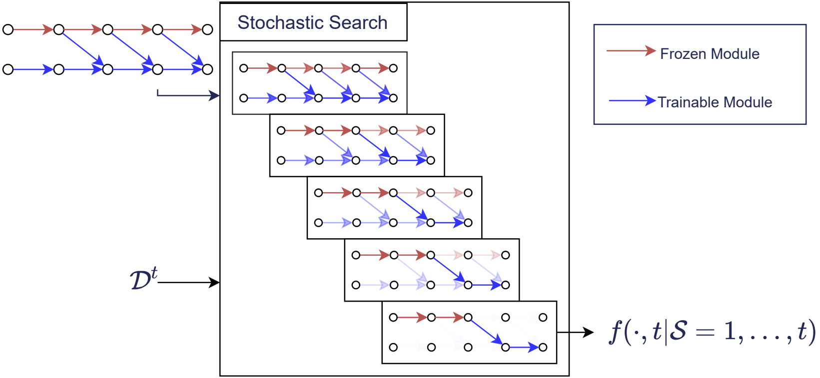


Modular Network





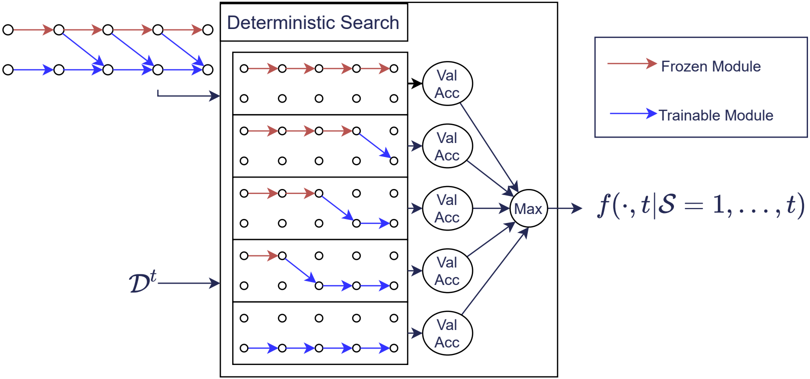
MNTDP-S



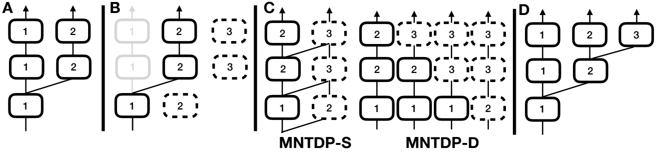
Stochastic Search
Frozen Module
Trainable Module
$\mathcal{D}^t$
$f(\cdot, t|\mathcal{S} = 1, \dots, t)$

SCIENCES SORBONNE UNIVERSITÉ

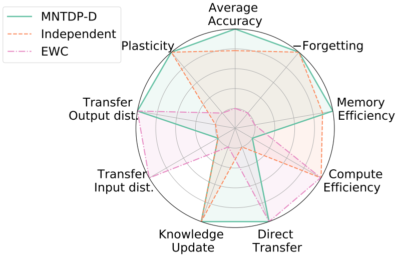
MNTDP-D



Modular Network + Task-Driven Prior



Radar plots on CTrL



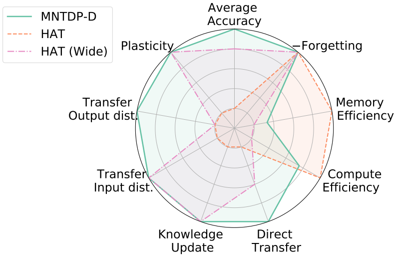


Figure 1: Comparison of various CL methods on the CTrL benchmark using Resnet (left) and Alexnet (right) backbones. MNTDP-D is our method.

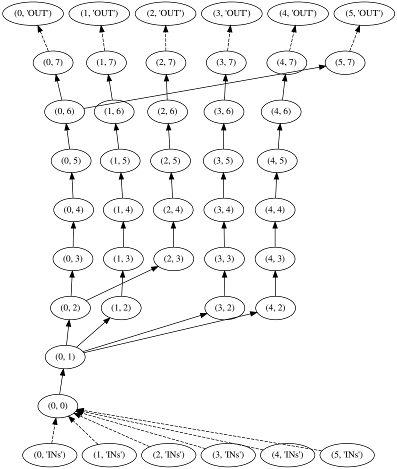


Figure 2: Global graph of paths discovered by MNDTP-D on the $\mathcal{T}(\mathcal{S}^{\text{out}})$ Stream. "INs" (resp. "OUT") nodes are the input (resp. output) of the path for each task.

$\mathcal{S}^{\text{long}}$

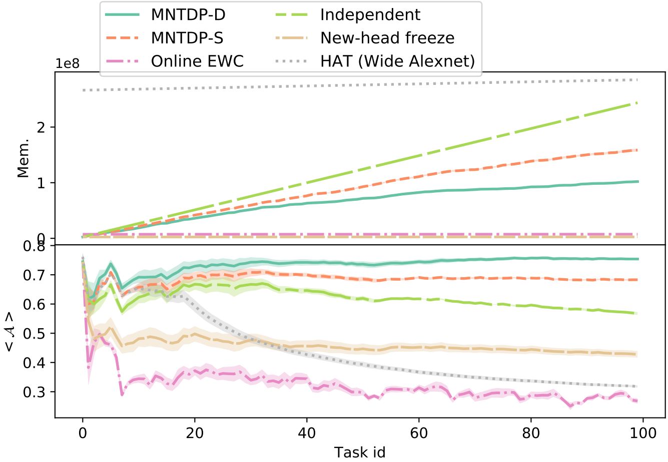


Figure 3: Evolution of $< \mathcal{A} >$ and Mem. on $\mathcal{S}^{\text{long}}$.

Krizhevsky, Alex (2009). "Learning Multiple Layers of Features from Tiny Images". In: *University of Toronto, technical report*.

Rebuffi, Sylvestre-Alvise, Hakan Bilen, and Andrea Vedaldi (2017). "Learning multiple visual domains with residual adapters". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon et al., pp. 506–516. URL: http://papers.nips.cc/paper/6654-learning-multiple-visual-domains-with-residual-adapters.