PODNet: Pooled Outputs Distillation for Small-Tasks Incremental Learning

Arthur Douillard
Matthieu Cord
Charles Ollion
Thomas Robert
Eduardo Valle

@Ar_Douillard
arthurdouillard.com

Machine Learning & Deep Learning for Information Access

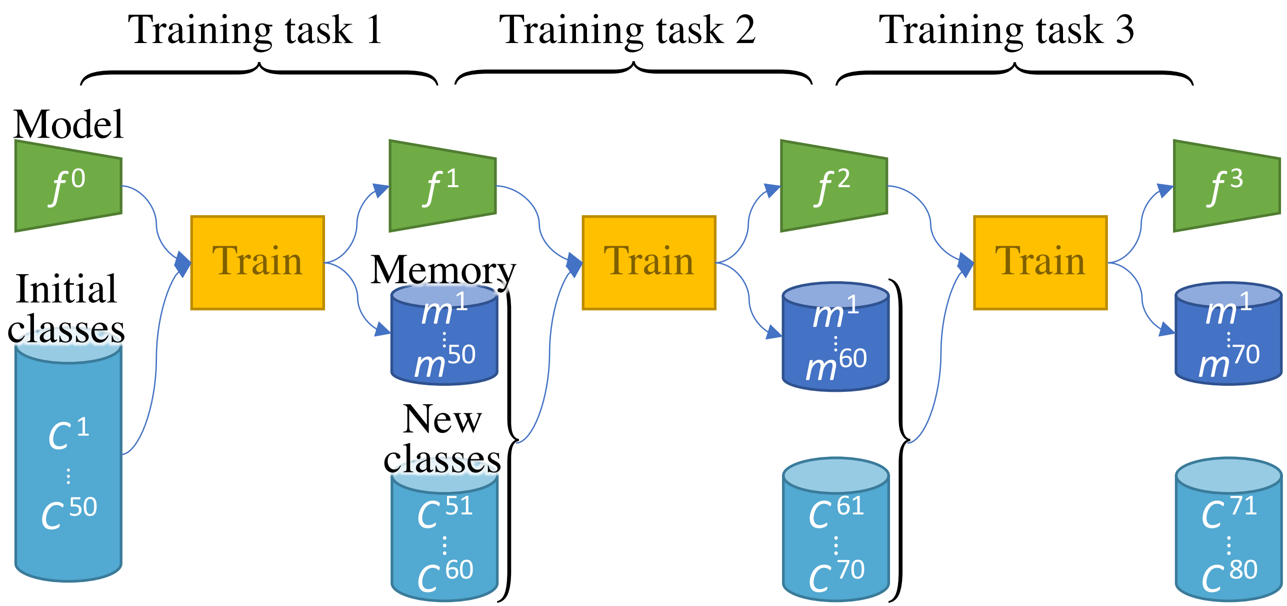
Setting

Incremental Learning

Each new task brings **new classes**

After each task, evaluation is done on **all seen classes**

Previous task data is available in a limited quantity in a **rehearsal memory**

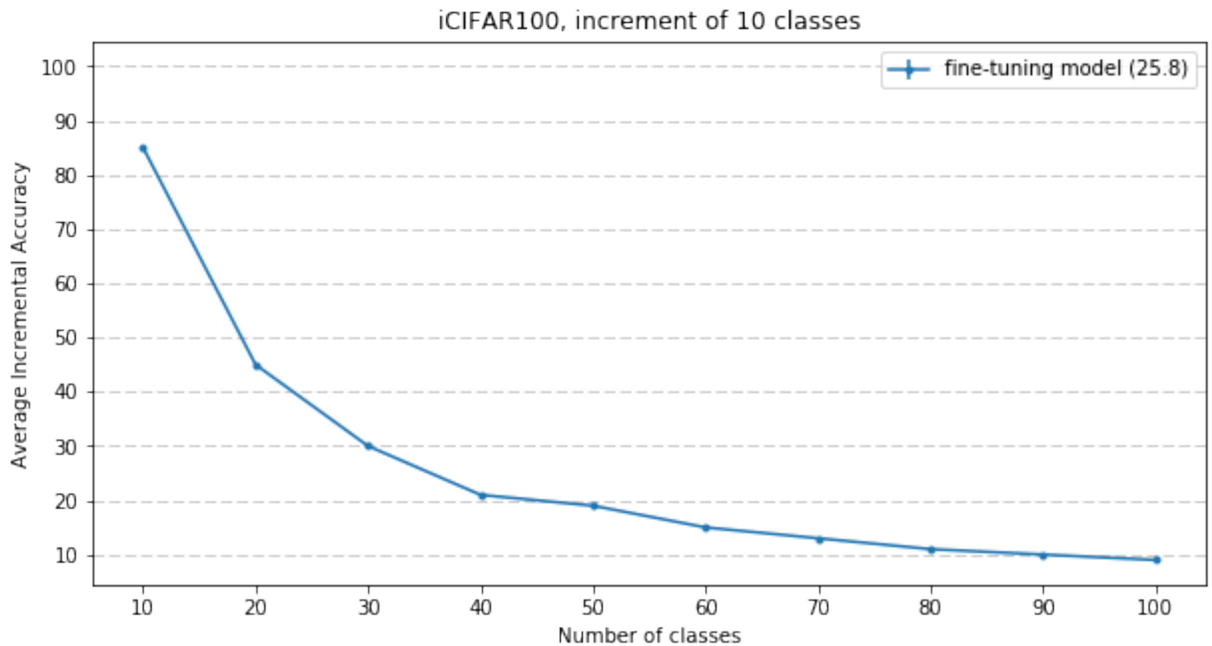


Also known as Continual Learning or Lifelong Learning

Catastrophic Forgetting

Learning new classes with few old classes data produce a

Catastrophic Forgetting

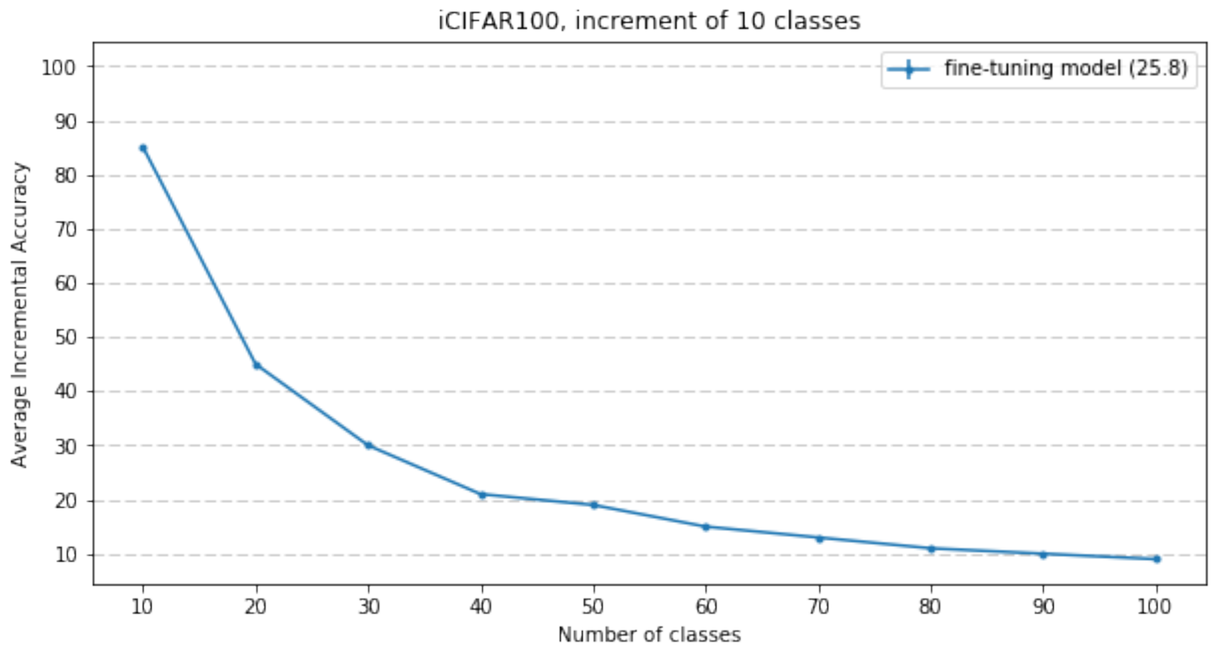


Tradeoff

Rigidity: not forgetting previous knowledge

vs

Plasticity: learning new knowledge



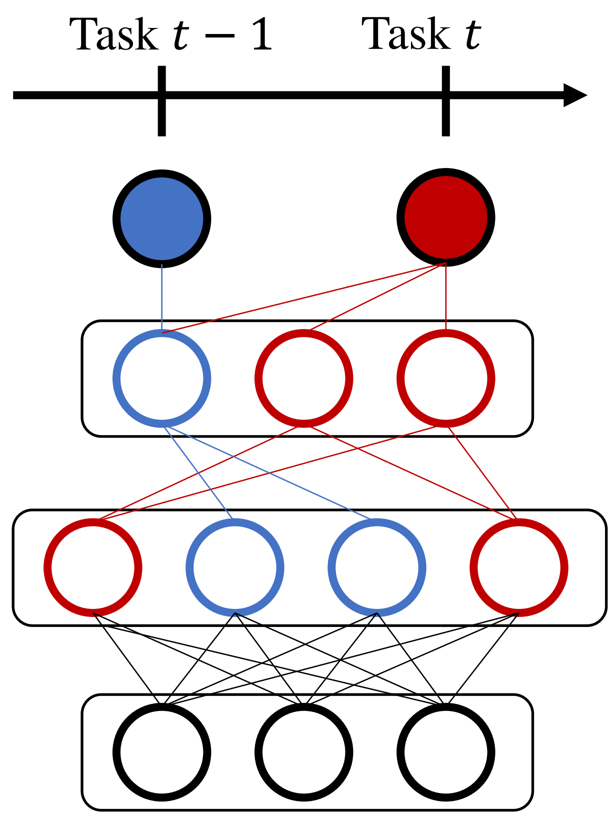
Existing Solutions

Sub-networks

One **sub-network** per task

Often requires in inference the **task id** to select the task-specific sub-network.

Sub-network can be uncovered via evolutionary algorithms (*Fernando et al, 2017*), sparsity (*Golkar et al, 2019*), or learned masks (*Hung et al, 2019*).

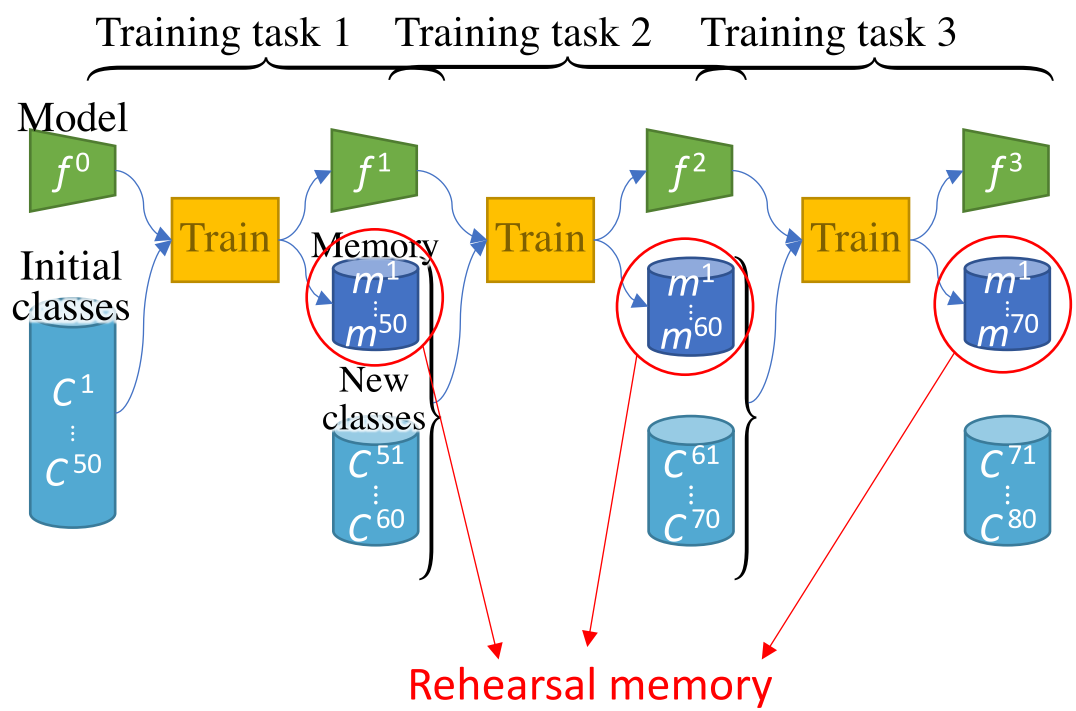


Two sub-networks & can co-exist in the same network

Rehearsal

Re-using a **limited amount** of previous task data (*Rebuffi et al, 2017*)

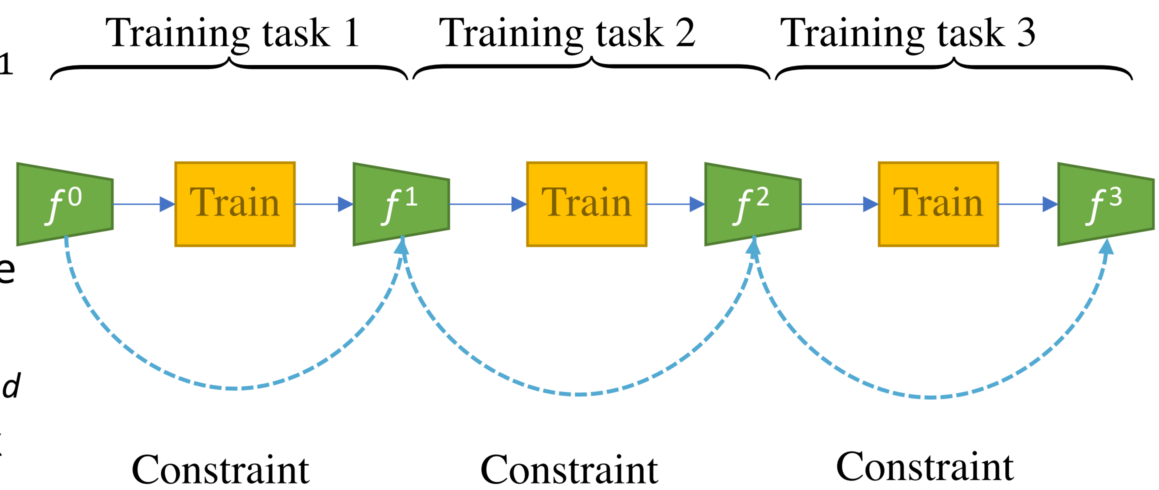
Or **generating** previous task data (*Shin et al, 2017*)



Distillation

Constrain the model f^t to be **similar** to the model f^{t-1}

Can enforce similarity on the **weights** (*Kirkpatrick et al, 2016*), on the **gradients** (*Lopez-Paz and Ranzato, 2017*), or the network **outputs** (*Li and Hoeim, 2016*).



Our model

Use **rehearsal** learning

Use **distillation** on the network outputs

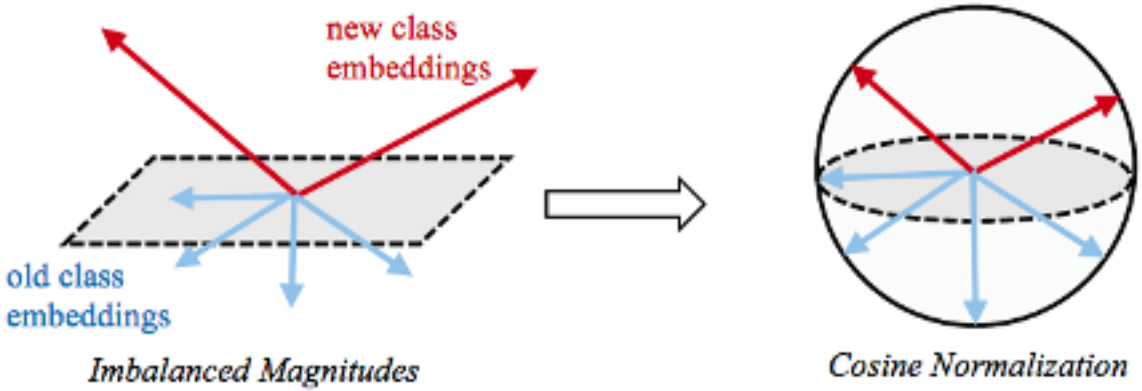
Introduce an **architectural change** on the classifier

Table of Content

LSC:
Local Similarity Classifier

Classifier

Based on a **cosine classifier**

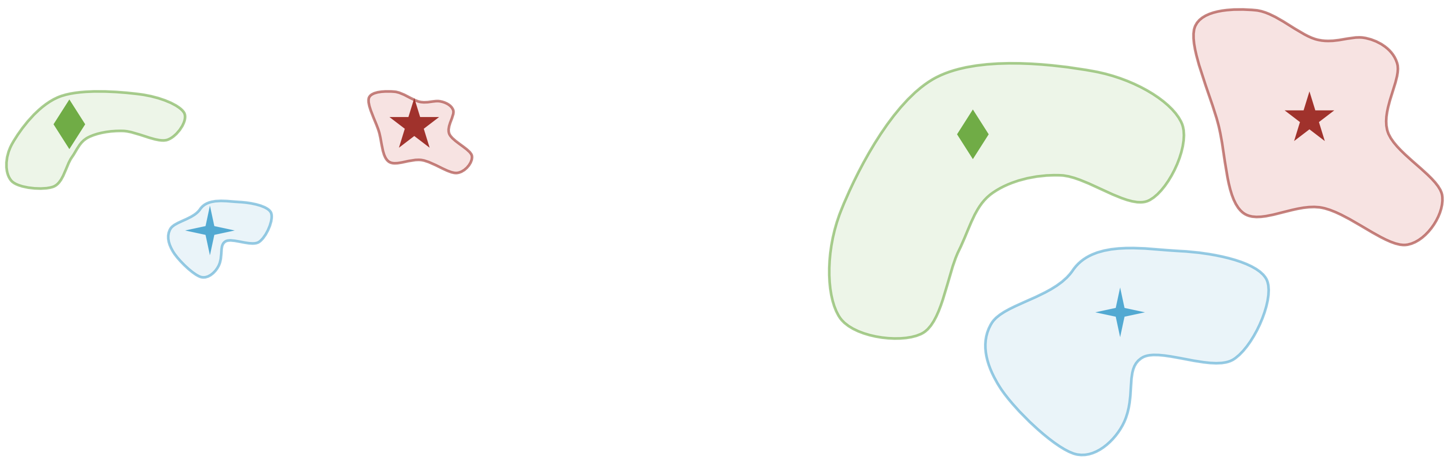


Each centroid represent the **majority mode** of its class

Classifier

Complex classes are made of **multiple modes**

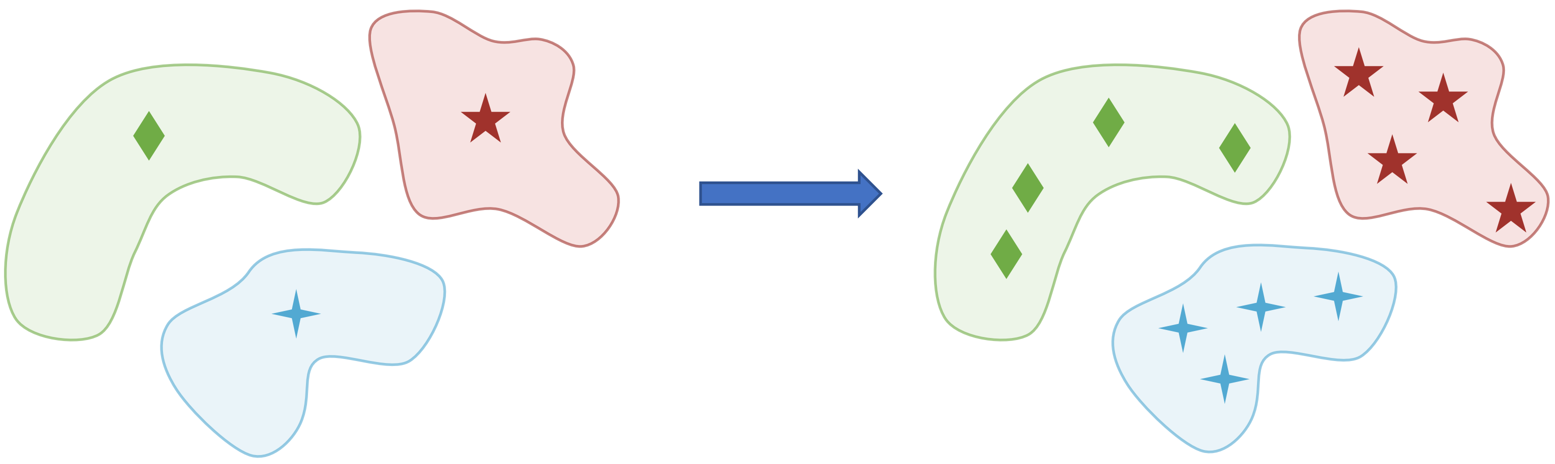
The incremental learning **distorts** class embeddings, making the majority mode a poor centroid



Task 1 Task N

Classifier

Modeling **multiple modes** per class → more robust to **distribution change**



One mode per class

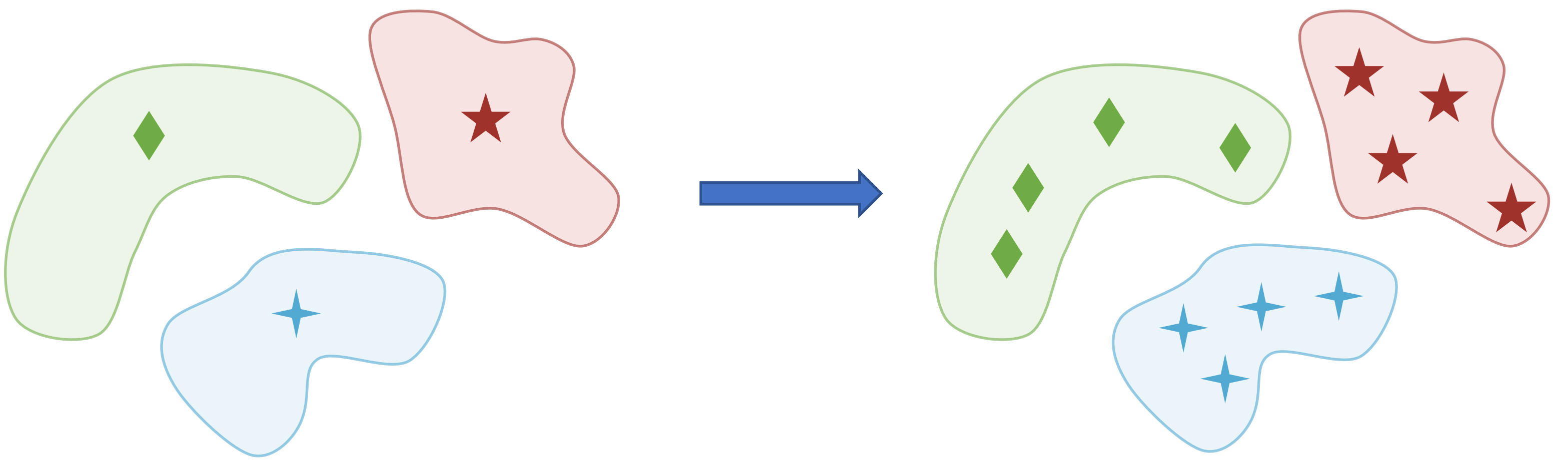
Four modes per class

Classifier

heuritech SCIENCES SORBONNE UNIVERSITÉ

Performance increase mainly because the **old classes are less forgotten**

+1.18pts and +1.51pts on CIFAR & ImageNet



One mode per class

Four modes per class

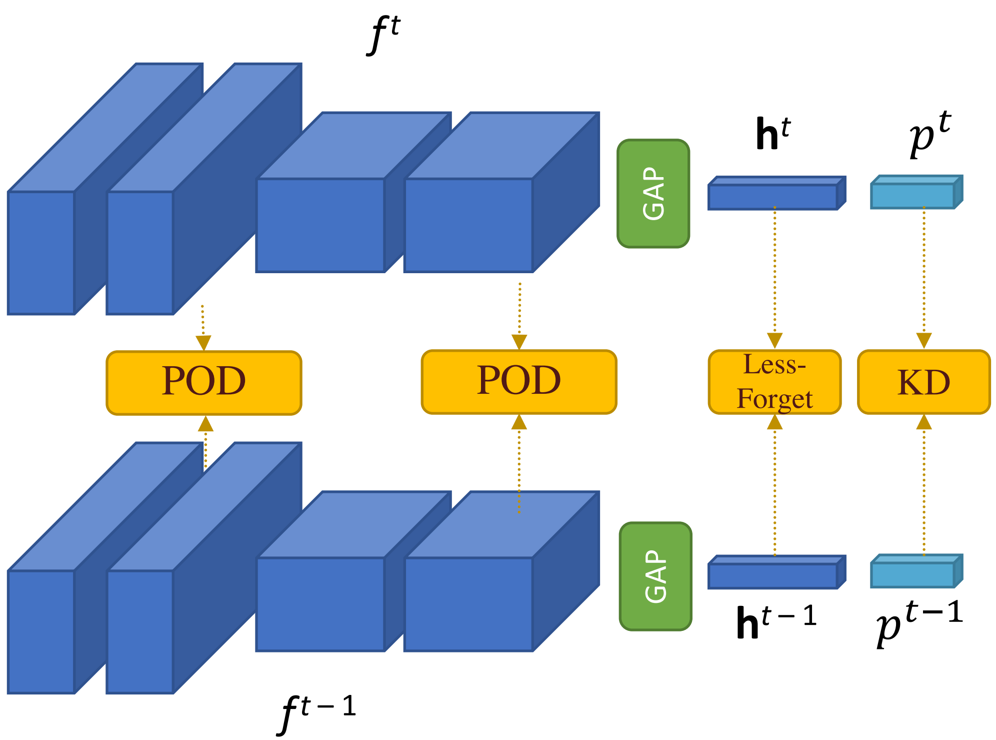
POD:
Pooled Outputs Distillation

Distillation

Knowledge Distillation constrains probabilities

Less-Forget constrains embeddings

POD constrains spatial features



POD Distillation

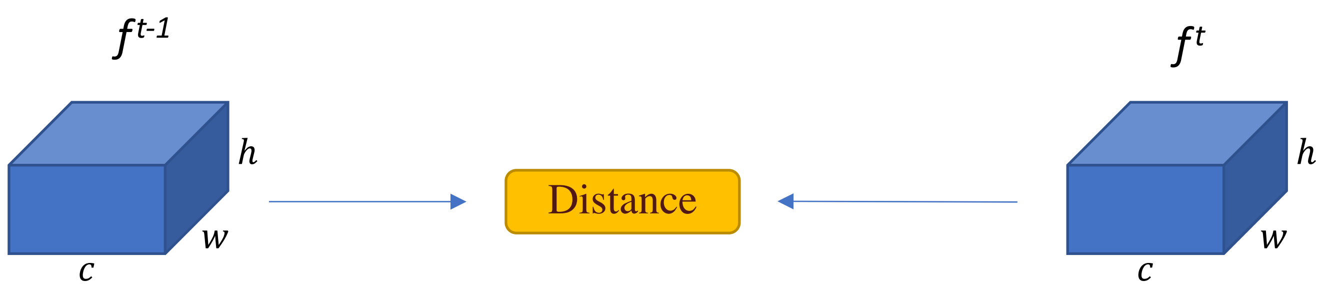
Naive distance between features doesn’t work

$c \times w \times h$ constraints

→ too **rigid**

→ sensitive to outliers

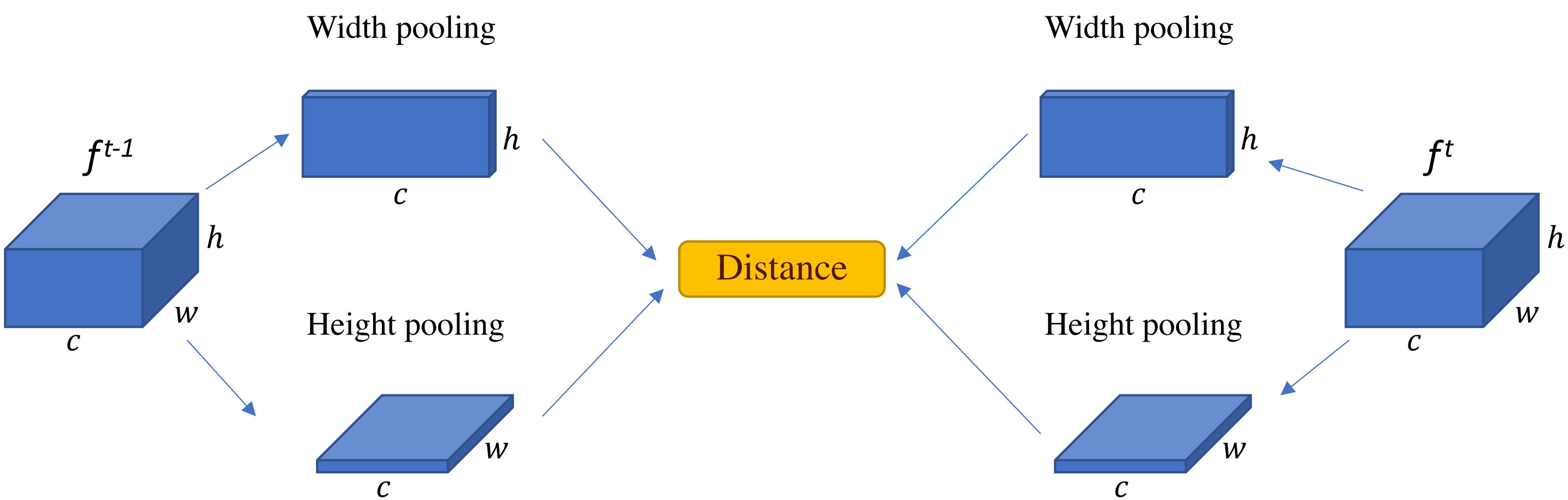
→ no spatial prior



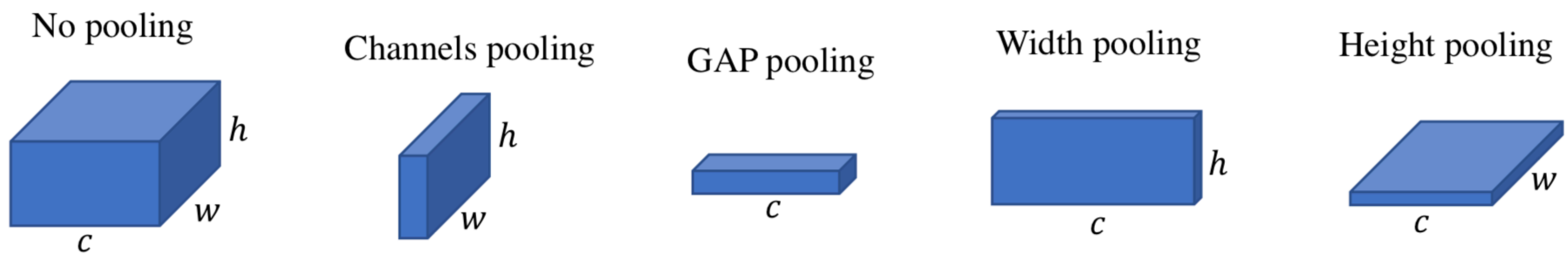
POD Distillation

Distance between **spatial statistics**

Balancing **rigidity** (not forgetting) and **plasticity** (learning)



POD Distillation



No pooling, distance directly on pixels

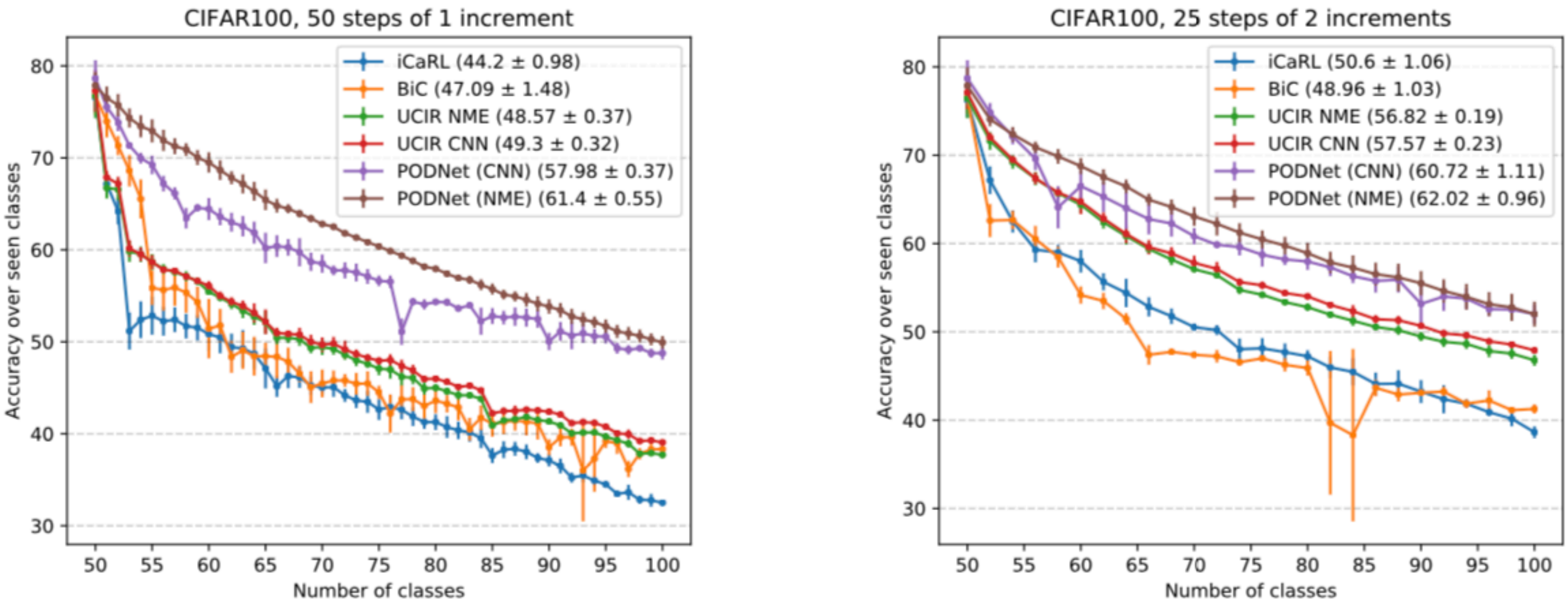
POD-width + POD-height

Loss	NME	CNN
None	53.29	52.98
POD-pixels	49.74	52.34
POD-channels	57.21	54.64
POD-gap	58.80	55.95
POD-width	60.92	57.51
POD-height	60.64	57.50
POD-spatial	**61.40**	**57.98**
GradCam [5]	54.13	52.48
Perceptual Style [14]	51.01	52.25

Results

Results

Outperforming SotA with **large amount of tasks** with little forgetting



(a) 50 steps, 1 class / step

(b) 25 steps, 2 classes / step

Results

	ImageNet100				Imagenet1000	
	50 steps	25 steps	10 steps	5 steps	10 steps	5 steps
New classes per step	1	2	5	10	50	100
iCaRL* [30]	—	—	59.53	65.04	46.72	51.36
iCaRL [30]	54.97	54.56	60.90	65.56	—	—
BiC [35]	46.49	59.65	65.14	68.97	44.31	45.72
UCIR (NME)* [13]	—	—	66.16	68.43	59.92	61.56
UCIR (NME) [13]	55.44	60.81	65.83	69.07	—	—
UCIR (CNN)* [13]	—	—	68.09	70.47	61.28	64.34
UCIR (CNN) [13]	57.25	62.94	67.82	71.04	—	—
PODNet (CNN)	**62.48**	**68.31**	**74.33**	**75.54**	**64.13**	**66.95**
	± 0.59	**± 2.45**	**± 0.93**	**± 0.26**		

Results

Table 4. Effect of the memory size per class M_{per} on the models performance. Results from CIFAR100 with 50 steps, we report the average incremental accuracy

M_{per}	5	10	**20**	50	100	200
iCaRL [30]	16.44	28.57	44.20	48.29	54.10	57.82
BiC [35]	20.84	21.97	47.09	55.01	62.23	**67.47**
UCIR (NME) [13]	21.81	41.92	48.57	56.09	60.31	64.24
UCIR (CNN) [13]	22.17	42.70	49.30	57.02	61.37	65.99
PODNet (NME)	**48.37**	**57.20**	**61.40**	**62.27**	**63.14**	63.63
PODNet (CNN)	**35.59**	**48.54**	**57.98**	**63.69**	**66.48**	**67.62**

Summary

1. LSC: Local Similarity Classifier
2. POD: Pooled Outputs Distillation
3. Experiments up to 50 tasks

Code is available!

https://github.com/arthurdouillard/incremental_learning.pytorch