On the importance of cross-task features for class-incremental learning

Albin Soutif--Cormerais, Marc Masana,
Joost Van de Weijer, Bartlomiej Twardowski

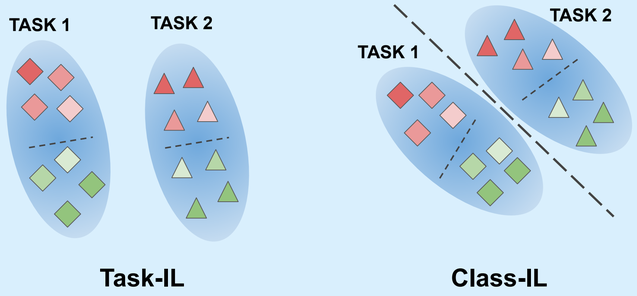
Centre de Visió per Computador
Universitat Autònoma de Barcelona
SAL SILICON AUSTRIA LABS
TU Graz
HUAWEI
HISILICON

Overview

We consider two settings in supervised continual learning:

- **task-incremental learning** → task-ID at test time
- **class-incremental learning** → no task-ID at test time

Therefore, class-IL requires the learner to perform task-inference too. We call the features that enable it to do so the **cross-task features.**



Our contributions are the following:

- we study the effect of not explicitly learning cross-task features on classical class-IL benchmarks,
- we design a new forgetting metric that better suits class-IL and show that forgetting is not the main cause of performance drop in this setting.

Learning, and not, of cross-task features

We believe that learning cross-task features is a hard task in continual learning, since they can only be learned using the data from the limited memory from old tasks. In order to study how important the learning of cross-task features actually is, we design two baseline using replay, one learns the cross-task features, while the other does not. During an additional finetuning step, both use the cross-entropy loss (2) to tune the classifier only. That way it is possible to evaluate both method on the class-IL task

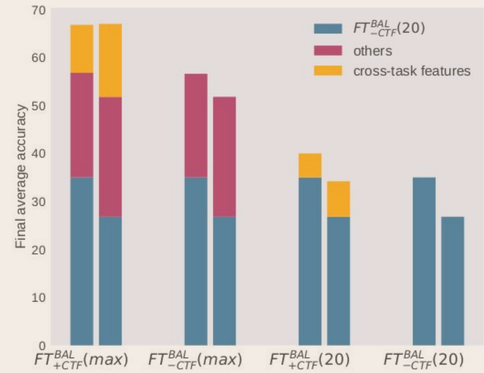
$$FT^{BAL}_{+CTF} \qquad \mathcal{L}_{CE}(\mathcal{B};\theta) = \sum_{(x,y)\in\mathcal{B}} -\log \frac{\exp f(x;\theta)_y}{\sum_{c\in C^T_\Sigma} \exp f(x;\theta)_c}, \qquad (2)$$

$$FT^{BAL}_{-CTF} \qquad \mathcal{L}_{CE\text{-}IT}(\mathcal{B};\theta) = \frac{1}{T}\sum_{t=1}^{T} \sum_{(x,y)\in\mathcal{B}^t} -\log \frac{\exp f(x;\theta)_y}{\sum_{c\in C^t} \exp f(x;\theta)_c}. \qquad (3)$$

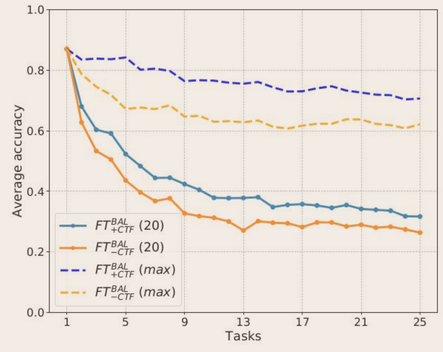
On the influence of cross-task features

Cross task features:

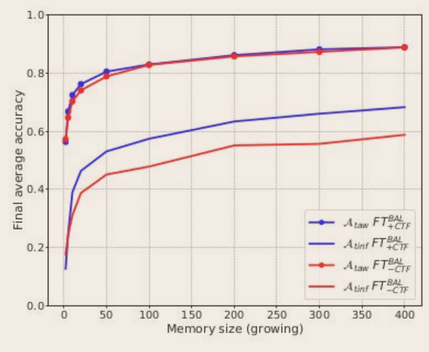
- They are learned in part by the method with limited memory (5 and 7%)
- The part in red is predominant and is not due to the absence of cross-task features. It can be addressed via better intra-task features.
- We suspect that **better knowledge transfer** (forward, backward transfer [1]) between tasks can contribute to filling the red part.



Final average accuracy on CIFAR100, 10 taskss (left) and 20 takss (right). Blue rectangle present the performance of the method that does not learn cross task features. The influence of cross-task features is then displayed in yellow. For the two splits, we see that this is not the most influent factor.



Average accuracy of both baseline and their counterparts using maximum memory on Imagenet-Subset (25 tasks). In comparison to CIFAR, while there are more tasks, cross-task features have less importance for this dataset..



Final task aware (dotted lines), and task-inference accuracies, CIFAR100 10 tasks, with a growing amount of memory available. After 50 exemplars per task, the gap in task inference between the two baseline remain stable. Better task-inference is then obtained with better intra-task features.

scenario	approach	growing memory size 20/cls	10/cls	5/cls	2/cls	max
10 tasks	FT^{BAL}_{+CTF}	40.55 ± 2.1	31.5 ± 3.0	19.1 ± 2.5	9.0 ± 0.69	66.8 ± 1.1
	FT^{BAL}_{-CTF}	34.8 ± 2.1	27.4 ± 1.5	20.1 ± 1.2	14.3 ± 0.5	56.6 ± 1.2
20 tasks	FT^{BAL}_{+CTF}	34.2 ± 1.9	26.6 ± 2.7	12.5 ± 1.6	4.4 ± 0.15	67.0 ± 1.5
	FT^{BAL}_{-CTF}	26.8 ± 2.6	20.1 ± 2.0	15.1 ± 1.7	9.2 ± 0.7	51.8 ± 2.3

Results on CIFAR100, 10 and 20 tasks split for different memory configurations. Results for maximum memory (500 exemplars per class) are given on the right.

Correlation between intra- and cross-task features

Unlike in the introductory example, in real image datasets, it is very likely that intra- and cross- task features are correlated. Thus, increasing the quality of intra-task features also increases task-inference capability of the network (see Fig. above).

Cumulative forgetting

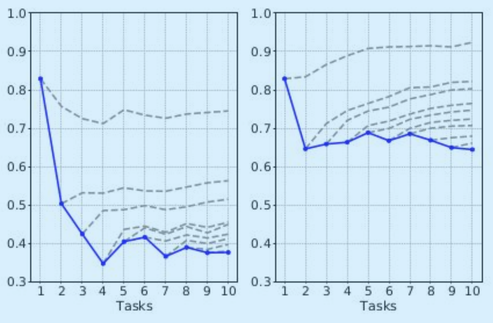
- The *forgetting metric from task-IL [2] is often misused in class-IL*. For instance, by computing it directly on the average accuracy, which results in very high forgetting.
- To design our new forgetting metric, we instead consider the sequential cumulative tasks, composed of all classes seen until task k, and compute our cumulative forgetting by only considering logits corresponding to these classes.

$\mathcal{T}^k_\Sigma = (C^k_\Sigma, D^k_\Sigma)$. The cumulative task k

$$\hat{y}_k(x;\theta) = \arg\max_{c\in C^k_\Sigma} f(x;\theta)_c\,, \qquad (4)$$

$$b^t_k = \frac{1}{|D^k_\Sigma|} \sum_{x,y\in D^k_\Sigma} 1_{\{y\}}\big(\hat{y}_k(x;\theta^t)\big)\,, \qquad (5)$$

$$f^t_k = \max_{i\in\{1,\dots,t\}} b^i_k - b^t_k\,. \qquad (6)$$



Cumulative accuracies for the baseline learning cross-task features, using 20 exemplars per class (felt), and using the maximum amount of memory (right). In the first case, cumulative accuracies remain stable, resulting in low cumulative forgetting. In the second case, they grow over time, which indicated positive backward transfer.

References

[1] Lopez-Paz, D. and Ranzato, M. A. Gradient episodic memory for continual learning. Advances in Neural Information Processing Systems (Neurips) 2017

[2] Chaudhry, A., Dokania, P. K., Ajanthan, T., and Torr, P. H. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In Proceedings of the European Conference on Computer Vision (ECCV),