


Preprint on arXiv:
https://arxiv.org/abs/2304.00933

“Knowledge Accumulation in Continually Learned Representations and the Issue of Feature Forgetting”

Timm Hess, Eli Verwimp*, Gido M. van de Ven, Tinne Tuytelaars*

* joint first author

Do representations forget catastrophically?

- Neural networks suffer from catastrophic forgetting “at the output level”. Is this also true at the level of representations?

- Recent studies imply an innate robustness to forgetting for representations:

  Davari et al. (2022, CVPR):
  “[…] in many commonly studied cases of catastrophic forgetting, the representations under naive finetuning approaches, undergo minimal forgetting, without losing critical task information.”

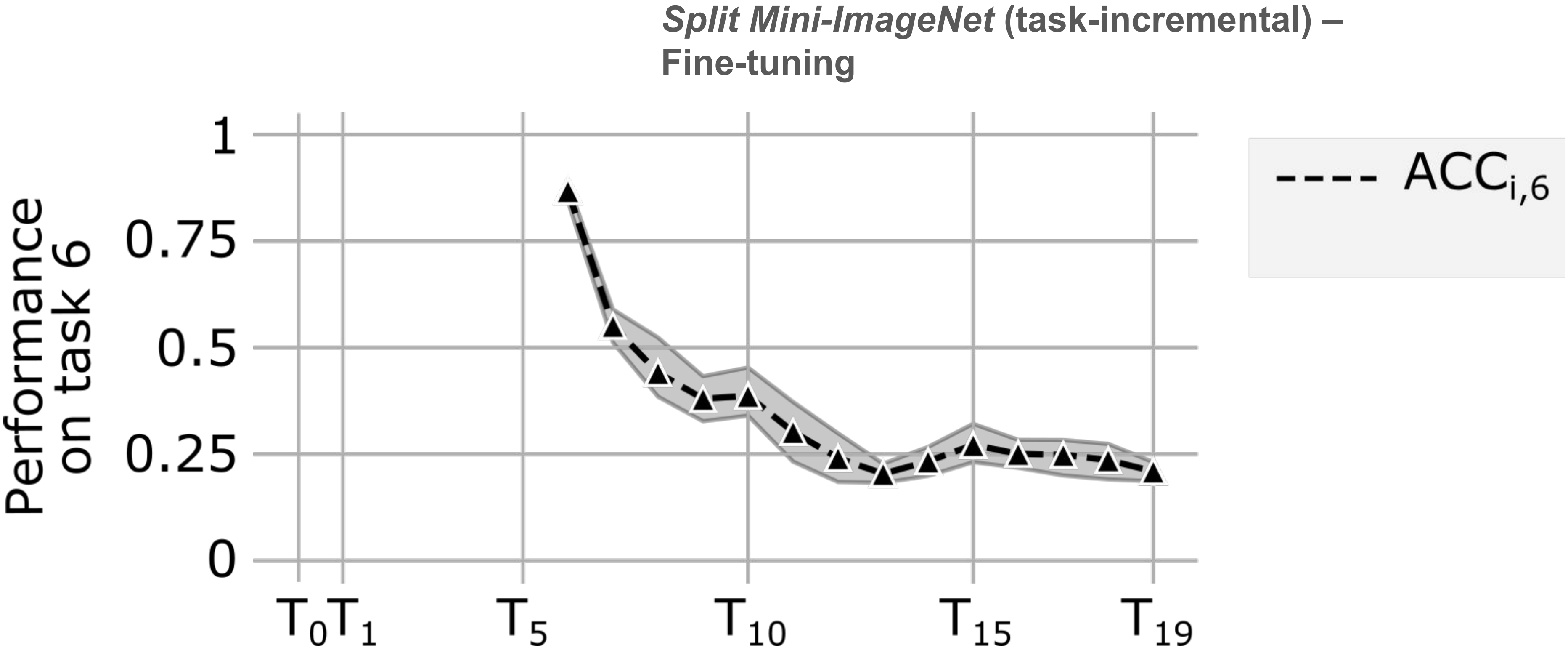
  Zhang et al. (2022, arXiv):
  “there seems to be no catastrophic forgetting in terms of representations.”

Measuring representation quality

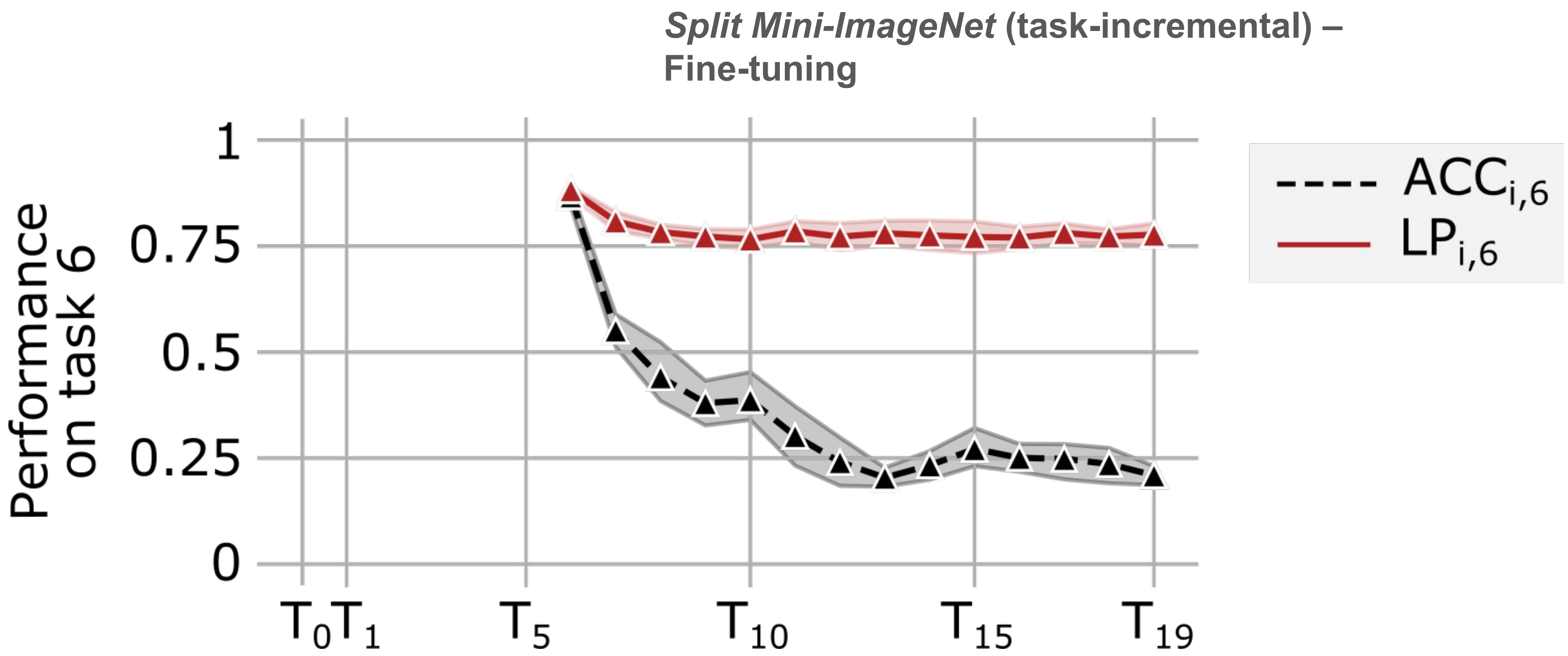
- We measure ‘representation quality’ using the metric *linear probe accuracy*, denoted $LP_{i,j}$

- After finishing training on task *i*, we train a new head for task *j* using all training data from task *j*, while freezing the representation layers

Comparing forgetting in representation and at output level

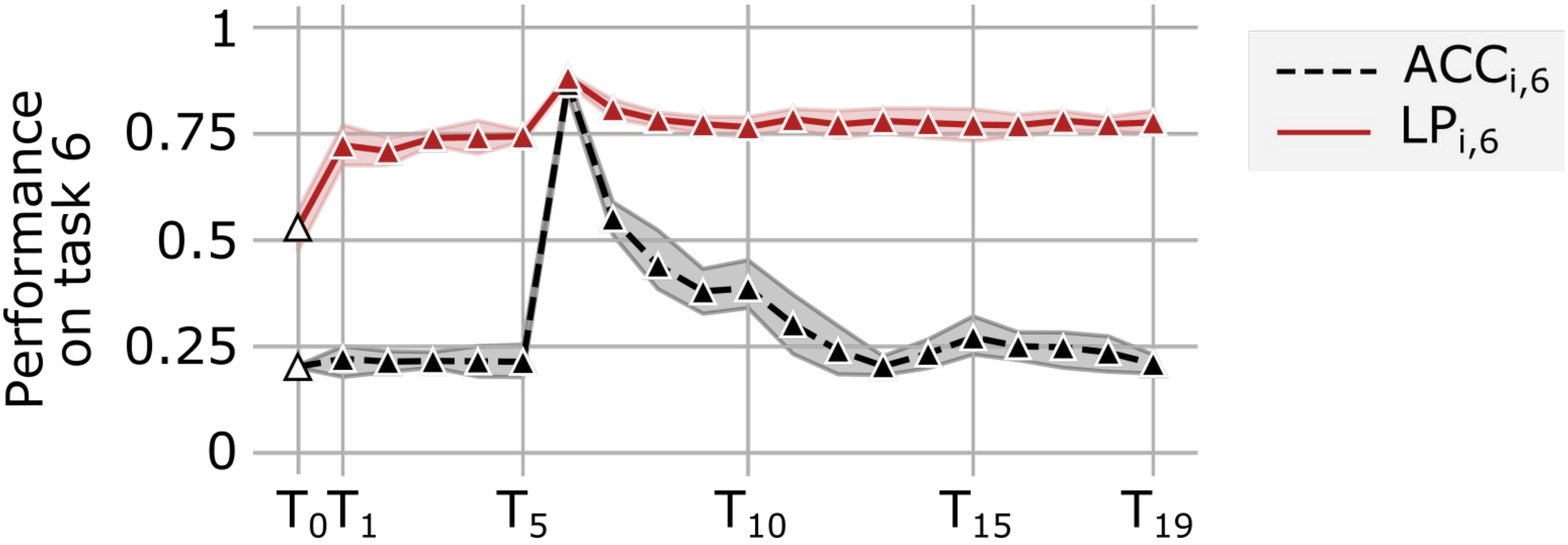


Comparing forgetting in representation and at output level



Comparing forgetting in representation and at output level

Split Mini-ImageNet **(task-incremental) – Fine-tuning**



Relative forgetting

- The proportion of new knowledge gained during training on task *i* that is lost when training further on other tasks:
  - with $r_{i,j}$ the performance on task *j* after finishing training task *i*

Knowledge lost when training further on other tasks

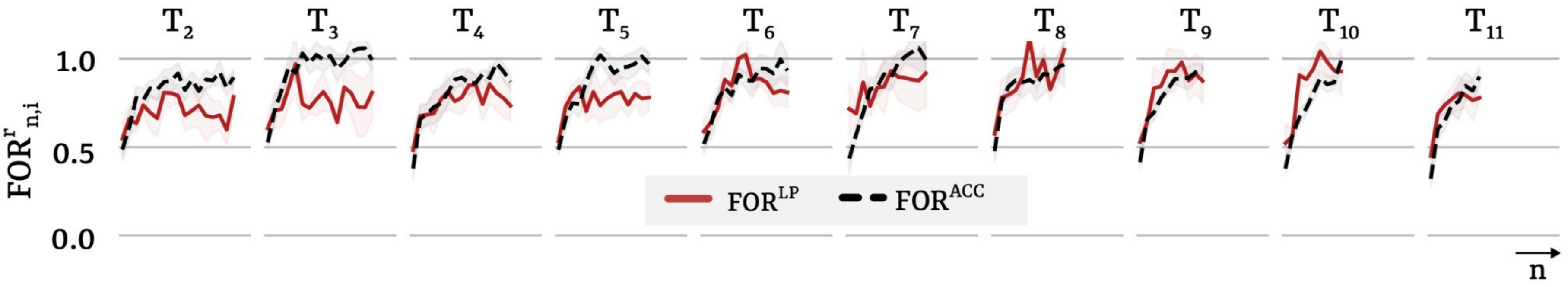
$$\mathrm{FOR}^{r}_{n,i} = \frac{r_{i,i} - r_{i+n,i}}{r_{i,i} - r_{i-1,i}}$$

New Knowledge gained during training task *i*

Relative forgetting: representations forget 'catastrophically'

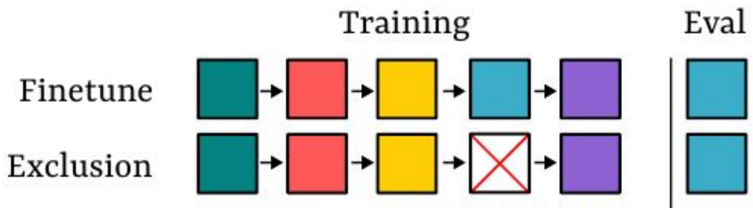
- The proportion of new knowledge gained during training on task *i* that is lost when training further on other tasks:
    - with $r_{i,j}$ the <u>performance</u> on task *j* after finishing training task *i*

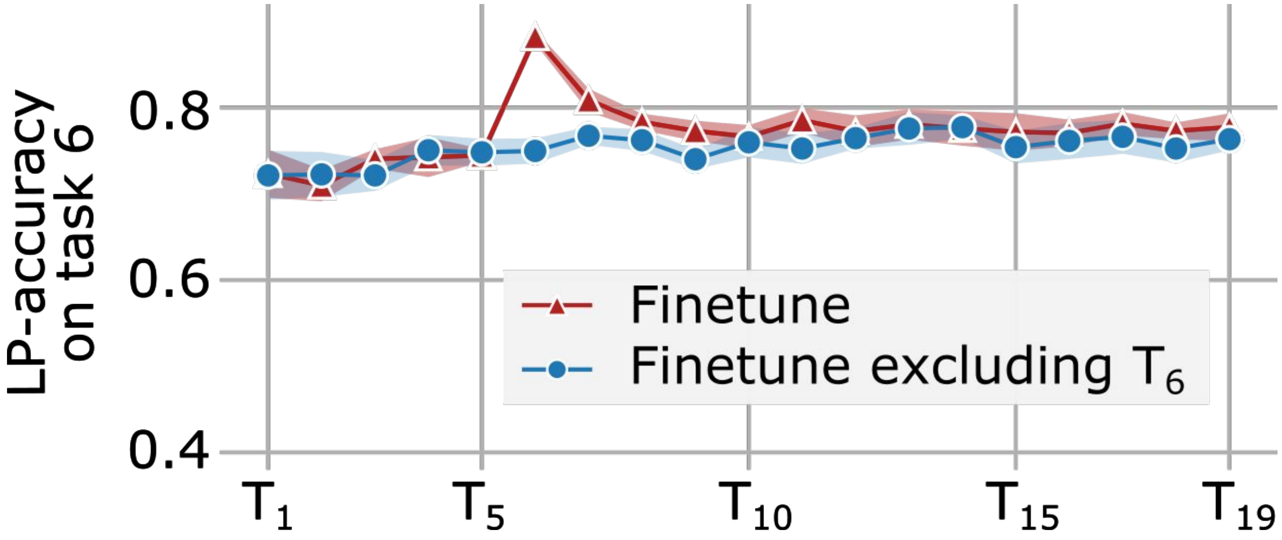
$$\mathrm{FOR}^{r}_{n,i} = \frac{r_{i,i} - r_{i+n,i}}{r_{i,i} - r_{i-1,i}}$$



Task exclusion baseline

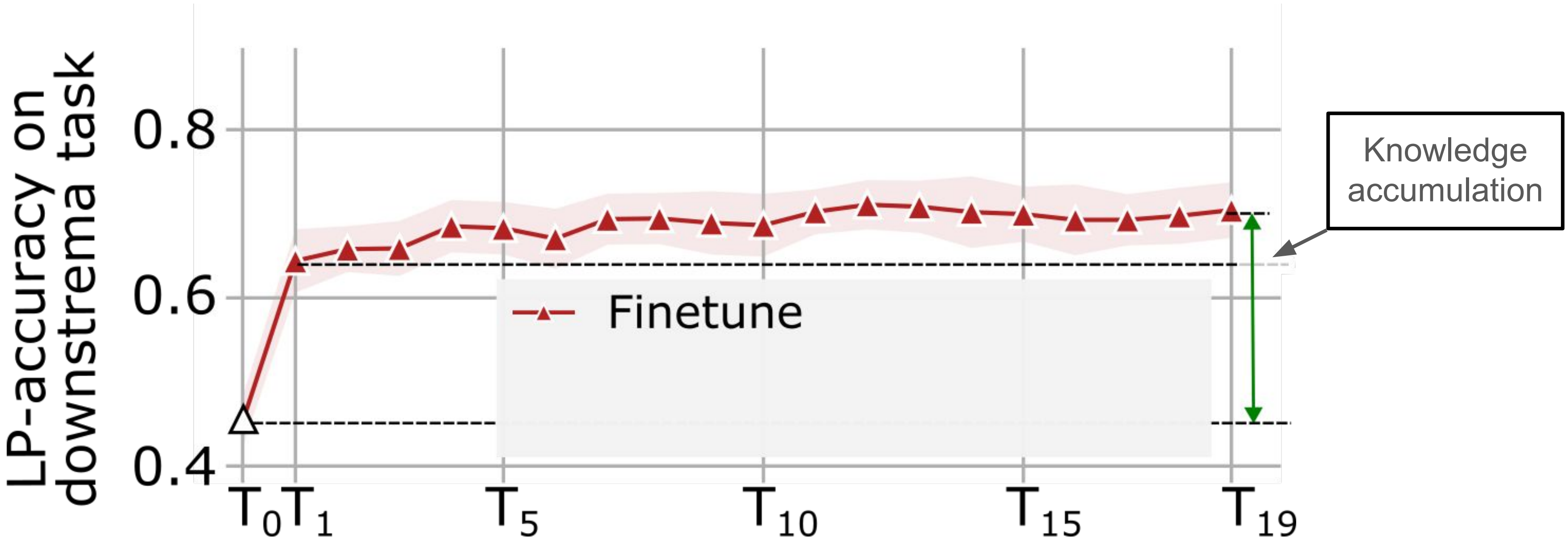
- In the end, for the representation quality for task 6, it does not matter whether or not the model is trained on task 6





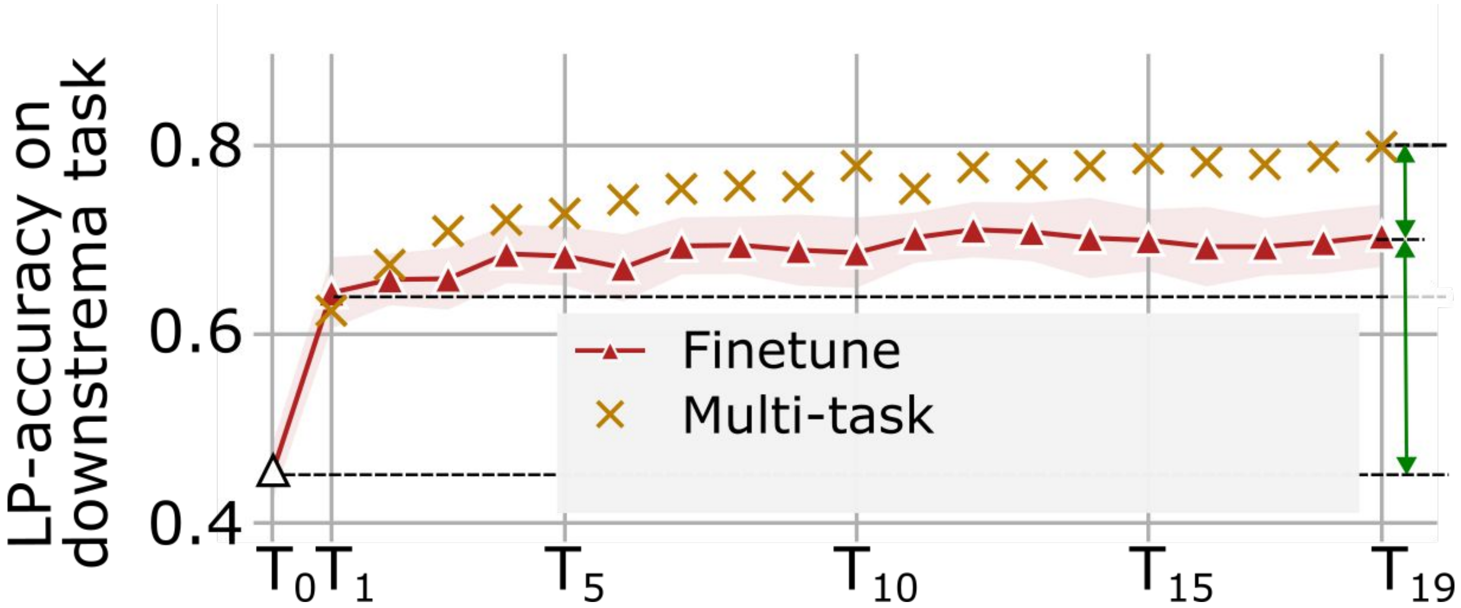
Is feature forgetting problematic?

- It has been argued representations only forget “task-specific” knowledge



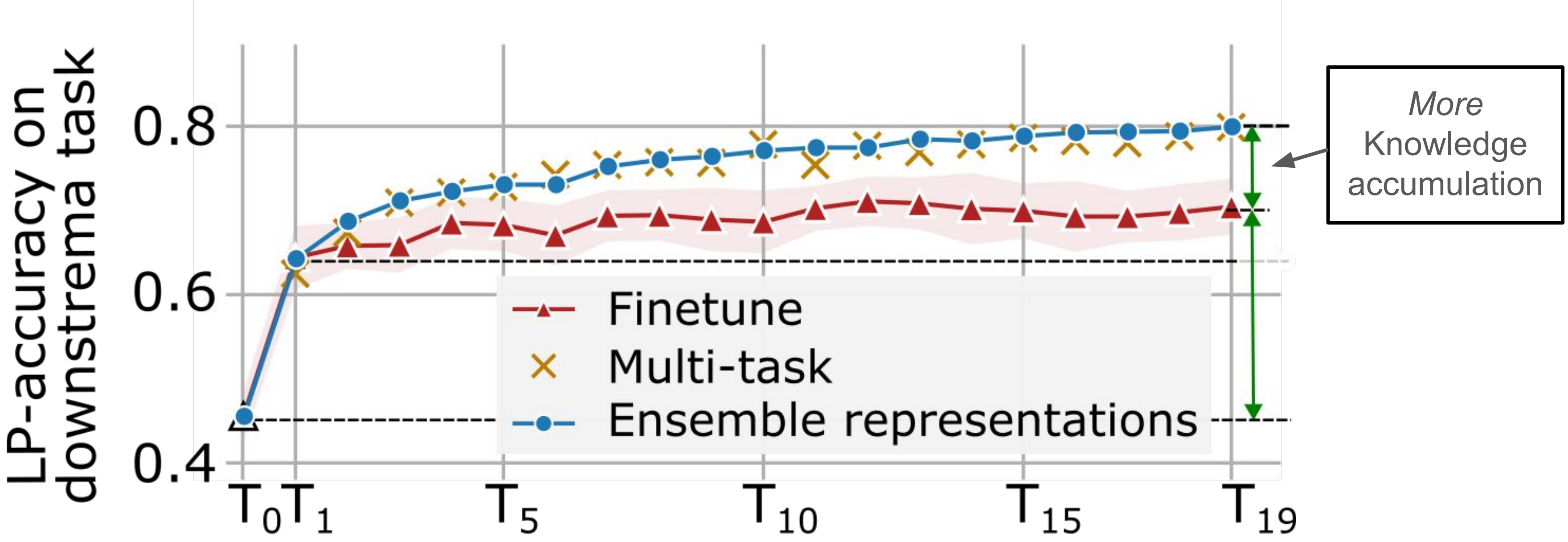
Is feature forgetting problematic?

- It has been argued representations only forget “task-specific” knowledge



Feature forgetting slows down knowledge accumulation

- If we keep everything the same, except we prevent forgetting, the amount of knowledge accumulation is substantially increased



Summary

- Representations do forget catastrophically
  - *Newly learned information* is forgotten similarly in the representation as at the output level
  - Uncovered by measuring forgetting in relative terms

- Such feature forgetting impairs knowledge accumulation
  - Demonstrated by using a representation ensembling baseline which learns in the exact way as fine-tuning, but does not forget

---

- For details: https://arxiv.org/abs/2304.00933

Funding acknowledgements

This project has been supported by funding from the European Union under the Horizon 2020 research and innovation program (ERC project KeepOnLearning, grant agreement No. 101021347) and under Horizon Europe (Marie Skłodowska-Curie fellowship, grant agreement No. 101067759).

Preprint on arXiv:
https://arxiv.org/abs/2311.14028

“Continual Learning of Diffusion Models with Generative Distillation”

Sergi Masip, Pau Rodríguez, Tinne Tuytelaars, Gido M. van de Ven

Diffusion models

- Powerful class of models
- Strong performance in many generative modelling tasks (*e.g.* image synthesis)

(source: [Ramesh et al., 2022])

- Training is very resource-demanding!

- It would be great if these models could be trained *continually*

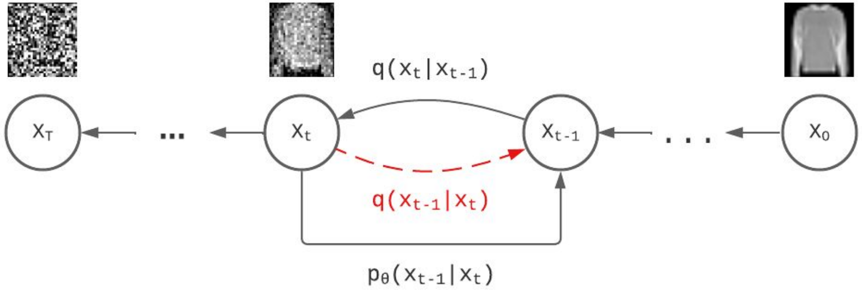
Continual learning of diffusion models

- Relatively unexplored

- A promising approach: generative replay
    - No need to store data
    - Replayed data will be diverse
    - Generative model is already available!

Generative replay for diffusion models

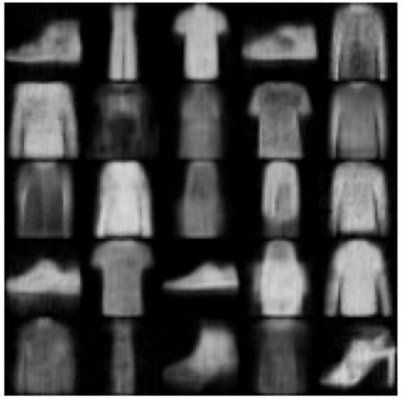
- Problem: sampling from a diffusion model is computationally expensive!



- Previous studies use limited number of samples, treat them as a replay buffer
  - Loses the benefit of diversity
  - Still need to store samples, still computationally costly
  - Disappointing performance (e.g., Zając et al., 2023; Smith et al., 2023)

Modification 1: use faster sampling techniques

- *Denoising Diffusion Implicit Models* (DDIM): permits sampling using a smaller number of denoising steps, trading computational efficiency for sample quality



DDIM (2 steps)

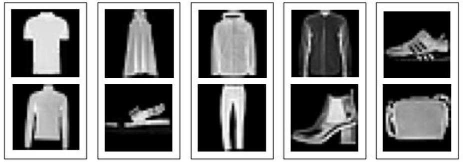
DDIM (10 steps)

DDIM (100 steps)

DDPM (1000 steps)

Standard generative replay with DDIM breaks down

Split Fashion MNIST

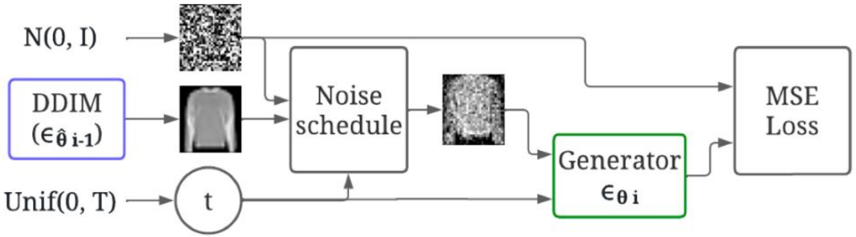


(a) Task 1 (b) Task 2 (c) Task 3 (d) Task 4 (e) Task 5

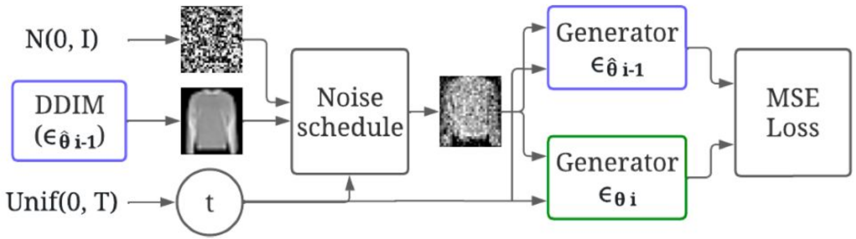
Modification 2: combine generative replay with distillation

- Observation: standard generative replay transfers knowledge only at the end point of the reverse process of the diffusion model

Generative replay



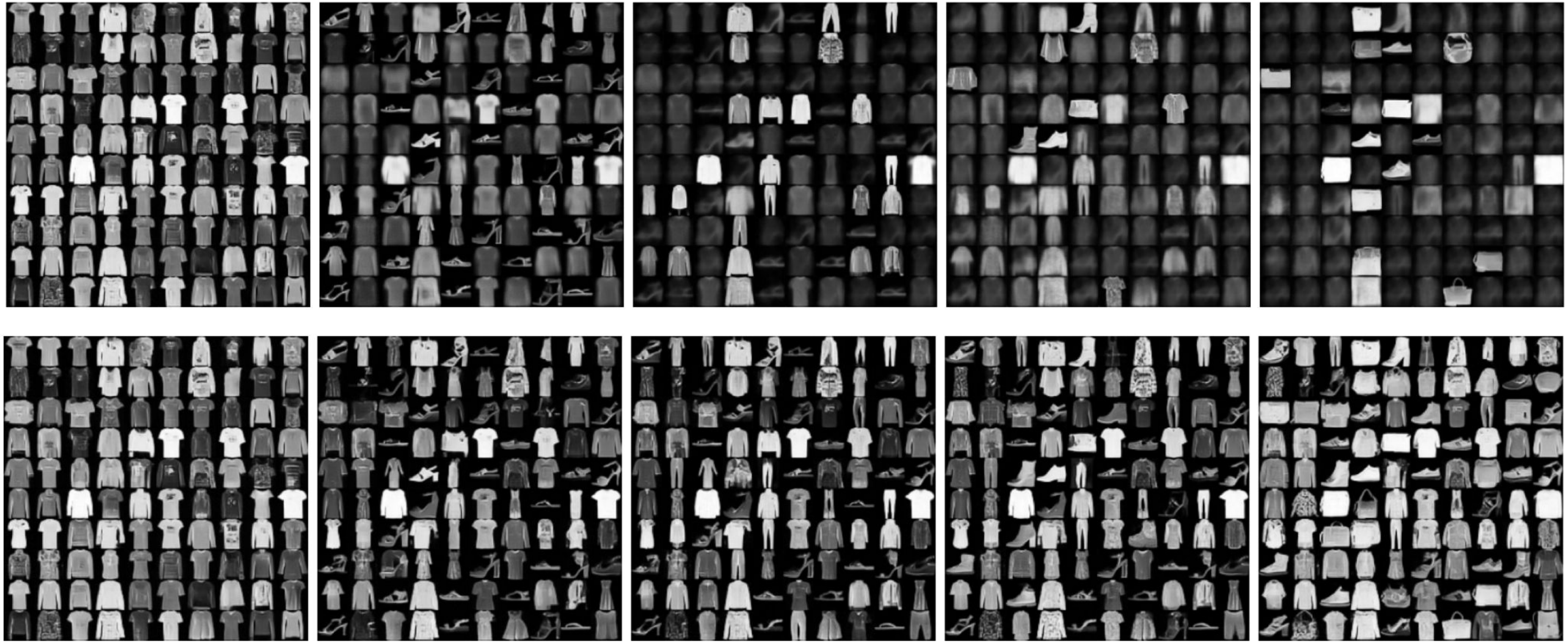
Generative distillation



- *Generative distillation* transfers knowledge *at every step* of the diffusion process

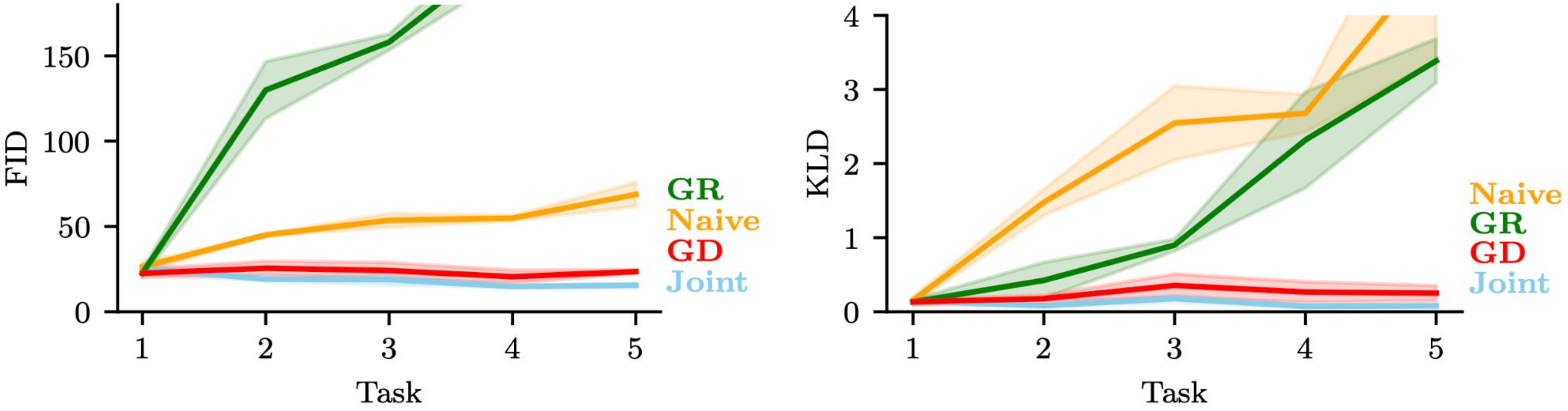
  (a similar distillation process is currently used to train a student to generate same quality images as its teacher but in fewer generation steps, e.g., Luhman and Luhman, 2021; Salimans and Ho, 2022)

Generative distillation markedly improves generative replay



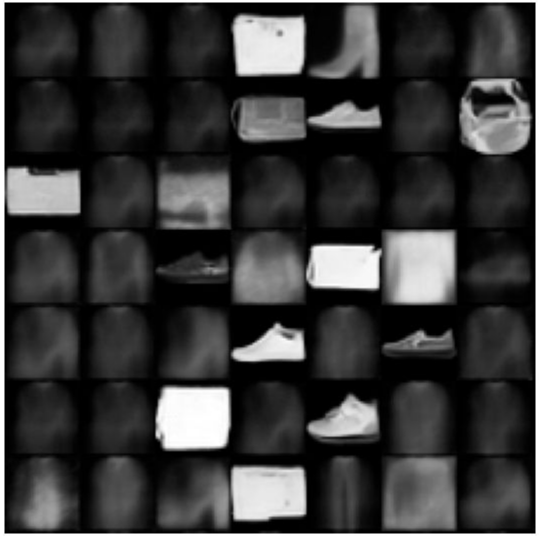
(a) Task 1 (b) Task 2 (c) Task 3 (d) Task 4 (e) Task 5

Generative distillation markedly improves generative replay



Generative distillation markedly improves generative replay

Joint training



Generative replay

Generative distillation

(after learning all 5 tasks of Split Fashion MNIST)

Summary

- Continually training a diffusion model with standard generative replay results in a catastrophic loss in its denoising qualities
- Including knowledge distillation into the generative replay process (i.e., *generative distillation*) mitigates this catastrophic forgetting and markedly enhances performance

---

- See the paper for details: https://arxiv.org/abs/2311.14028
- Code: https://github.com/Atenrev/difussion_continual_learning

Funding acknowledgements

This project has been supported by funding from the European Union under the Erasmus+ Traineeship program, under the Horizon 2020 research and innovation program (ERC project KeepOnLearning, grant agreement No. 101021347) and under Horizon Europe (Marie Skłodowska-Curie fellowship, grant agreement No. 101067759).

Erasmus+

erc

MSCA