Improving Feature Generalizability With Multitask Learning In Class Incremental Learning

IEEE ICASSP 2022

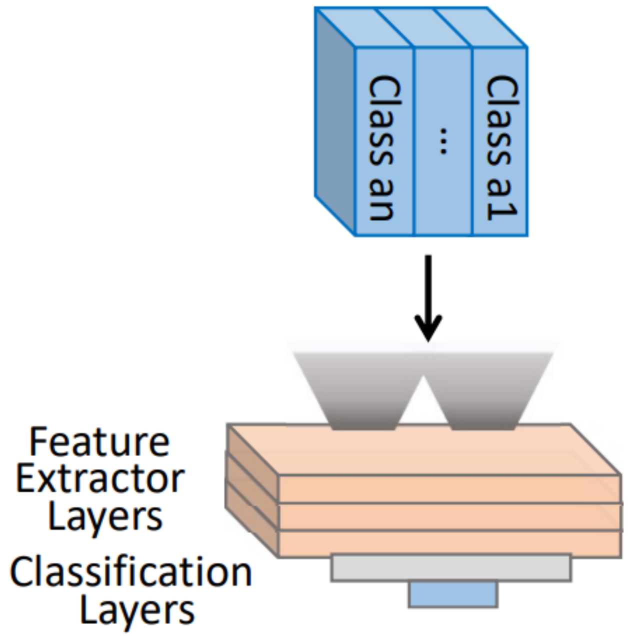
Dong Ma, **Chi Ian Tang**, and Cecilia Mascolo

dongma@smu.edu.sg, cit27@cl.cam.ac.uk

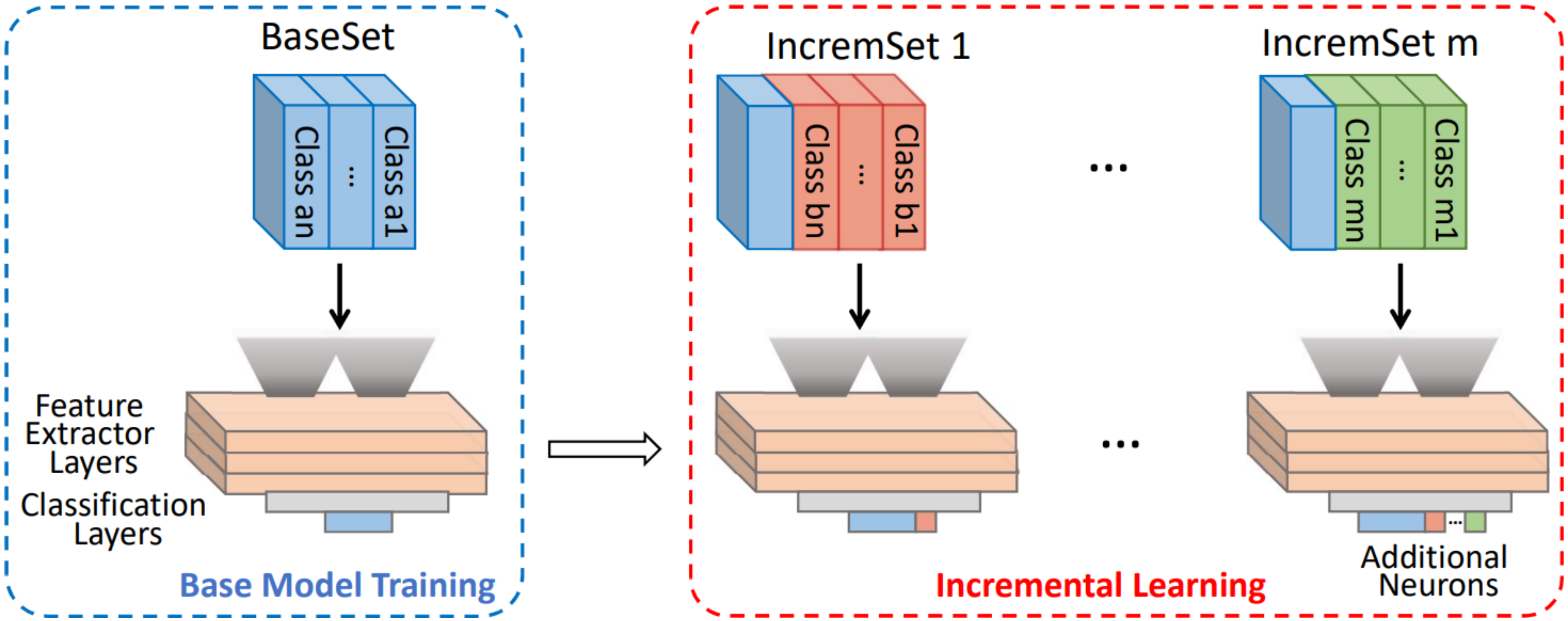
Background

Conventional Deep Learning

- Trained on fixed dataset
- Lacking flexibility of adapting to new data



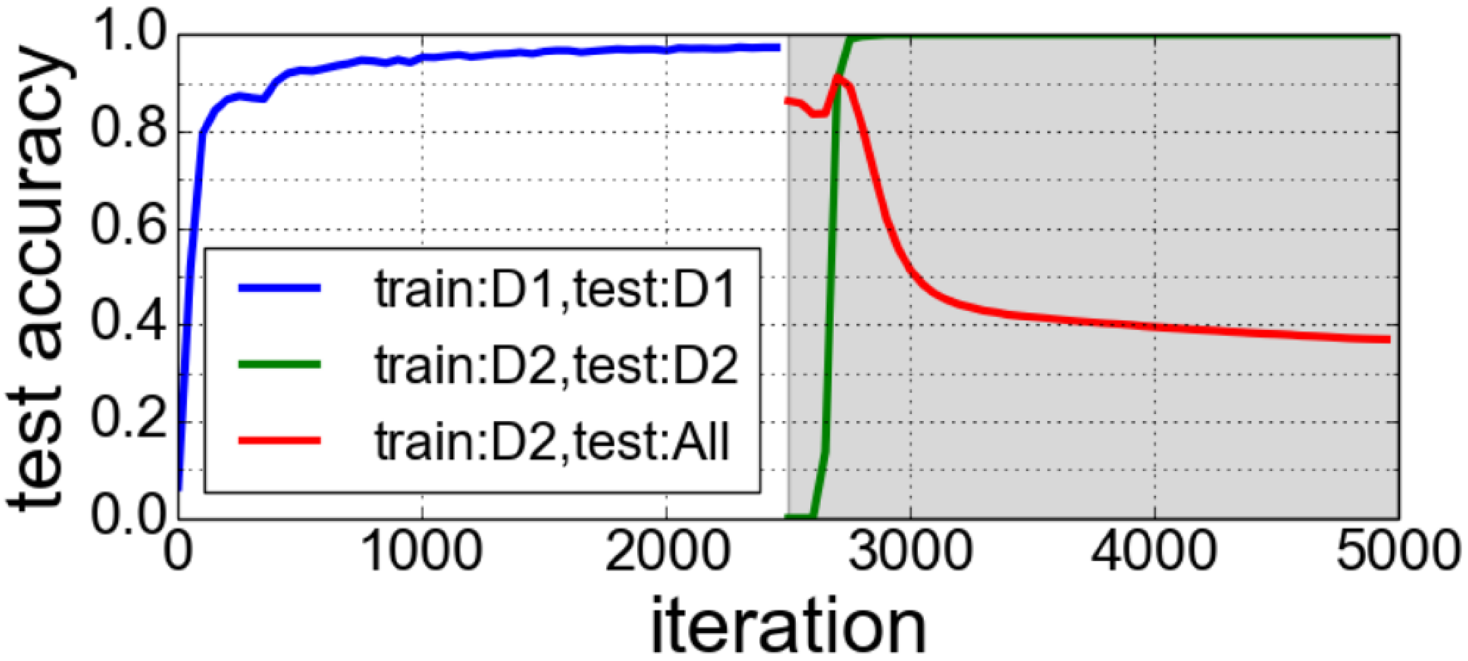
Class Incremental Learning, CIL



- Retain the acquired knowledge while learning new concepts
- 2 stages: Base Model Training, Incremental Learning

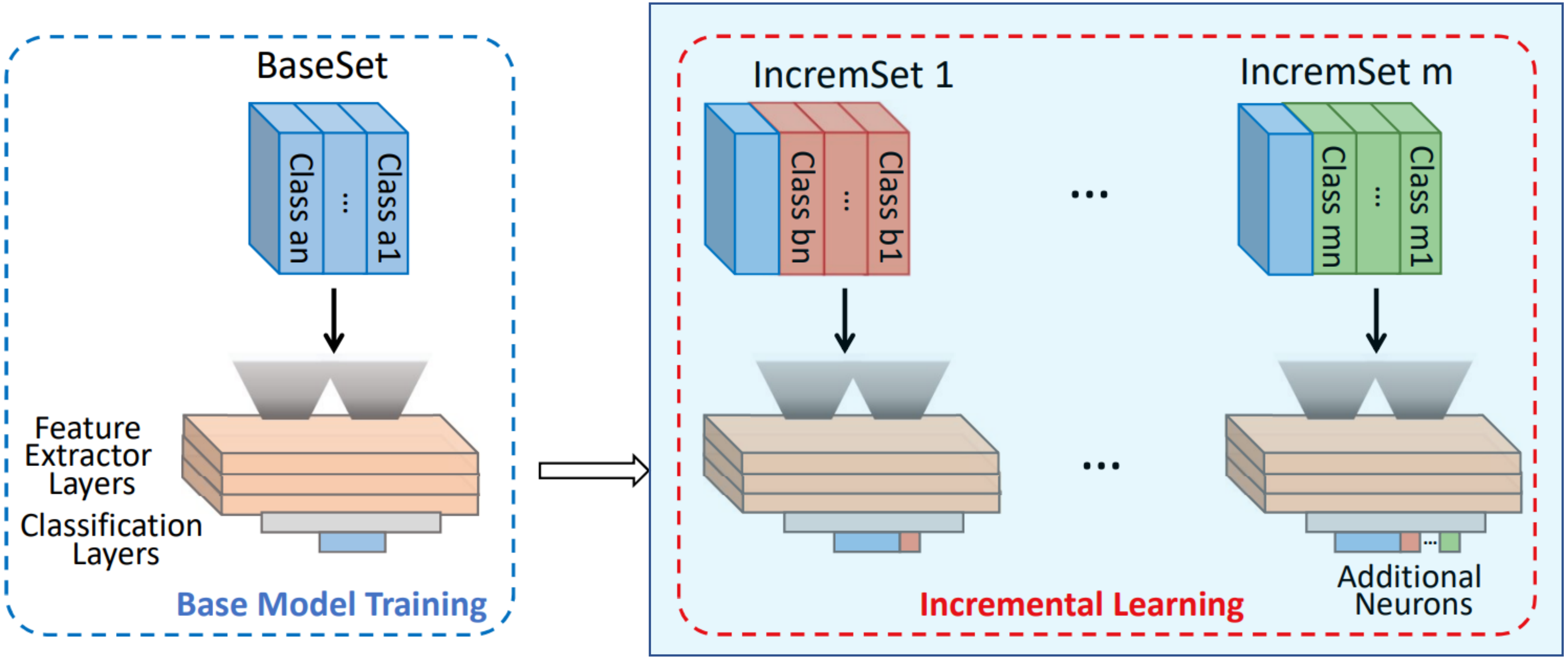
Catastrophic Forgetting in CIL

- Overfitting to new data
    - Imbalance between the old and new class data



[Pfülb, 2018.]

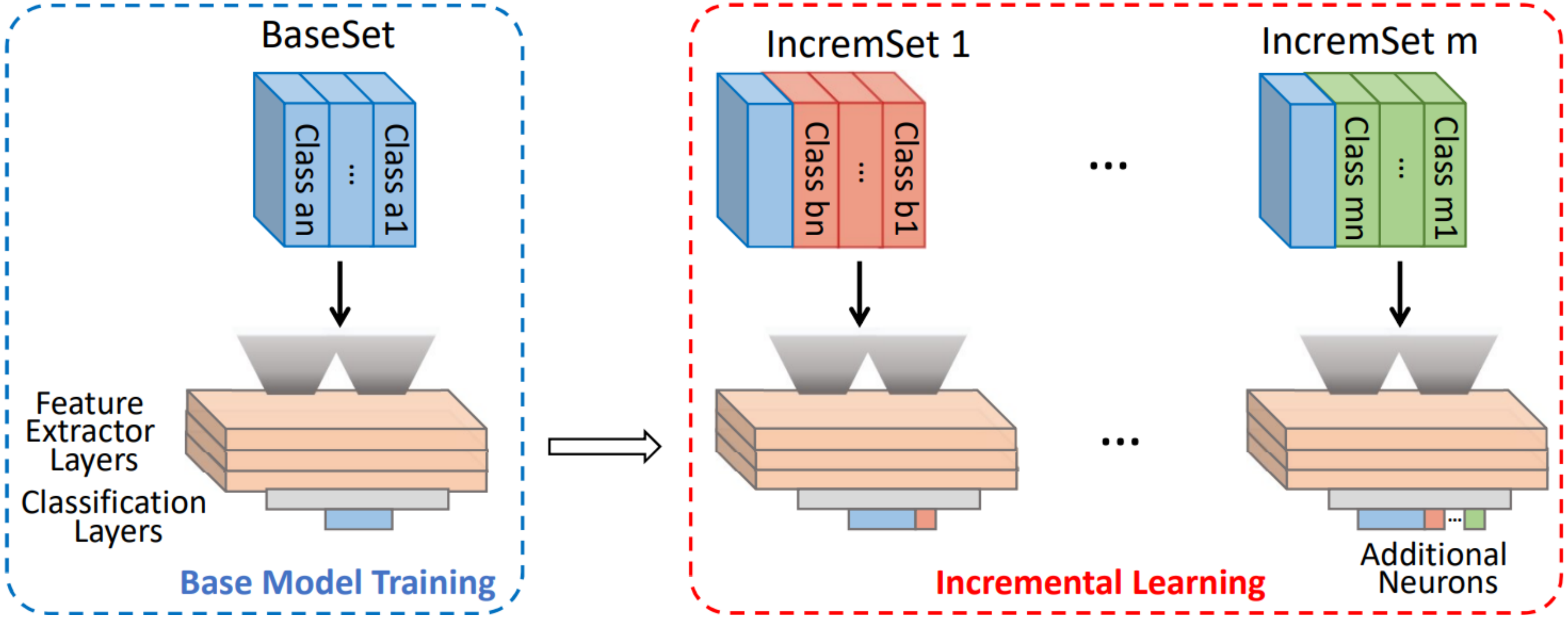
Existing Solutions



- Existing approaches mainly focus on the incremental learning stage

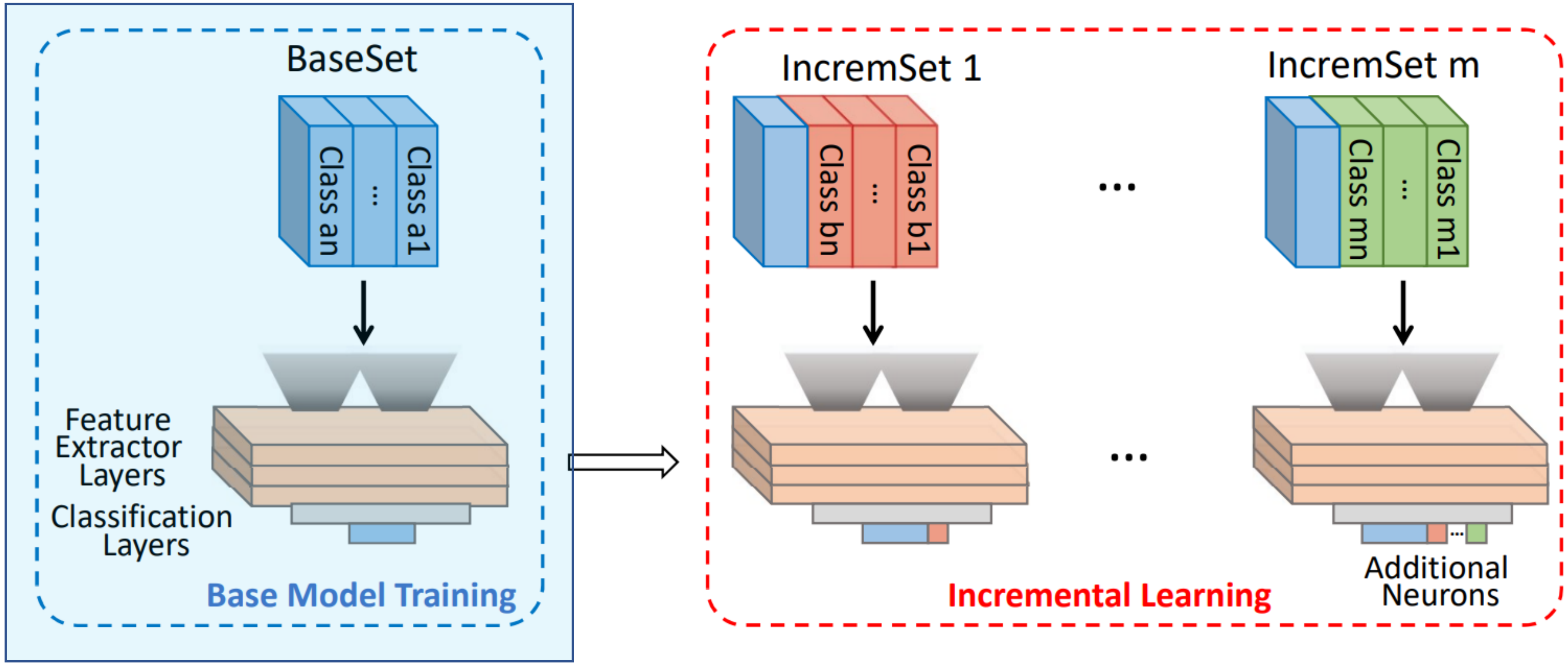
Observation

- Different initial seed results in different weights
- Which set of weight is better for CIL?



Hypothesis

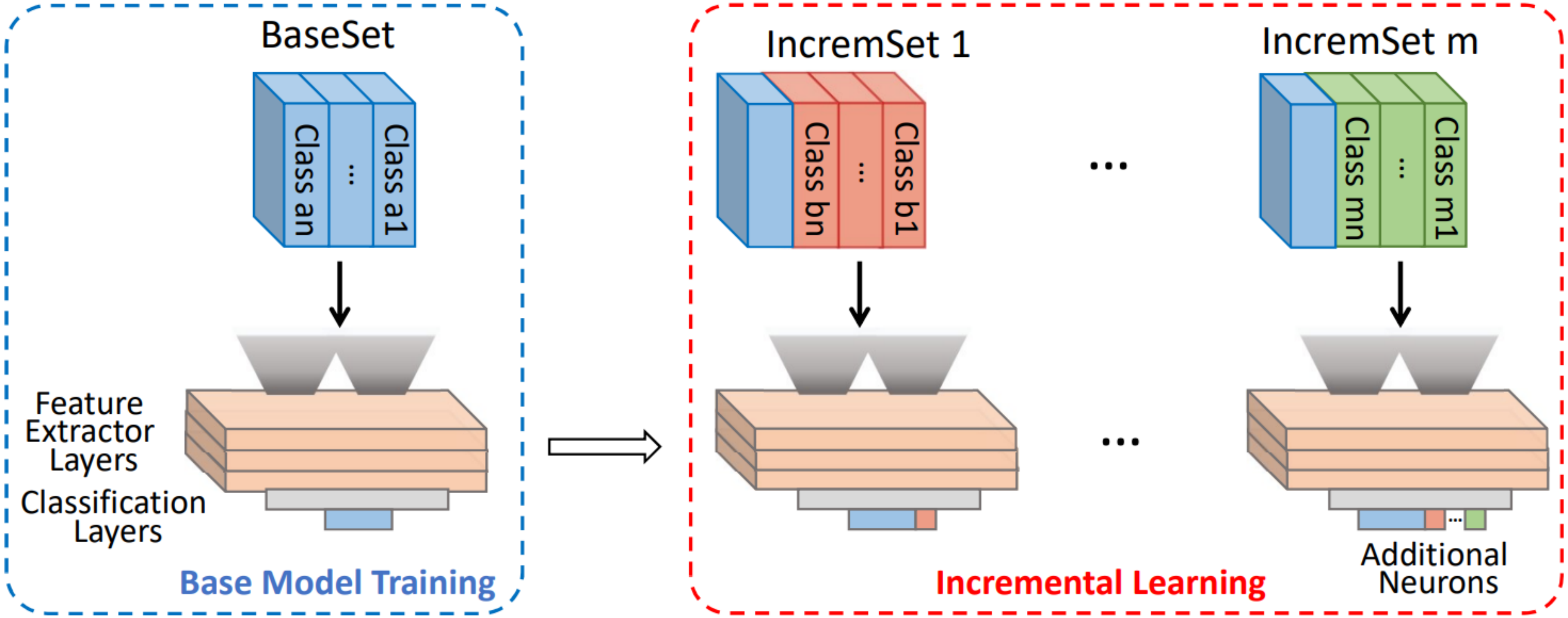
It is possible to alleviate catastrophic forgetting by training a more transferable base model



Approach

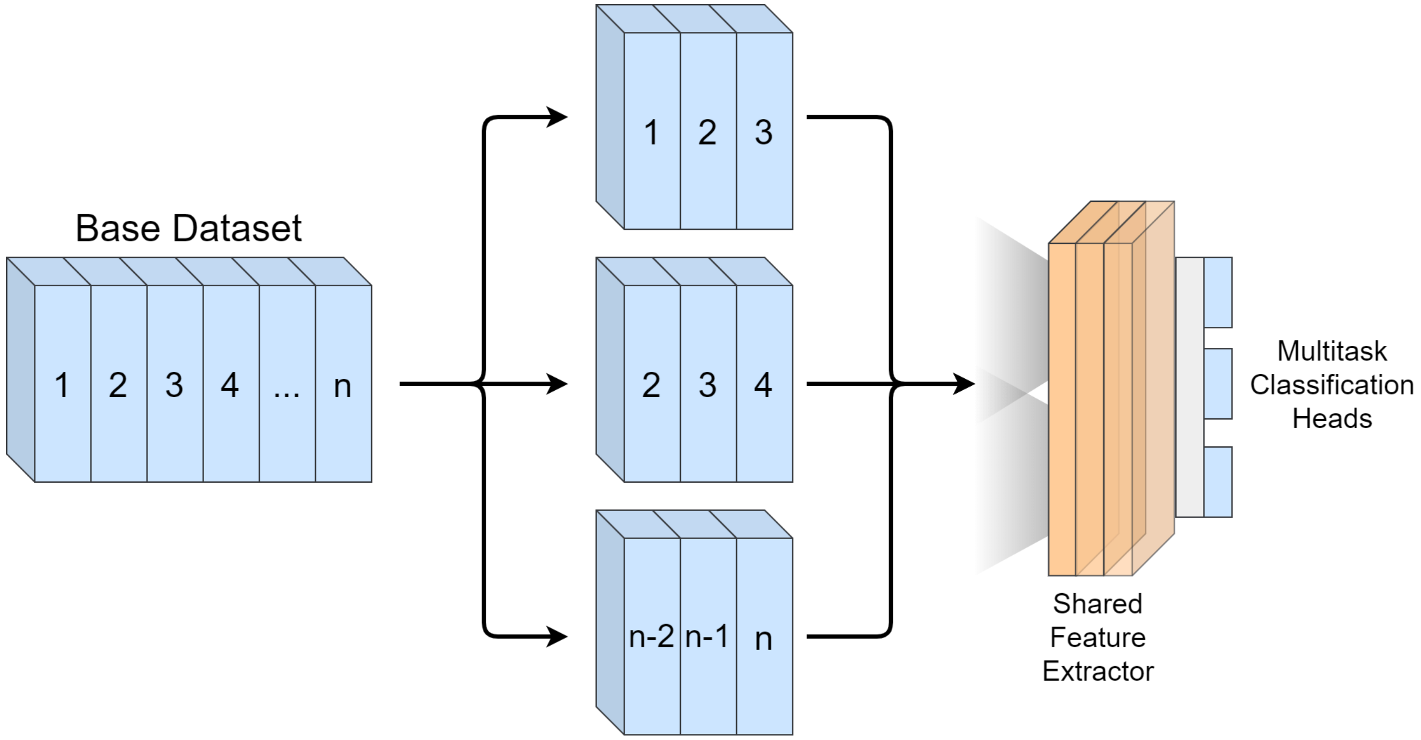
Multitask Learning - Intuition

- CIL requires model to retain previous knowledge
- Idea: Simulate incremental learning



Multitask Learning

- Decompose the base task
- Trained with a shared backbone
- Find weights which can solve all tasks at once



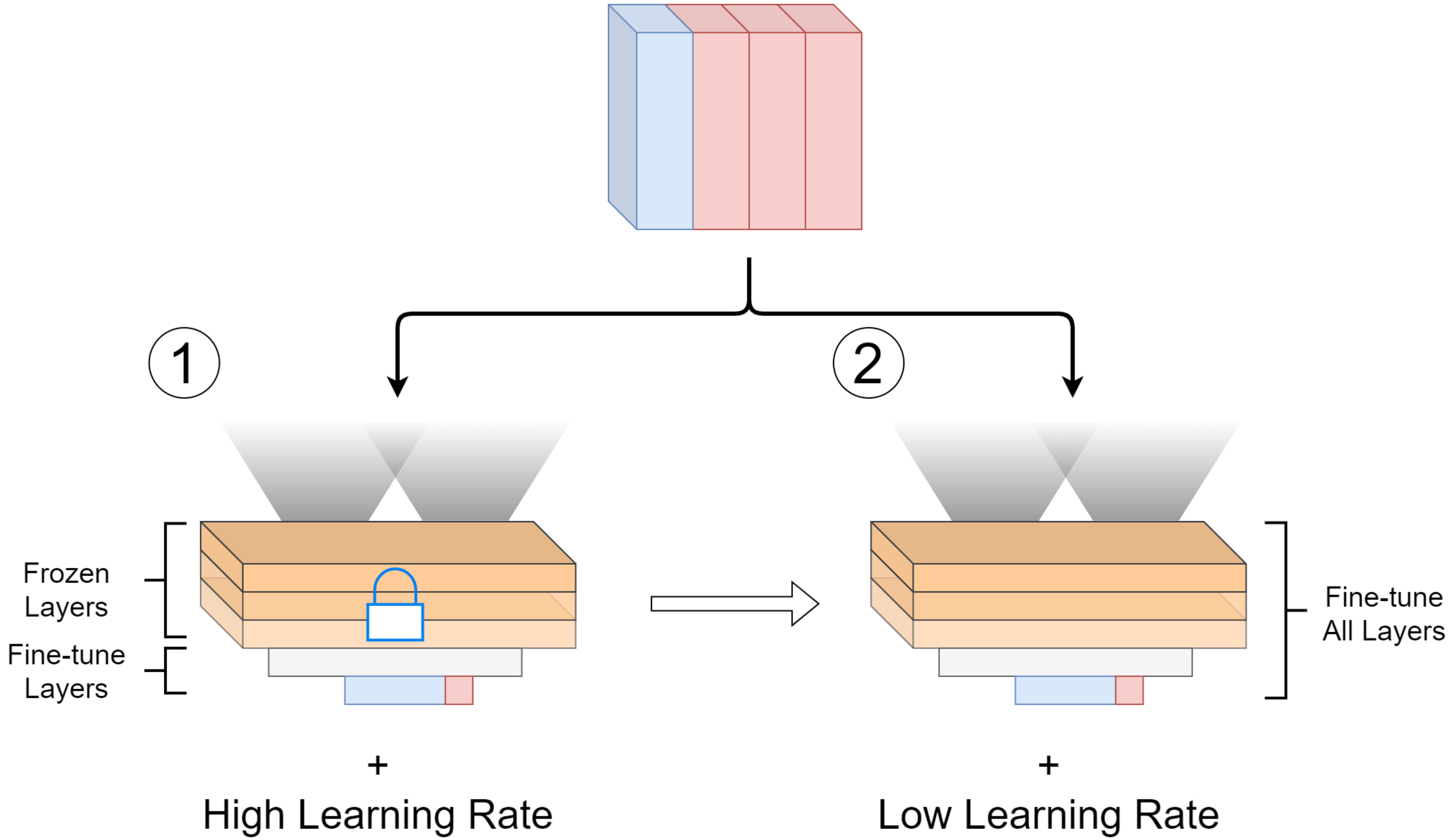
Selection of Tasks

- Many valid sub-tasks $N \mapsto 2^N - 1$
- Difficulty $\leftrightarrow$ Diversity

- Explored along 2 directions
    1. Number of classes
    $\{1,2,3,4,5\} \to (\{1,2\},\{2,3,4\},\{1,2,3,4,5\})$
    2. Subset of classes
    $\{1,2,3,4,5\} \to (\{1,2,3\},\{2,3,4\},\{1,2,4\})$

Fine-tuning Strategy

- High learning rate → large changes in weights
- 2-step fine-tuning strategy during incremental learning:



Evaluation

Datasets

- UrbanSound8K
  - 10 environment sound events
  - Split into 4 (base), 2, 2, 2 classes
- Google Speech Commands (GSC)
  - 20 core keyword classes
  - Split into 5 (base), 3, 3, 3, 3, 3 classes

Baseline

- Previous state-of-the-art results by Mittal et al. (2021)
- Cross Entropy (CE) + Knowledge Distillation (KD)
- Balanced exemplar set
- We only changed base model training

Essentials for Class Incremental Learning

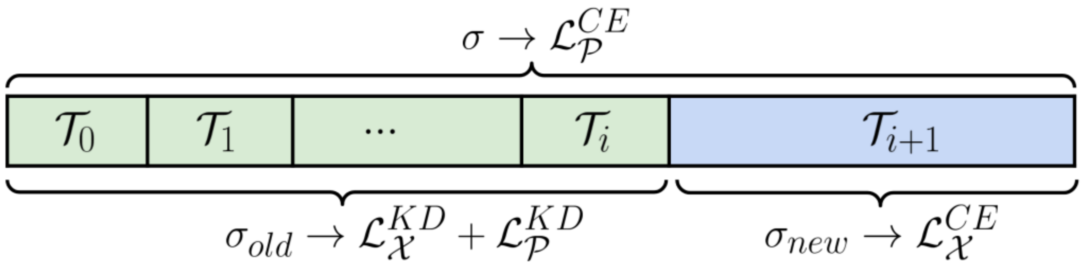
Sudhanshu Mittal Silvio Galesso Thomas Brox
University of Freiburg, Germany
mittal,galessos,brox@cs.uni-freiburg.de

Abstract

Contemporary neural networks are limited in their ability to learn from evolving streams of training data. When trained sequentially on new or evolving tasks, their accuracy drops sharply, making them unsuitable for many real-world applications. In this work, we shed light on the causes of this well known yet unsolved phenomenon – often referred to as catastrophic forgetting – in a class-incremental setup. We show that a combination of simple

classes, a forgetting constraint to keep previous knowledge while learning new tasks, and a learning system that balances old and new classes. Although several methods have been proposed to address each of these components, there is not yet a common understanding of best practices.

Prabhu *et al.* [22] provides an overview over current state of continual learning methods for classification. It shows that a simple greedy balanced sampler-based approach (GDumb) can outperform various specialized formulations in most of the continual learning settings, how-



[Mittal, 2021.]

Task Selection

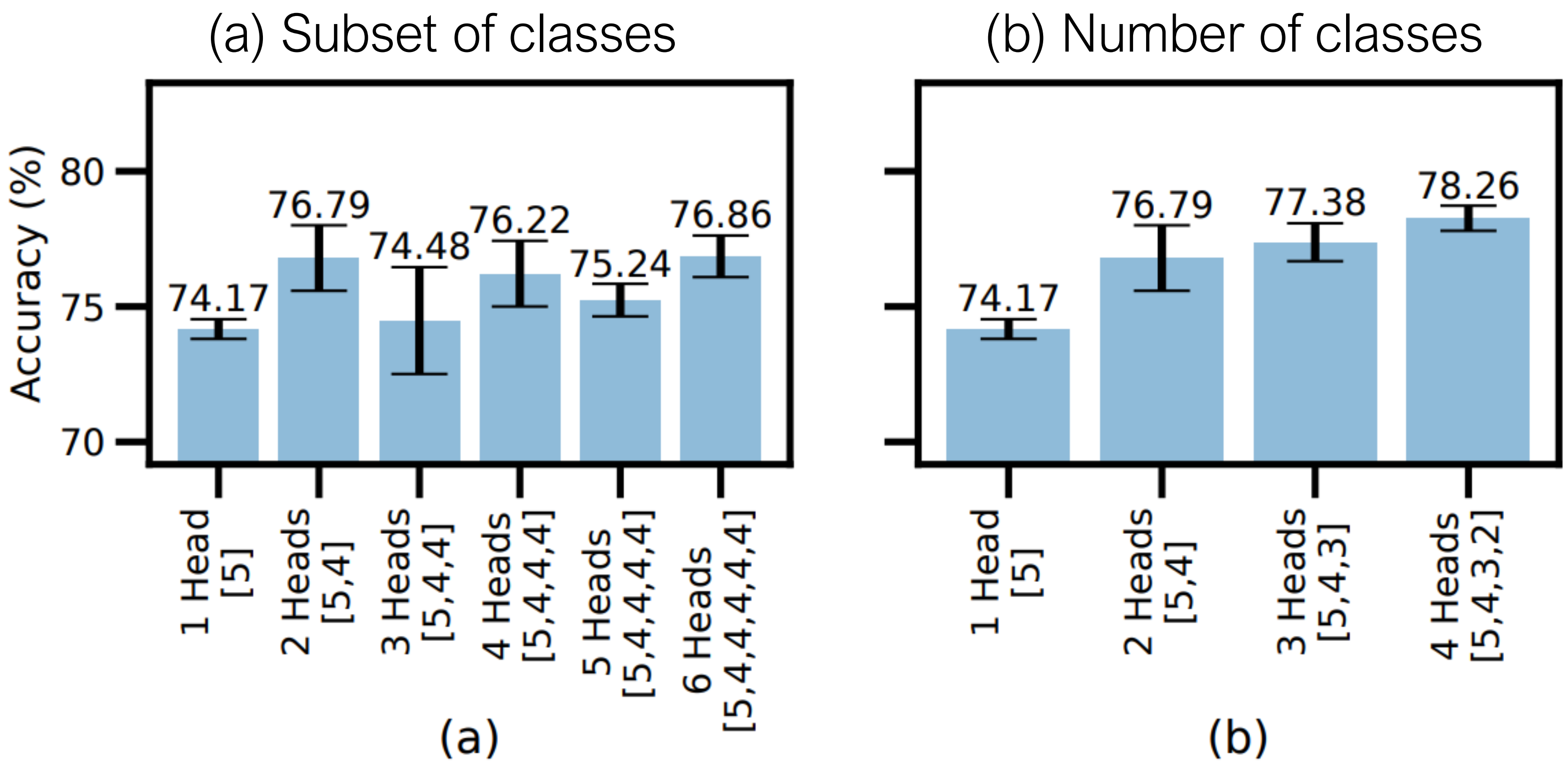
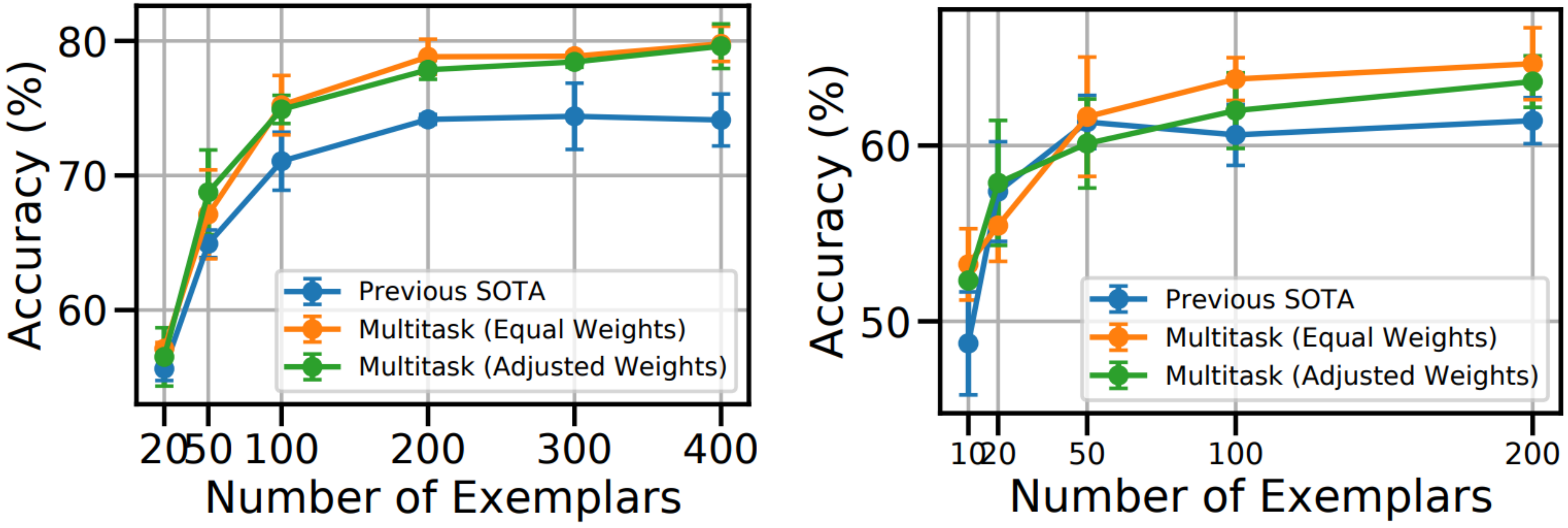


Fig. 4. Impact of task creation.

More diverse → Higher generalizability

Number of Exemplar Comparison to SOTA



(a) GSC (b) UrbanSound8K

Fig. 5. Impact of exemplar quantity.

More exemplars → Higher Performance

Multitask > Non-multitask base training

Effect of Losses

Table 2. Effect of different losses on the incremental learning performance. CE refers to Cross Entropy, KD refers to Knowledge Distillation, N refers to New samples, and O refers to Old samples (exemplars).

# of class	5	8	11	14	17	20	Avg
CE_N	96.97	60.23	43.79	35.73	29.46	26.22	**39.09**
CE_N+KD_N	97.08	60.75	43.34	37.43	36.20	34.85	**42.52**
CE_N+CE_O	97.25	88.77	79.07	72.88	72.82	57.27	**74.16**
CE_N+CE_O +KD_N	96.78	84.65	78.27	77.91	73.60	72.55	**77.39**
CE_N+CE_O +KD_O	97.12	85.72	80.64	78.99	74.30	71.93	**78.32**
CE_N+CE_O+ KD_N+KD_O	97.35	87.92	81.47	77.66	73.80	73.27	**78.82**

Improvement comes from use of exemplar

Knowledge distillation has limited effect

Conclusion

1. Hypothesis: more transferable feature representation might be beneficial to CIL
2. Introduced multitask learning to the base model training
3. Improves average incremental accuracy by up to 5.5%
4. Opens the door to improving the quality of base model in incremental learning

References

- Pfülb, B., Gepperth, A., Abdullah, S., & Kilian, A. (2018, October). Catastrophic forgetting: still a problem for DNNs. In *International conference on artificial neural networks* (pp. 487-497). Springer, Cham.
- Mittal, S., Galesso, S., & Brox, T. (2021). Essentials for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3513-3522).

More Information

Chi Ian Tang

cit27@cl.cam.ac.uk

iantangc.github.io

Dong Ma

dongma@smu.edu.sg

dongma.info