A System 1 and System 2 Perspective on Continual Learning for Practical Implementation

Vivek Chavan*[1,2], Oliver Heimann[1], and Jörg Krüger[1,2]

[1] Fraunhofer IPK
[2] Technical University of Berlin, Germany

Abstract. Deep Learning based approaches have led to several breakthroughs in Computer Vision (CV) research. However, catastrophic forgetting and the plasticity-rigidity dilemma remain a fundamental obstacle in our path to machine intelligence. Continual Learning (CL) research aims to address this issue through replay, model augmentation, regularisation and other techniques. In this paper, we analyse CL using the Dual Process theory and related frameworks. We extend concepts and theories from studies on human reasoning and learning to ongoing CL developments. We argue that Continual Learning is a System 2 problem, and derive insights about the practical application of existing approaches. We develop our hypotheses and verify their validity on publicly available and varied datasets. Our approach involves encoding contextual semantic information and robust pretraining. Experiments show that these techniques can improve model performance under diverse scenarios. The code is available at: https://github.com/Vivek9Chavan/ADCM

Keywords: Continual Learning · System 1 · System 2 · Explainability · Pretraining

1 Introduction

The recent advancements in Computer Vision (CV) and pattern recognition have been fuelled by the rapid adoption of Neural Network (NN) based architectures for [22, 28, 36]. This has been possible due to the availability of large amounts of data [21, 57] and compute [2, 9]. Despite this, *Catastrophic Forgetting* remains a fundamental problem in NN-based architectures when retrained on new data after their initial training [25, 26].

Continual Learning (CL) or *Incremental Learning* (IL) is a growing area of research and has introduced several novel approaches for mitigating this issue [19, 45]. The goal is to reduce catastrophic forgetting (*improve model rigidity*) while enabling learning from new data (*improve model plasticity*). Empirically, we define CL as a Machine Learning (ML) problem, where the model $f(D_i)$

*Correspondence: vivek.chavan@ipk.fraunhofer.de

has to learn from new incoming tasks ($t = 0, 1...i...N$), which introduce new streams of data ($D_0, D_1, ...D_i, ..D_N$) such that the model performs well on all data $\sum_{t=0}^{i} D_t$ at a given task i. Real-world adoption and implementation of CL approaches have been steadily growing [17, 67].

In this paper, we approach CL from the lens of System 1 and System 2 thinking [30, 31]. While we intend this to be a loose analogy, we observe that CL is inherently a System 2 ML problem, and the current solutions often aim to mimic System 2 behaviour in NN-based models explicitly.

2 Related Work

Continual Learning. Catastrophic Forgetting and plasticity-rigidity trade-off have been studied since the adoption of NN architectures [25]. Van de ven et al. [63] generally categorise IL into three categories based on the new incoming streams of data, viz., Task, Domain and Class-IL (CIL). The latter introduces new classes or object groups during CL and is considered more challenging. Methods such as EWC [32] and packNet [43] are employed for Task and Domain-IL. Rebuffi et al. [53] introduced a Nearest Mean Exemplars (NME) strategy for CIL, along with Herding, an approach for selecting samples for memory replay during retraining. Several other approaches have been proposed for CIL in Convolutional Networks (CNNs), including UCIR [29], Weight Aligning [72], PODNet [23], DER [69], AANet [40], FOSTER [64], MEMO [75], among others. Transformer-based [22, 61] IL approaches have also been suggested [24]. Several IL methods utilise Knowledge Distillation [20] based loss in conjunction with the traditional cross-entropy loss to retain learning from previous tasks [7, 45]. Additionally, regularisation-based techniques and model augmentation are also employed. Approaches such as Mnemonics [42], RMM [41] have been proposed as an alternative to herding, for selecting exemplars. Newer methods also implement CIL without rehearsal memory [6, 51, 76], which employ an ensemble of strategies for effective representation learning. Certain approaches such as DER and FOSTER are able to retain learning from old tasks better and struggle with learning new tasks. Conversely, iCaRL and PODNet learn new data more easily and struggle with previous tasks [17].

System 1 and System 2. There are several parallels between the current advancements in Artificial Intelligence (AI) and Cognitive Psychology [8, 55]. Analysing human judgement and choice is an active and interdisciplinary area of research [1, 30]. The Dual Process Theory [16] studies thinking and decision-making as an interplay between *System 1* and *System 2* processes. System 1 is often fast, intuitive, automatic and based on heuristics. Whereas, System 2 is rational, slow, and involves careful and *attentive* processing of the data. In practice, the two processes work in tandem, and System 2 typically operates with System 1 functioning in the background [31].

Machine Learning and the Dual Process Theory. It is hypothesized that the current paradigm of ML largely resembles System 1 behaviour [27]. There is a growing focus on approaches capable of System 2 actions [38, 66].

The *attention mechanism* mimics cognitive behaviour of focusing on relevant aspects of external stimuli [48, 68]. Strategies such as chain-of-thought prompting [65] encourage *reasoning* in Large Language Models (LLMs) [12]. Ongoing research on the development of World Models by pairing vision and language shows promising results [11,39]. However, such innovations currently require substantial compute and large-scale training data [38]. Novel approaches attempt to address this issue by combining Deep Learning with techniques such as Tree Search [3], planning [59] and Symbolic AI [10, 27].

3 Methods

Continual Learning as a System 2 Problem. CL investigation is often influenced by the evolutionary ability of humans and animals to adapt and actively react to changing stimuli (data) [62]. Concepts from Hebbian Learning, Complementary Learning Systems (CLS) theory [46] and the Dual Process Theory have inspired recent works [4, 52, 71]. Catastrophic Forgetting is a consequence of the fast and unconscious (System 1) processing of data by neural networks. One of the most common approaches for preserving the learning from previous tasks in IL is *replay*. This may be via rehearsal memory storage [53], generative replay [58] or indirectly processed via regularisation [13]. This is similar to the cognitive steering model in psychology, where effortful, associative simulation (System 2) aids in aligning new incoming data with remote memory [37]. De Neys [47] showed that System 1 is automatic and works independently of working memory. However, System 2 processes are impeded by the reduction in memory. In the absence of a working memory, System 1 takes over, which results in self-bias. NN-based models exhibit similar behaviour when learning continually [25, 49]. A significant difference between biological cognition and conventional ML is that humans process information multimodally, with a rich context that aids the visual input [60]. Moreover, research suggests, that humans have two memory representations: *verbatim* traces include exact and literal surface information, such as words, and *gist* traces contain representations of an event's semantic features [54]. It is theorised that verbatim and gist information are stored simultaneously in a parallel fashion [44]. Although the Fuzzy Trace Theory was developed for studying verbal information and reasoning, we extend this concept to processing visual information in our work.

For ML, the verbatim data include image representations, e.g., as stored exemplars (for CIL). We conjecture that incorporating semantic information could aid the model in learning continually. Such context-based information can be extracted via (Class Activation Maps) CAMs immediately after the task is introduced. This approach replicates a System 2 adjacent slow and effortful operation, where the meaningful explanations extracted from the model aid the deliberate processing of the data. This approach requires storing additional data in the form of class activation masks. This increases the required storage and computational requirements. The CAM data could be stored as pixel-wise single-

channel images with the same resolution as the original images or as lower-resolution images (e.g. by averaging the intensity values across patches).

Encoding Contextual Semantic Information. To address catastrophic forgetting, we encode contextual (gist) information via the ground truth masks, in the form of a weighted loss function, $\mathcal{L}_{\text{gist}}$, which encourages the model to *focus* on the important regions in the image, as shown in Equation 1.

$$\mathcal{L}_{\text{gist}} = \frac{\lambda}{N} \sum_{n=1}^{N} \left(\frac{1}{\sum M^{(n)}} \sum_{i} M_i^{(n)} \cdot \mathcal{L}(\hat{y}^{(n)}, y^{(n)}) \right) \tag{1}$$

The total loss for a batch of N examples is the mean of individual losses. Here, $\mathcal{L}(\hat{y}^{(n)}, y^{(n)})$ is the cross-entropy loss for the n-th example, and $\sum M^{(n)}$ is the sum of weights for the n-th example, normalising the weighted loss. The mask, which indicates the region of interest, is denoted as M where M_i is the weight for the i-th pixel. In this case, the mask can be considered as a weight vector where each entry corresponds to a certain region's importance in the image. The scalar λ is a hyperparemeter.

Grad-CAM [73] produces a visualization map CAM_{ij}^c by applying a ReLU to the weighted sum of feature maps A^k, where weights α_k^c are computed as the global-average-pooled gradients of the class score y^c with respect to A^k:

$$CAM_{ij}^c = \text{ReLU} \left(\sum_k \left(\frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \right) A_{ij}^k \right) \tag{2}$$

Here, Z denotes the number of spatial locations in the feature maps, and (i, j) indexes spatial locations. The Grad-CAM, CAM_{ij}^c, is then a weighted combination of feature maps, followed by a ReLU function to only retain positive influences. To convert CAM^c into a usable mask M, we normalise the CAM output.

$$M = \frac{CAM^c - \min(CAM^c)}{\max(CAM^c) - \min(CAM^c)} \tag{3}$$

M is then processed to compute $\mathcal{L}_{\text{gist}}$, as discussed. The gist loss is appended to other terms based on the implementation (e.g. the POD distillation loss [23]).

Pretraining Schemes and Transfer Learning Although academic CIL setups involve learning on data from scratch [53], real-world implementation benefits from transfer learning. When comparing the pretraining schemes for CIL, we observe that Self-Supervised pretraining (DINO [15] and VICRegL [5]) outperforms supervised pretraining given the same model (ResNet50) and pretraining dataset (ImageNet1k [21]). SSL has been shown to produce robust and generalised feature representations [15, 18, 50]. A suitable analogy is a Chess Grandmaster [31], who develops deeper and generalised pattern recognition skills compared to a novice. As a result, the Grandmaster can make decisions instinctively without significant effort and adapts to changing situations more easily.

Algorithm 1: Incremental Learning with Contextual Information via Replay

Input: Dataset D, Replay Memory R, Model, Hyperparameters

for $t = 0$ **to** N **do**
  if $t = 0$ **then**
    if *masks available* **then**
      $\mathcal{L}_{\text{gist}} \leftarrow$ use Equation 1 `// Gist loss for initial task`
    else
      $\mathcal{L}_{\text{CE}} \leftarrow \mathcal{L}(\hat{y}^{(n)}, y^{(n)})$ `// Cross-entropy loss`
    $R \leftarrow \text{SelectExemplars}(D_t)$ `// Select and store exemplars`
  else
    if *no masks available* **then**
      for *images in* R **do**
        $\alpha_k^c \leftarrow \frac{1}{Z} \sum_i \sum_j g_{y^c}^k(i,j)$ `// Importance weights`
        $CAM_{ij}^c \leftarrow$ use Equation 2 `// Grad-CAM`
        $M^{(n)} \leftarrow$ use Equation 3 `// Normalize Grad-CAM`
        $R \leftarrow R \cup \{M^{(n)}\}$ `// Update replay memory`
      $\mathcal{L}_{\text{gist}} \leftarrow$ use Equation 1 for img in R `// Gist loss`
      $\mathcal{L}_{\text{CE}} \leftarrow \mathcal{L}(\hat{y}^{(n)}, y^{(n)})$ for D_t `// Cross-entropy loss`
    else
      $\mathcal{L}_{\text{gist}} \leftarrow$ use Equation 1 for $D_t + R$ `// Gist loss`
  $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{gist}} + \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{other}}$ `// Total loss`
  $R \leftarrow \text{SelectExemplars}(D_t)$ `// Update exemplars`

4 Experiments

Setup. We use the FACIL [45] and PyCIL [74] libraries for our experiments. A consistent set of hyperparameters are used for all tests. We use ResNet18 as the base architecture for Experiments 1 and 2, and ResNet50 for Experiments 2 and 3. Herding is employed for exemplar selection for replay. All experiments were performed on the same workstation (details available in Supplementary Material).

Experiment 1: Encoding Contextual Information for ImageNet [21] and InVar-100 [17] datasets. We incorporate the gist information via CAMs (Equations 2 and 3) during experience replay for older tasks. Table 1 gives the results and Fig. 1 shows the plot for the two CIL approaches.

Experiment 2: Pretraining Schemes and Transfer Learning. Pretrained ResNet50 models were incrementally trained on different datasets with the same setup as §3. Table 2 shows a summary of the results for different implementations. Fig.2 shows the top-1 accuracy results for PODNet with varying task sequences as a challenging scenario. A clear separation between old class results can be seen for DINO and VICRegL pretrained models.

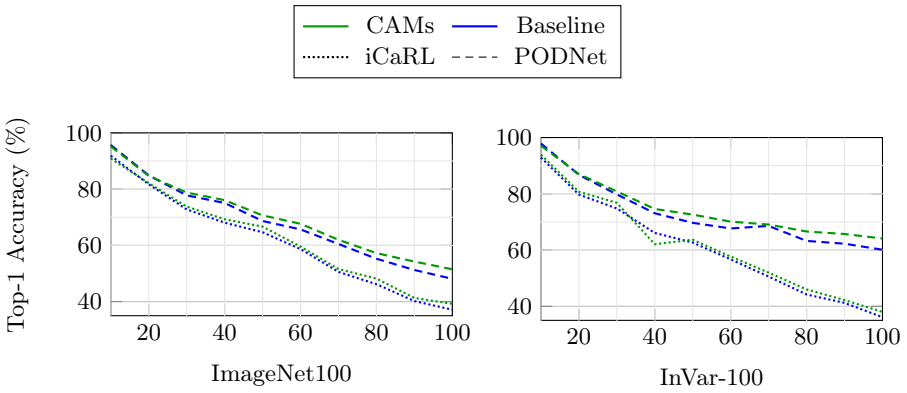

Fig. 1: Top-1 Accuracy vs. total number of classes for the two CIL approaches on the ImageNet100 (left) and InVar-100 (right) datasets: 10 tasks, 20 allowed exemplars per class. CAMs is when the gist information is derived from class activations and provided for old data only. Baseline is the standard training. Further details are in Table 1. The results show clear improvements from adding contextual information.

Table 1: A comparison of the three approaches on encoding contextual information; with CAMs and Baseline for **ImageNet100 (Top)** and **InVar-100 (Bottom)** datasets. Δ_{old} and Δ_{new} are average changes in the incremental accuracy on old and new tasks, respt. The results confirm that adding Gist information can help in improving model plasticity as well as rigidity. The table shows average incremental accuracy values (%).

Approach	**#exemplars**	**iCaRL**		Δ_{old}	Δ_{new}	**PodNet**		Δ_{old}	Δ_{new}
	(per Class)	5 tasks	10 tasks			5 tasks	10 tasks		
CAMs	20	54.7	52.8	+1.0	+0.2	67.84	64.21	+2.1	-0.1
Baseline	20	53.6	51.9	0	0	65.21	62.89	0	0
CAMs	20	59.12	52.32	+1.3	-0.2	74.88	71.16	+1.7	+0.1
Baseline	20	57.24	51.20	0	0	70.81	68.80	0	0

Table 2: Average Incremental Accuracy (%) for different datasets using ResNet50 pretrained on ImageNet1K using Supervised Learning, DINO and VICRegL. SSL pretrained models (DINO and VICRegL) outperform supervised models in each scenario. Details of old and new class performance for some cases are shown in Fig. 2

	Supervised			**DINO** [15]			**VICRegL** [5]		
	PODNet	DER	iCaRL	PODNet	DER	iCaRL	PODNet	DER	iCaRL
CIFAR-100 [35]	48.04	47.21	32.74	**50.70**	**48.54**	**36.14**	47.32	48.01	35.93
Stanford Cars [34]	67.03	68.95	51.2	**74.99**	**72.20**	**75.71**	71.15	71.86	74.86
MVIP Subset [33]	56.51	55.42	41.2	**63.39**	**61.24**	**48.14**	58.35	58.62	46.45
InVar-100 [17]	72.58	71.56	62.12	**80.68**	78.26	**68.67**	76.90	**80.39**	65.26
DIMO Subset [56]	63.63	61.27	51.80	**66.01**	**65.12**	**56.10**	64.97	62.85	54.02

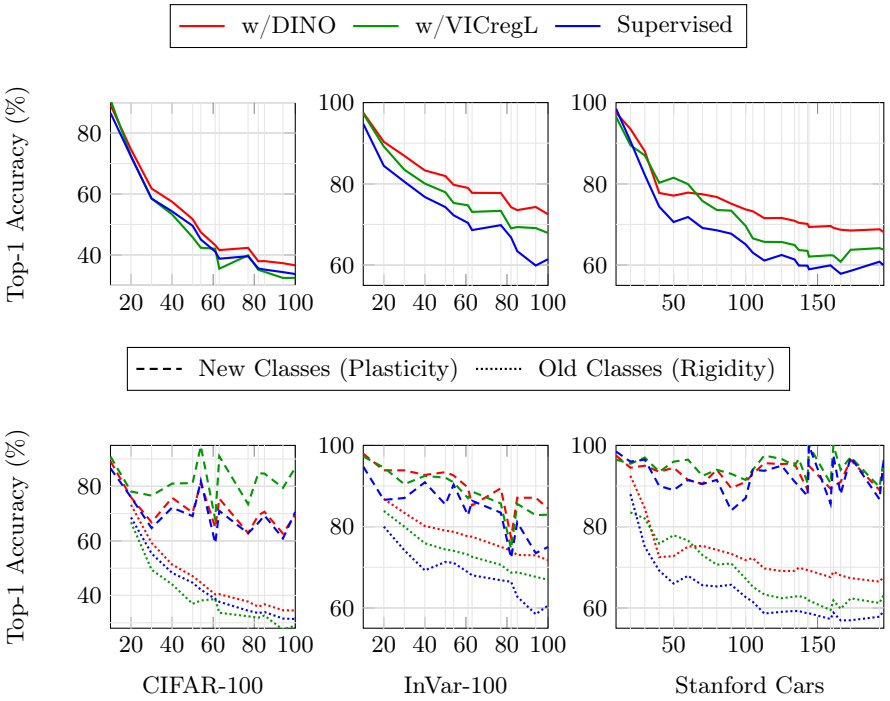

Fig. 2: Accuracy for CIFAR100, InVar-100 and Stanford-Cars Datasets for PODNet implementation with ResNet50 model pretrained on the ImageNet1K dataset. 10 exemplars per class were stored. The X-axis shows the total number of classes during the respective incremental learning tasks (shown by the vertical lines). **Top:** Accuracy on all classes. **Bottom:** Accuracy for old classes and new classes in the dataset. While the performance in new classes is similar for the three approaches, a significant gap is seen in maintaining accuracy in old classes. Additional details are given in Table 2.

5 Discussion

Overview. The *gist* information extracted from the model explanations improves the model performance on old classes, thereby reducing catastrophic forgetting. Semantic contextual information may also be derived from other sensors, e.g. via eye gaze derived from eye-tracking cameras on egocentric devices. We expect, that introducing context-rich multi-modality would further improve the performance of such lifelong learning systems. Similarly, robust SSL pretraining correlates with better performance in IL, especially on old data. In terms of Systems 1 and 2, it must be reiterated that the two always work together during practical implementation, and are aided by each other [31]..

Computational Footprint. It is crucial to consider the computational requirements when developing lifelong learning systems. System 2 is known for being *slow, effortful and demanding* [30]. When processing and storing contextual semantic information, the increase in storage and computational load would be proportional to the dataset size and number of classes. Using model explanations such as CAMs improves the retention of old data, which may offset the increase in compute.

Limitations. We discussed aspects of Continual Learning using System 1 and System 2 concepts. Class-IL was mainly used as the key benchmark task since it presents a challenging scenario w.r.t. catastrophic forgetting. There is room for further exploration of other IL scenarios and setups within this framework. Additionally, CL research involves several other areas in Computer Vision, which we did not address. There is a need for additional research on other forms of gist information that can assist in lifelong learning. We also tested our approach on limited IL methods, datasets and task sequences due to compute constraints.

6 Conclusion

In this paper, we re-contextualised Continual Learning as a System 2 problem and inspected various aspects of the current state-of-the-art with a focus on practical implementation. We extended the analogy of *verbatim* and *gist* traces to augment model training with meaningful semantic information, which improves performance on old data and can also improve learning on new tasks. Model explanations, in the form of CAMs, can provide such meaningful information and encourage the model to focus on the relevant regions in the images. Similarly, self-supervised pretraining improves performance on old data. We test the hypotheses on publicly available datasets under different setups. The System 1 and System 2 framework presents further potential to drive the fields of Continual Learning, AI and Computer Vision forward.

Acknowledgments. This work is funded by the German Federal Ministry of Education and Research (BMBF) and the German Aerospace Center (DLR) under the KIKERP project 0118S23055C in the KI4KMU program.

References

1. Agnoli, F.: Development of judgmental heuristics and logical reasoning: Training counteracts the representativeness heuristic. Cognitive Development **6**(2), 195–217 (1991). https://doi.org/https://doi.org/10.1016/0885-2014(91)90036-D, https://www.sciencedirect.com/science/article/pii/088520149190036D 2
2. Alabdulmohsin, I.M., Neyshabur, B., Zhai, X.: Revisiting neural scaling laws in language and vision. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 22300–22312. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/8c22e5e918198702765ecff4b20d0a90-Paper-Conference.pdf 1
3. Anthony, T., Tian, Z., Barber, D.: Thinking fast and slow with deep learning and tree search. Advances in neural information processing systems **30** (2017) 3
4. Arani, E., Sarfraz, F., Zonooz, B.: Learning fast, learning slow: A general continual learning method based on complementary learning system. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=uxxFrDwrE7Y 3
5. Bardes, A., Ponce, J., LeCun, Y.: Vicregl: Self-supervised learning of local visual features. In: NeurIPS (2022) 4, 6, 17
6. Belouadah, E., Popescu, A., Kanellos, I.: Initial classifier weights replay for memoryless class incremental learning. In: 31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020. BMVA Press (2020), https://www.bmvc2020-conference.com/assets/papers/0743.pdf 2
7. Belouadah, E., Popescu, A., Kanellos, I.: A comprehensive study of class incremental learning algorithms for visual tasks. Neural Networks **135**, 38–54 (2021) 2
8. Benchetrit, Y., Banville, H., King, J.R.: Brain decoding: toward real-time reconstruction of visual perception. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=3y1K6buO8c 2
9. Bengio, Y., LeCun, Y.: Scaling learning algorithms towards AI. In: Large Scale Kernel Machines. MIT Press (2007) 1
10. Bottou, L.: From machine learning to machine reasoning. CoRR **abs/1102.1808** (2011), http://arxiv.org/abs/1102.1808 3
11. Brooks, T., Peebles, B., Homes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C.W.Y., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), https://openai.com/research/video-generation-models-as-world-simulators 3
12. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020). https://doi.org/10.48550/ARXIV.2005.14165, https://arxiv.org/abs/2005.14165 3
13. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline (2020) 3
14. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments (2020) 17

15. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision (ICCV) (2021) 4, 6, 17
16. Chaiken, S., Trope, Y.: Dual-process theories in social psychology. Guilford Press (1999), includes bibliographical references and indexes 2
17. Chavan, V., Koch, P., Schlüter, M., Briese, C.: Towards realistic evaluation of industrial continual learning scenarios with an emphasis on energy consumption and computational footprint. In: Proceedings of the International Conference on Computer Vision (ICCV) (2023) 2, 5, 6, 15
18. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9620–9629. IEEE Computer Society, Los Alamitos, CA, USA (oct 2021). https://doi.org/10.1109/ICCV48922.2021.00950, https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00950 4, 17
19. Chen, Z., Liu, B., Brachman, R., Stone, P., Rossi, F.: Lifelong Machine Learning. Morgan & Claypool Publishers, 2nd edn. (2018) 1
20. Cho, J.H., Hariharan, B.: On the efficacy of knowledge distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4794–4802 (2019) 2
21. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848 1, 4, 5, 15
22. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy 1, 2
23. Douillard, A., Cord, M., Ollion, C., Robert, T., Valle, E.: Podnet: Pooled outputs distillation for small-tasks incremental learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XX. Lecture Notes in Computer Science, vol. 12365, pp. 86–102. Springer (2020). https://doi.org/10.1007/978-3-030-58565-5_6, https://doi.org/10.1007/978-3-030-58565-5_6 2, 4, 15
24. Douillard, A., Ramé, A., Couairon, G., Cord, M.: Dytox: Transformers for continual learning with dynamic token expansion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 2
25. French, R.: Catastrophic interference in connectionist networks: Can it be predicted, can it be prevented? In: Cowan, J., Tesauro, G., Alspector, J. (eds.) Advances in Neural Information Processing Systems. vol. 6. Morgan-Kaufmann (1993), https://proceedings.neurips.cc/paper/1993/file/28267ab848bcf807b2ed53c3a8f8fc8a-Paper.pdf 1, 2, 3
26. Goodfellow, I.J., Mirza, M., Da, X., Courville, A.C., Bengio, Y.: An empirical investigation of catastrophic forgeting in gradient-based neural networks. In: Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings (2014), http://arxiv.org/abs/1312.6211 1
27. Goyal, A., Bengio, Y.: Inductive biases for deep learning of higher-level cognition. Proceedings of the Royal Society A **478**(20210068) (2022). https://doi.org/10.1098/rspa.2021.0068, http://doi.org/10.1098/rspa.2021.0068 2, 3

28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 1
29. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 2
30. Kahneman, D.: A perspective on judgment and choice: mapping bounded rationality. American Psychologist **58**(9), 697–720 (Sep 2003). https://doi.org/10.1037/0003-066X.58.9.697, https://doi.org/10.1037/0003-066X.58.9.697, pMID: 14584987 2, 8, 16
31. Kahneman, D.: Thinking, fast and slow. Farrar, Straus and Giroux, New York (2011), https://www.amazon.de/Thinking-Fast-Slow-Daniel-Kahneman/dp/0374275637/ref=wl_it_dp_o_pdT1_nS_nC?ie=UTF8&colid=151193SNGKJT9&coliid=I3OCESLZCVDFL7 2, 4, 8, 16
32. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017) 2
33. Koch, P., Schlüter, M., Briese, C., Chavan, V.: Mvip: A dataset for industrial part recognition (2023). https://doi.org/http://dx.doi.org/10.24406/fordatis/300, https://fordatis.fraunhofer.de/handle/fordatis/358?mode=full&locale=en, fraunhofer Fordatis 6
34. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13). Sydney, Australia (2013) 6, 15
35. Krizhevsky, A.: Learning multiple layers of features from tiny images pp. 32–33 (2009), https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf 6, 15
36. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 25. Curran Associates, Inc. (2012), https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf 1
37. Kumaran, D., Summerfield, J.J., Hassabis, D., Maguire, E.A.: Tracking the emergence of conceptual knowledge during human decision making. Neuron **63**(6), 889–901 (2009). https://doi.org/10.1016/j.neuron.2009.07.030, https://doi.org/10.1016/j.neuron.2009.07.030 3
38. LeCun, Y.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review **62** (2022) 2, 3
39. Liu, H., Yan, W., Zaharia, M., Abbeel, P.: World model on million-length video and language with ringattention (2024) 3
40. Liu, Y., Schiele, B., Sun, Q.: Adaptive aggregation networks for class-incremental learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2544–2553 (2020) 2
41. Liu, Y., Schiele, B., Sun, Q.: Rmm: Reinforced memory management for class-incremental learning. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 3478–3490. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper/2021/file/1cbcaa5abbb6b70f378a3a03d0c26386-Paper.pdf 2

42. Liu, Y., Su, Y., Liu, A.A., Schiele, B., Sun, Q.: Mnemonics training: Multi-class incremental learning without forgetting. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 12245–12254 (2020) 2
43. Mallya, A., Lazebnik, S.: Packnet: Adding multiple tasks to a single network by iterative pruning. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 7765–7773. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00810, http://openaccess.thecvf.com/content_cvpr_2018/html/Mallya_PackNet_Adding_Multiple_CVPR_2018_paper.html 2
44. Mandler, G.: Recognizing: The judgment of previous occurrence. Psychological Review **87**(3), 252–271 (1980). https://doi.org/10.1037/0033-295X.87.3.252, https://doi.org/10.1037/0033-295X.87.3.252 3
45. Masana, M., Liu, X., Twardowski, B., Menta, M., Bagdanov, A.D., van de Weijer, J.: Class-incremental learning: survey and performance evaluation on image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 1, 2, 5
46. McClelland, J.L., McNaughton, B.L., O'Reilly, R.C.: Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychological review **102**(3), 419 (1995) 3
47. Neys, W.D.: Dual processing in reasoning: Two systems but one reasoner. Psychological Science **17**(5), 428–433 (2006). https://doi.org/10.1111/j.1467-9280.2006.01723.x, https://doi.org/10.1111/j.1467-9280.2006.01723.x, pMID: 16683931 3
48. Niu, Z., Zhong, G., Yu, H.: A review on the attention mechanism of deep learning. Neurocomputing **452**, 48–62 (2021) 3
49. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. CoRR **abs/1802.07569** (2018), http://arxiv.org/abs/1802.07569 3
50. Parthasarathy, N., Eslami, S.M.A., Carreira, J., Henaff, O.J.: Self-supervised video pretraining yields strong image representations (2023), https://openreview.net/forum?id=8onXkaNWLHA 4
51. Petit, G., Popescu, A., Schindler, H., Picard, D., Delezoide, B.: Fetril: Feature translation for exemplar-free class-incremental learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3911–3920 (January 2023) 2
52. Pham, Q., Liu, C., Hoi, S.: Dualnet: Continual learning, fast and slow. Advances in Neural Information Processing Systems **34** (2021) 3
53. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017) 2, 3, 4, 15
54. Reyna, V., Brainerd, C.: Fuzzy-trace theory: An interim synthesis. Learning and Individual Differences **7**(1), 1–75 (1995). https://doi.org/https://doi.org/10.1016/1041-6080(95)90031-4, https://www.sciencedirect.com/science/article/pii/1041608095900314, special Issue: Fuzzy-Trace Theory 3
55. Ritter, S., Barrett, D.G., Santoro, A., Botvinick, M.M.: Cognitive psychology for deep neural networks: A shape bias case study. In: International conference on machine learning. pp. 2940–2949. PMLR (2017) 2
56. Roovere, P.D., Moonen, S., Michiels, N., Wyffels, F.: Dataset of industrial metal objects (2022) 6, 15

57. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models (2022). https://doi.org/10.48550/ARXIV.2210.08402, https://arxiv.org/abs/2210.08402 1
58. Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 2994–3003. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017) 3
59. Silver, T., Chitnis, R., Curtis, A., Tenenbaum, J.B., Lozano-Pérez, T., Kaelbling, L.P.: Planning with learned object importance in large problem instances using graph neural networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 11962–11971 (2021) 3
60. Snowden, R.J., Snowden, R., Thompson, P., Troscianko, T.: Basic vision: an introduction to visual perception. Oxford University Press (2012) 3
61. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf 2
62. van de Ven, G.M., Siegelmann, H.T., Tolias, A.S.: Brain-inspired replay for continual learning with artificial neural networks. Nature Communications **11**(1), 4069 (2020). https://doi.org/10.1038/s41467-020-17866-2, https://doi.org/10.1038/s41467-020-17866-2 3
63. van de Ven, G.M., Tuytelaars, T., Tolias, A.S.: Three types of incremental learning. Nature Machine Intelligence **4**, 1185–1197 (2022) 2
64. Wang, F.Y., Zhou, D.W., Ye, H.J., Zhan, D.C.: Foster: Feature boosting and compression for class-incremental learning. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV. pp. 398–414. Springer (2022) 2
65. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022) 3
66. Weston, J., Sukhbaatar, S.: System 2 attention (is something you might need too) (2023) 2
67. Wistuba, M., Ferianc, M., Balles, L., Archambeau, C., Zappella, G.: Renate: A library for real-world continual learning (2023) 2
68. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 2048–2057. PMLR, Lille, France (07–09 Jul 2015), https://proceedings.mlr.press/v37/xuc15.html 3
69. Yan, S., Xie, J., He, X.: Der: Dynamically expandable representation for class incremental learning. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3013–3022. IEEE Computer Society, Los Alamitos, CA, USA (jun 2021). https://doi.org/10.1109/CVPR46437.2021.00303, https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00303 2, 15

70. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning (2021) 17
71. Zenke, F., Poole, B., Ganguli, S.: Improved multitask learning through synaptic intelligence. CoRR **abs/1703.04200** (2017), http://arxiv.org/abs/1703.04200 3
72. Zhao, B., Xiao, X., Gan, G., Zhang, B., Xia, S.T.: Maintaining discrimination and fairness in class incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 2
73. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016) 4
74. Zhou, D.W., Wang, F.Y., Ye, H.J., Zhan, D.C.: Pycil: A python toolbox for class-incremental learning (2021) 5
75. Zhou, D.W., Wang, Q.W., Ye, H.J., Zhan, D.C.: A model or 603 exemplars: Towards memory-efficient class-incremental learning. In: ICLR (2023) 2
76. Zhu, K., Zhai, W., Cao, Y., Luo, J., Zha, Z.J.: Self-sustaining representation expansion for non-exemplar class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9296–9305 (2022) 2

Appendix

A. Experimental Setup

Table 3 gives the details of the configuration for the tests in Sections 3 (Methods) and 4 (Experiments). We use the same setup for all our Machine Learning (ML) trainings for a fair and unbiased comparison.

Table 3: Details of the parameters and setup used for the experiments presented in the paper.

Parameter	**Value**
Train-Val Split	80/20
Optimizer	SGD
lr start	0.1
lr end	0.0001
weight decay	0.0005
Batch Size	64
Transforms: Train	Resize: (224, 224), RandomHorizontalFlip
Transforms: Val	Resize: (256, 256), CenterCrop
System Memory	48GB
CPU Cores	12
GPU Count	1
GPU type	NVIDIA RTX A6000
Python version	3.8.18

B. Public Datasets

The methods presented in §3 in the paper were verified on diverse publicly available datasets. Here we provide some additional details and justifications for using the datasets.

- **ImageNet100:** We create a subset from the ImageNet1k dataset [21] comprising 100 randomly selected classes. Since this dataset is widely used and studied, it provides proof of generalisability for our methods. We use the dataset for Experiment 1.
- **InVar100:** This dataset contains industrial objects in different contexts and backgrounds [17]. We use this dataset for Experiments 1, 2, 3 and 4; since the different subcategories represent different visual challenges and contexts that we analyse for Continual Learning scenarios. The objects themselves tend to be visually similar, presenting a fine-grained classification problem.
- **CIFAR-100:** This is a widely studied and benchmarked dataset. It enables quick testing due to its small size [35]. We use the dataset in Experiment 4. Unlike the previously established protocol for Continual Learning with the dataset [53], we do not use the ResNet34 model. ResNet50 serves as the backbone, due to the wider availability of models pretrained using supervised and self-supervised learning methods. We do not employ the dataset for Experiment 1, since given the small size of the images, it is not practical for extracting *Gist* information in the form of class activations.
- **Stanford Cars:** This dataset represents a fine-grained classification challenge and serves as a good case study for domain-specific applications [34]. We use the dataset for Experiment 4.
- **DIMO:** The dataset contains real and synthetic images of industrial metal objects [56]. It also serves as a domain-specific and fine-grained categorisation problem. We use the dataset in Experiment 4.

C. Class Incremental Learning Methods

We implement and test our methods on the iCaRL [53] and PODNet [23] Class-IL approaches in Experiments 1 and 2. Similar to the datasets, we would argue that the inferences we draw are general and can be extended to other Continual Learning scenarios and methods. We could not experiment and test more methods due to resource constraints. For Experiment 4, we also use the DER [69] implementation, since the hypothesis (also discussed in §6) is more general and widely applicable. The number of exemplars for experience replay is kept constant per class, unlike other common approaches (e.g. 2000 total samples). We would argue that allowing a higher memory for the initial tasks (i.e. $2000/Num_{classes}$) skews the performance results when comparing the effect of methods and augmentations that we propose.

D. Continual Learning as a System 2 Problem

In the paper, we use the Dual Process Theory [30, 31] and other frameworks in cognitive studies as a starting point for our work. It is important to reiterate that the comparisons are meant as a loose analogy, and we do not claim that human cognition and AI models work based on the same principles. Furthermore, we implemented some concepts from human thinking and reasoning for our approach, which deals with Computer Vision. We believe there is substantial scope for research and development in Continual Learning and Computer Vision based on such analogies. However, this may not be true in each scenario and cannot be accepted as universally applicable. As stated in the paper, a System 2 problem inherently requires System 1 processes and does not function independently in practice. Moreover, through processes such as model pretraining and learning from multiple contexts, we show that there is a synergy between System 1 and System 2 behaviour in the framework of ML and Continual Learning. As such, it is rather difficult to establish a clear separation between System 1 and System 2. We are happy to revise the comparisons and analogies further for more clarity based on the feedback.

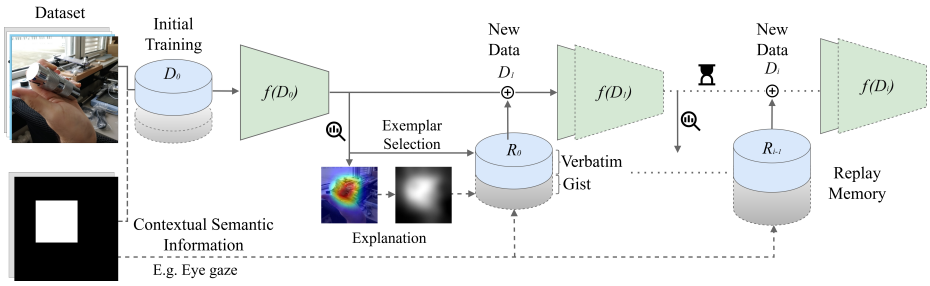

Fig. 3: Proposed approach to incorporate domain-specific semantic knowledge to encourage the model to remember the previous tasks while effectively learning the new tasks. The gist knowledge can be explicitly encoded to improve model plasticity and rigidity. Alternatively, information extracted from the model training (such as CAMs) can be used to improve performance on older tasks.

Encoding Contextual Information

We discussed the use of *Gist* information as context-based semantic information for improving performance during Continual Learning. The results for the ImageNet100 and InVar-100 datasets in Experiment 1 in §4. We conjecture that the *gist* information can be provided in other forms and modalities, which remains to be verified. However, we do not intend for context-based semantic information to only represent importance masks or class activations.

Pretraining Schemes and Transfer Learning

We explore the use of different pretraining schemes for Continual Learning scenarios in §3.3 in Methods and Experiments 3 and 4 in §4 in the paper. The conjecture is that self-supervised pretraining is more robust and encourages the model to learn better and more generalised features from, which transfer better for challenging scenarios such as Class Incremental Learning. Improved performance is seen on old tasks, which indicates lower catastrophic forgetting when using SSL-pretrained models. We also verify this for a simple use case in Experiment 3 where the ResNet50 model is fine-tuned on the different subcategories of the InVar-100 dataset (Fig. 7 and 8 in the paper). We visualise the features for the last fully connected layer for the fine-tuned models from each task and observe that the key representations of the target object are better visualised in the SSL-pretrained model. This only serves as a rough visual conformation for the presented inferences. We do not provide such visualisations for additional objects and methods due to lack of space and to avoid redundancy.

We also tested other self-supervised pretraining approaches, viz., MoCo V3 [18], Barlow Twins [70], SwAV [14]. However, we did not notice statistically improved performance over their supervised counterpart. We believe the features learnt by VICRegL [5] and DINO [15] are more general and superior. There is additional scope for further research on the intersection of Continual Learning and pretraining schemes.

Experiment 3: Fine-tuning on different Subsets of the InVar-100 dataset. Starting with ResNet50 models pretrained on ImageNet1K using Supervised Learning, VICRegL, and DINO, we fine-tune on different subsets of the InVar-100 dataset. Task 0 introduces data with *white background*, Task 1 introduces data with *stationary setup*, Task 2 introduces data with *handheld* objects, and Task 4 introduces images with *cluttered background*. We introduce all classes during each task. No replay memory was stored from previous tasks. As new tasks are introduced, the SSL pretrained models outperform the supervised model (Fig. 4). We inspect the feature activations from the last layer of the models for different classes and observe that the SSL models learn better representations that are retained as the visual complexity of the tasks increases. Fig. 5 shows an example.

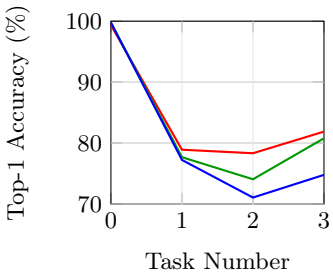

Fig. 4: Fine-tuning results on various subcategories of the InVar-100 dataset on pretrained Resnet50. No replay memory from previous tasks was allowed. SSL pretrained models perform better.

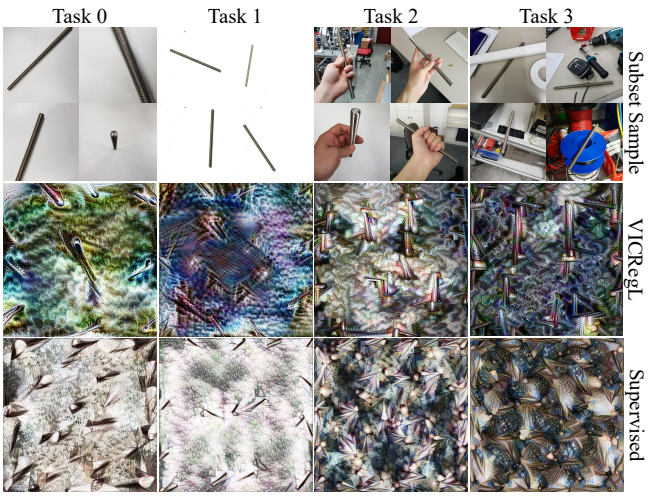

Fig. 5: Feature visualisations for a class index from InVar-100. SSL pretrained model shows better retention of the core features of the object.
Top: Samples from the subset for a class from the dataset.
Middle: Feature visualisation from ResNet50 pretrained using VICRegL.
Bottom: Feature visualisations from ResNet50 pretrained via supervised learning.

D. Computational Overhead Analysis

In this section, we detail the computational cost and model parameters associated with incorporating CAMs into our continual learning framework. We utilise the ResNet18 architecture as our base model and conduct experiments on the InVar-100 dataset. After each task, we store 20 exemplars per class, resulting in 2000 exemplars in total.

Storage Requirements

Exemplar Images: Each exemplar image is of size 224×224 pixels with 3 colour channels (RGB), leading to a storage requirement of approximately 150 KB per image (assuming 8 bits per pixel per channel). For 2,000 exemplars, the total storage for images amounts to approximately 300 MB.

Class Activation Maps (CAMs): CAMs are stored as single-channel saliency maps corresponding to the important regions in each image. If stored at the same resolution as the input images (224×224), each CAM requires approximately 50 KB of storage. Thus, storing CAMs for all exemplars adds $\sim$100 MB, increasing the storage overhead by approximately 33

To mitigate this overhead, we can store CAMs at a lower resolution. For instance, downsampling CAMs to 56×56 pixels reduces the storage requirement to approximately 3 KB per CAM. This results in a total of 6 MB for all CAMs, which is a negligible increase of 2% in the overall storage requirement.

Computational Cost

CAM Generation: Generating CAMs involves an additional forward and backward pass to compute gradients for the target class scores. However, since CAMs are generated only once per exemplar after the task is introduced, the

computational overhead is amortised over the entire training process. For 2000 exemplars, the one-time cost is manageable and does not significantly impact the overall training time.

Training Overhead: Incorporating CAMs into the loss function introduces additional computations during training. Specifically, the calculation of the *gist* loss $\mathcal{L}_{\text{gist}}$ requires element-wise multiplication between the loss and the CAM-derived masks. This operation is computationally less intensive compared to the standard forward and backward passes.

Inference Efficiency

During inference, CAMs are not utilized, and thus there is no additional computational overhead or latency introduced in the deployment phase. The model size remains the same as the base ResNet18 architecture, ensuring that inference speed and resource requirements are unaffected.