ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240466

Zero-Waste Machine Learning

Tomasz Trzcinski[a,b,*], Bartłomiej Twardowski[a,d,e], Bartosz Zieliński[a,c], Kamil Adamczewski[a] and Bartosz Wójcik[a,c]

[a]IDEAS NCBR
[b]Warsaw University of Technology
[c]Jagiellonian University
[d]Computer Vision Center, Barcelona, Spain
[e]Universitat Autonoma de Barcelona, Barcelona, Spain
ORCID (Tomasz Trzcinski): https://orcid.org/0000-0002-1486-8906, ORCID (Bartłomiej Twardowski): https://orcid.org/0000-0003-2117-8679, ORCID (Bartosz Zieliński): https://orcid.org/0000-0002-3063-3621, ORCID (Kamil Adamczewski): https://orcid.org/0000-0002-2917-4392, ORCID (Bartosz Wójcik): https://orcid.org/0000-0002-1100-4176

Abstract. Today, both science and industry rely heavily on machine learning models, predominantly artificial neural networks, that become increasingly complex and demand more computing resources to be trained. In this paper, we will look holistically at the efficiency of machine learning models and draw the inspirations to address their main challenges from the green sustainable economy principles. Instead of constraining some computations or memory used by the models, we will focus on reusing what is available to them: computations done in the previous processing steps, partial information accessible at run-time, or knowledge gained by the model during previous training sessions in continually learned models. This new research path of zero-waste machine learning can lead to several research questions related to efficiency of contemporary neural networks - how machine learning models can learn better with less data? How they select relevant data samples out of many? Finally, how can they build on top of already trained models to reduce the need for more training samples? Here, we explore all the above questions and attempt to answer them.

1 Introduction

Today, both science and industry heavily depend on machine learning models, especially artificial neural networks, which are becoming increasingly complex and require substantial computational resources. This trend is evident in a wide range of applications, from medical image processing to robotics. However, the most significant computational demands of machine learning models are seen in large high-energy physics experiments, where an enormous amount of data is generated and analyzed. For example, the ALICE experiment at CERN's Large Hadron Collider (LHC), the world's largest and most powerful particle accelerator, gathers several petabytes of data every hour, which is processed by numerous machine learning models.

The computations run by machine learning models to process this increasing amount of data come at an enormous price of long processing time, high energy consumption and large carbon footprint generated by the computational infrastructure [32]. Existing approaches to reduce this burden are either focused on constraining the optimization with a limited budget of computational resources [17] or they attempt to compress models [15].

In this paper, we look holistically at the **efficiency of machine learning models** and draw inspiration to address their main challenges from the green sustainable economy principles. Instead of limiting training of machine learning models, we ask a different question: how can we make the best out of the resources and information and computations that we already have access to? Instead of constraining the number of computations or memory used by the models, we focus on reusing what is available to them: computations done in the previous processing steps, partial information accessible at run-time or knowledge gained by the model during previous training sessions in continually learned models. We look at the research problem of efficient machine learning from the computation recycling perspective and propose methods to overcome its main challenges. Driven by this assumption, we have initiated a new research path of **zero-waste machine learning** focused on saving computations of machine learning models and reducing their impact on resource usage. This research is currently done in our research group at IDEAS NCBR, a Polish publicly-funded AI Center, and this paper aims to summarize recent works of this group.

To explore the landscape of zero-waste machine learning, we ask the following research questions:

- *How to optimize the resource usage of machine learning models by enforcing zero-waste policy and computations recycling?*
- *How to train neural network-based representations efficiently by conditioning their computations?*
- *How to accumulate knowledge efficiently within and beyond continually learned models?*
- *How can we evaluate machine learning models' efficiency from the resource recycling perspective?*

We attempt to answer the above questions in the remainder of this paper. We define *zero-waste machine learning* as a research path towards more efficient machine learning models, focused specifically on two aspects of re-using available resources: conditioning computations and acquiring knowledge in continually trained models. In the

* Corresponding Author. Email: tomasz.trzcinski@ideas-ncbr.pl

last part of this work, we look at the real-life use cases of zero-waste machine learning in autonomous robots and explore how the wide range of constraints they face can be used to inspire new approaches towards efficiency. We conclude this article with some general observatations and paths towards more resource-aware development of new machine learning models.

2 Conditional computations

Conditional computation is a technique in deep learning where only a subset of a model's components are activated during a forward pass, rather than using the entire network. This approach reduces computational costs and improves efficiency without significantly compromising performance. Our work in conditional computation focuses on two main paradigms: Mixture of Experts (MoE) and Early-Exit strategies. Below we outline our recent works in this area.

In the Mixture of Experts paradigm, conditional computation is achieved by routing different inputs to different subsets of experts, instead of engaging all experts for every input. This selective activation allows the model to scale effectively, as only the most relevant experts are used, thereby conserving computational resources.

Early exits, another form of conditional computation, enable the network to terminate the forward pass early for certain inputs if sufficient confidence is achieved in the predictions. This mechanism is particularly useful when some inputs are easier to process than others, allowing the model to save computation time by exiting early for simpler inputs.

Zero Time Waste: Recycling Predictions in Early Exit Neural Networks [37]. In this work we initiated our research on zero-waste machine learning methods. We proposed Zero Time Waste (ZTW), an efficient early-exit approach especially suitable for application to pre-trained models. While early-exiting is an established method to accelerate classification models [31], common early-exit methods discard information from previous classification heads if the classifier confidence is low. By reusing that information instead, ZTW is able to achieve state-of-the-art performance-vs-compute results.

Exploiting Activation Sparsity with Dense to Dynamic-k Mixture-of-Experts Conversion. While Mixture of Experts models were previously used mostly for scaling up the parameter count of the models [6], a recent work of Zhang et al. [40] has shown that static models can be converted to MoEs to improve execution time. Our work makes further progress to improve such converted MoE models by: 1) showing that the efficiency of the conversion can be significantly enhanced by enforcement of activation sparsity in the base model; 2) proposing Expert Contribution Routing, a novel objective for the training of the gating networks, which are now tasked to predict the output norm of each expert for the given input; 3) introducing dynamic-k gating, which allows the model to appropriately distribute its computational budget between easy and hard inputs; 4) extending the proposed conversion scheme to any linear layers such as multi-head attention projections. The conversion and router training scheme is presented in fig:d2dmoe. The proposed method, D2DMoE (Dense to Dynamic-k Mixture of Experts), outperforms existing approaches on common NLP and vision tasks, allowing us to save up to 60% of inference cost without significantly affecting model performance.

Adaptive Computation Modules: Granular Conditional Computation For Efficient Inference. The ability of a model to adjust the computational cost on a granular, per-token basis was also the inspiration for Adaptive Computation Module (ACM), which we recently proposed. An ACM consists of a sequence of *learners* that progressively refine the output of their preceding counterparts. An additional gating mechanism determines the optimal number of learners to execute for each token, which results in a spatially varying computational load for images, as shown in 1. Furthermore, in contrast to MoE layers, the learner-based approach limits the number of possible combinations of learners that can be executed for a token. This allows for an efficient GPU implementation that has an execution time that scales linearly with the number of executed floting-point operations (FLOPs). Our evaluation of transformer models in computer vision and speech recognition demonstrates that substituting layers with ACMs significantly reduces inference costs without degrading the downstream accuracy for a wide interval of user-defined budgets.

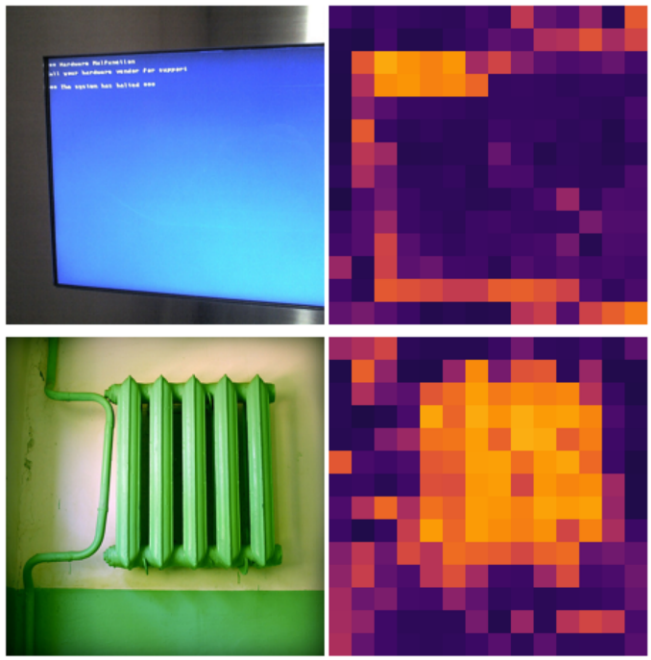

Figure 1: ACM-based model adapts its computational load to spend its resources on the semantically important regions of the input.

Scaling Laws for Fine-Grained Mixture of Experts. [14] This paper discusses the efficiency of Mixture of Experts models in reducing the computational cost of Large Language Models (LLMs). We introduce a new hyperparameter, granularity, to precisely control the size of experts. We further establish scaling laws to estimate the performance for fine-grained MoE models considering training tokens, model size, and granularity. The laws allow us to find optimal parameters for a given computational budget. The findings show that MoE models consistently outperform dense Transformers, especially as model size and training budget increase. The paper argues against the common practice of setting the size of MoE experts to match the feed-forward layer size. The results indicate that higher granularity leads to better performance and efficiency at any computational budget with the efficiency gap compared to vanilla transformers and non-granular MoE models increasing for higher budgets. Thus, we challenge previous claims that dense models might surpass MoE at large scales. The work provides practical guidance for selecting optimal training hyperparameters, emphasizing the importance of granularity and training duration for compute-efficient large language models. The paper was published at ICLM 2024.

SADMoE: Exploiting Activation Sparsity with Dynamic-k

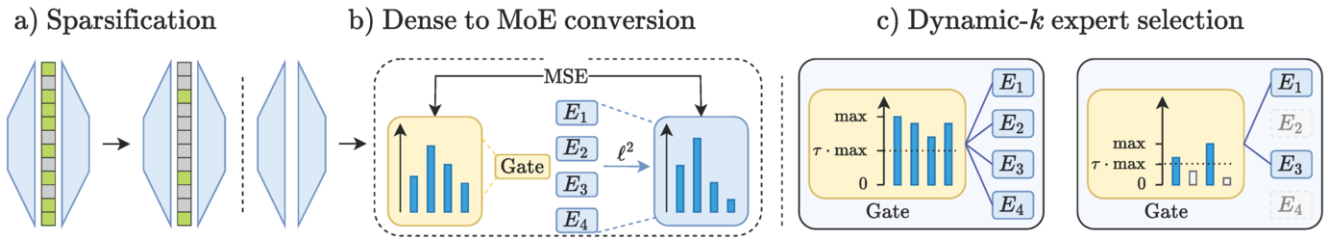

Figure 2: D2DMoE finetunes a static model to enforce activation sparsity, then splits the weights into experts, and finally trains the router to predict the norm of each expert. In inference, a threshold hyperparameter is used to skip the computation of experts with a relatively low predicted output norm.

Gating [34]. In this paper we introduce the Sparsified Activation Dynamic-k Mixture of Experts (SADMoE) approach to enhance the efficiency of Transformer models by leveraging activation sparsity. We propose transforming parts of the network into Mixture of Experts layers and enforcing higher activation sparsity levels, leading to significant computational savings. Moreover, we introduce a dynamic-k expert selection rule, 4which adjusts the number of executed experts based on the input's complexity, optimizing computational resources. This method is further extended to multi-head attention projections, yielding additional efficiency gains. Experimental results on various NLP and vision tasks demonstrate that SADMoE can reduce inference costs by up to 60% without significantly impacting model performance. This innovative approach addresses the high computational demands of Transformer models, making them more accessible for deployment in resource-constrained environments.

Joint or Disjoint: Mixing Training Regimes for Early-Exit Models. Our recent yet unpublished work on mixing training regimes explores the efficiency mechanism of early exits in deep neural networks, which allows terminating the network's forward pass before processing all its layers. The study proposes a novel mixed training approach, where the backbone network is initially trained on its own, followed by joint training with internal classifiers. The research categorizes training strategies into disjoint, joint, and mixed regimes and performs both theoretical (based on mutual information, loss landscape, and numerical rank) and empirical analyses to gauge their performance and efficiency across various architectures and datasets. The findings highlight that the mixed regime generally offers the best balance between computational cost and accuracy, suggesting its suitability for a wide range of input complexities and resource constraints. The paper's contributions provide valuable insights into optimizing the training of early-exit models for more efficient deep learning systems.

3 Acquiring knowledge in continual learning

The second pillar of zero-waste machine learning, Continual Learning (CL), has emerged as a critical paradigm in the field of deep learning, aiming to enable models to learn from a continuous stream of data without forgetting previously acquired knowledge. However, a significant challenge that remains is the phenomenon of catastrophic forgetting [22], where a model, when fine-tuned on new data, tends to overwrite its previously acquired knowledge. The issue occurs universally across different neural network architectures and can severely limit the effectiveness of models in real-world, where incoming data distribution changes and i.i.d assumption is often too strong.

Continual learning methods enable models to retain and integrate new information over time without forgetting previously learned tasks. These methods can be grouped into three main categories: regularization-based methods, architecture-based models, and replay-based methods [26, 21, 4, 36]. Depending on the application scenario, one continual learning method may outperform others. However, while many methods excel at mitigating catastrophic forgetting in specific scenarios, their efficiency and sustainability aspects are often overlooked.

In the zero-waste machine learning research line, we address this gap by focusing on the efficiency of CL methods. We emphasize the necessity of CL in fostering energy-efficient deep learning models and pose the following questions for CL methods: *How can we efficiently accumulate knowledge in a sustainable and energy-efficient manner, not just preventing forgetting from old tasks?* Furthermore, we explore beyond traditional CL: *Can we continue to efficiently adapt the model after deployment, even during test time, without any labels?*

Below, we present a set of our works that address the above questions. Starting with classical continual learning methods, such as class-incremental learning, we propose a new approach to efficiently incorporate new classes into the model. We then revisit self-supervised and supervised model training to enhance knowledge transferability and accumulation. We discuss how to improve the efficiency of generative replay methods using simple techniques, and how to estimate data drift with adversarial attacks in continual learning. Finally, we propose one of the most promising zero-waste machine learning techniques – using model merging to address knowledge accumulation in continual learning settings.

Divide and not forget: Ensemble of selectively trained experts in Continual Learning [30]. A popular strategy for continual learning involves increasing a number of parameters during incremental updates [38, 2]. Yet, this can result in indefinite parameter growth, and therefore, we have experimented with a variant where the number of parameters is increased but kept fixed during the incremental updates. Our method SEED: *Selection of Experts for Ensemble Diversification* fine-tunes only one, the most optimal expert for a considered task. However, during inference, all experts contribute by performing Bayes classification as presented in Figure 3. During the training, we do not increase computation requirements, as we still only train the selected expert. Additionally, the number of trainable parameters can be limited even more by sharing the first few layers of the experts' network. That gives the method a lot of flexibility and

versatile application possibilities. SEED presents outstanding results in both scenarios: where the first task is big (scenario promoting stability) and learning from scratch on multiple small tasks (scenario promoting plasticity).

FeCAM: Exploiting the heterogeneity of class distributions in exemplar-free continual learning [9]. Recent advancement in foundation models changed the approach of training big neural networks, where we start already from a good pre-trained model. This approach gained much attention in class-incremental learning where the backbone is not trained at all. It is used as an already good feature extractor, and the focus is how to continually train the classifier. In this work, we explore prototypical networks for CIL, which generate new class prototypes using the frozen feature extractor and classify the features based on the Euclidean distance to the prototypes. In an analysis of the feature distributions of classes, we show that classification based on Euclidean metrics is successful for jointly trained features. However, when learning from non-stationary data, we observe that the Euclidean metric is suboptimal and that feature distributions are heterogeneous. To address this challenge, we this work revisit the anisotropic Mahalanobis distance and propose a new method FeCAM for CIL. In addition, this work show that modeling the feature covariance relations is better than previous attempts at sampling features from normal distributions and training a linear classifier. Unlike existing methods, our approach generalizes to both many- and few-shot CIL settings, as well as to domain-incremental settings. Interestingly, without updating the backbone network, our method obtains state-of-the-art results on several standard continual learning benchmarks.

Adapt Your Teacher: Improving Knowledge Distillation for Exemplar-free Continual Learning [35]. Knowledge distillation is a highly effective regularization technique for continual learning. Several methods prevent forgetting by regularizing the outputs of the new model with the old model, either without using exemplars [16, 3, 1, 13] or with a small memory buffer [29, 3, 11, 5, 1]. In this work, we investigate improving knowledge distillation in continual learning by also adapting the teacher model during training. This adaptation, achieved through batch norm updates, enhances multiple methods, leading to more stable knowledge accumulation in class-incremental learning. This results in better CKA alignment of features and overall accuracy during the learning session.

Looking through the past: better knowledge retention for generative replay in continual learning [12]. Generative replay addresses catastrophic forgetting in continual learning by recalling old data through a generative model, but it is often complex and inefficient. In this work, we explore applying generative replay in the feature space rather than the input image space. However, this approach faces performance degradation due to generated features differing from the original ones in the latent space. To address this, we propose three modifications: incorporating latent space distillation between current and previous models to reduce feature drift, using latent matching to improve feature alignment, and cycling generations through previous models to enhance data reconstruction. Together, these modifications outperform existing generative replay methods across various scenarios with more efficient way and allows application to bigger scale scenarios (not only small image datasets like in many generative-replay based methods).

Revisiting Supervision for Continual Representation Learning [19]. In this work, we focus how representations can build continually and accumulate the knowledge efficiently. Most CL research focusing on supervised continual learning for image classification [21, 36]. Recently, unsupervised continual learning has gained attention for its ability to build robust representations without labeled data, which can leverage vast amounts of emerging unlabeled data [7, 8, 18]. Studies have shown that unsupervised approaches, particularly self-supervised learning (SSL), develop more resilient representations compared to supervised learning, despite the counter-intuitive notion that having more information (labels) should be beneficial. This discrepancy is partly attributed to the transferability gap, which recent findings suggest is improved by using a multi-layer perceptron (MLP) projector in SSL. Inspired by these advancements, this work revisits the role of supervision in continual representation learning, arguing that human annotations should enhance rather than hinder representation quality. We aim to leverage these insights to improve task transferability in continual learning.

Resurrecting Old Classes with New Data for Exemplar-Free Continual Learning [10]. Exemplar-free methods in continual learning are among the most intriguing and challenging. Without access to old data, feature drift can cause saved prototypes of old classes to become inaccurate. Methods like SDC [39] estimate this drift using only new data and counteract it accordingly. Our work improves upon this with method ADC, which utilizes adversarially generated samples. Specifically, ADC perturbs current samples so their embeddings align with old class prototypes in the previous model's embedding space. We then estimate the drift in the embedding space from the old to the new model using these perturbed images and adjust the prototypes accordingly. This approach leverages the transferability of adversarial samples between feature spaces in continual learning. Generating these images is simple and computationally efficient. Our experiments show that ADC better tracks prototype movement in the embedding space and outperforms existing methods on several standard continual learning benchmarks.

MagMax: Leveraging Model Merging for Seamless Continual Learning [20]. Model merging addresses the challenge of accumulating knowledge from pre-trained models without additional training. Typically, specific conditions are required for model merging to be effective. In this work, we explore how this approach can solve the continual learning problem. Unlike traditional continual learning methods that focus on minimizing forgetting during task training, MagMax combines sequential fine-tuning with maximum magnitude weight selection for efficient knowledge integration across tasks. A key contribution of this work is an extensive examination of model merging techniques, showing that simple methods like weight averaging and random weight selection perform surprisingly well in various continual learning scenarios. Importantly, MagMax introduces a novel model-merging strategy that facilitates the continual learning of large pre-trained models for successive tasks. Model merging in CL opens new possibilities for very efficient model preparation for the following task without training at all, and reusing already pre-trained models. From that perspective, this line of work is zero-waste and energy-efficient.

AR-TTA: A Simple Method for Real-World Continual Test-Time Adaptation [33]. This work is going beyond classical continual learning framework and tackle the problem of continual model adaptation during the test-time, without knowing any labels and assumption what the input data shift will be. In this paper, we have prepared a new challenging dataset for test-time adaptation methods, by utilizing recent autonomous driving dataset and benchmarked existing SOTA methods against it. As visualized in Fig. 4, existing methods excel at synthetic benchmarks (CIFAR-C), but lack robustness to novel scenarios (CLAD-C). The proposed AR-TTA method enhance well-established self-training framework by incorporating a small memory buffer to increase model stability and at the same

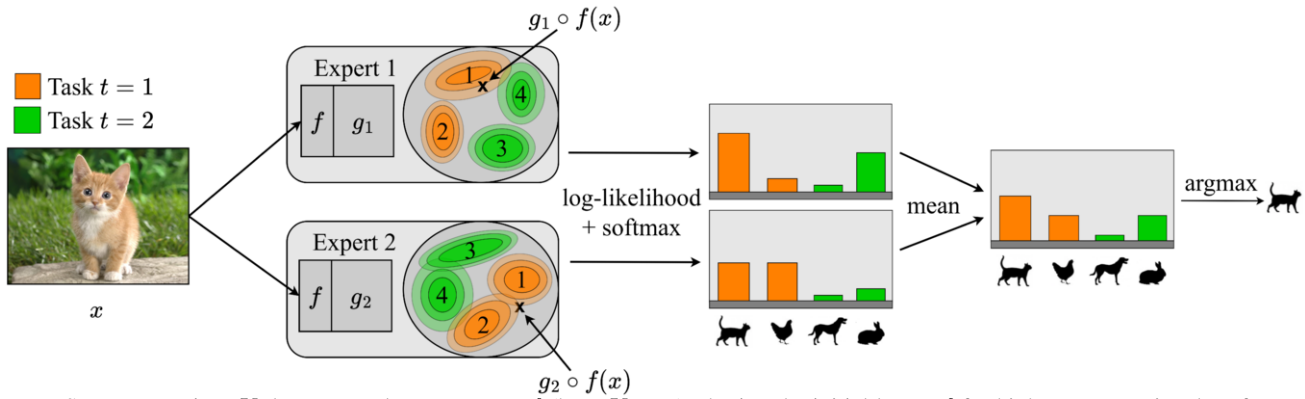

Figure 3: SEED comprises K deep network experts $g_k \circ f$ (here $K = 2$), sharing the initial layers f for higher computational performance. f are frozen after the first task. Each expert contains one Gaussian distribution per class $c \in C$ in his unique latent space. In this example, we consider four classes, classes 1 and 2 from task 1 and classes 3 and 4 from task 2. During inference, we generate latent representations of input x for each expert and calculate its log-likelihoods for distributions of all classes (for each expert separately). Then, we softmax those log-likelihoods and compute their average over all experts. The class with the highest average softmax is considered as the prediction.

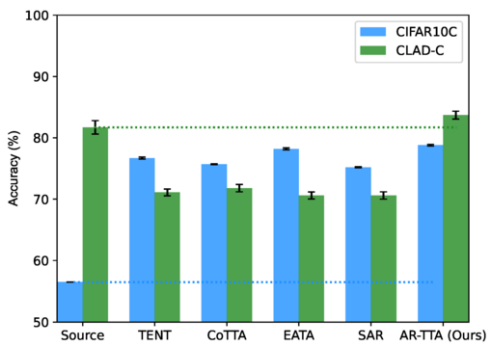

Figure 4: In [33] continual test-time adaptation methods were evaluated on synthetic (CIFAR-10C) and realistic (CLAD-C) domain shifts. This simple reality check presents that multiple state-of-the-art methods while being good on corrupted CIFAR-10 dataset cope to overpass source model (no adaptation at all) while being tested on more realistic CLAD-C dataset. The proposed AR-TTA method is the only one that consistently allows to improve over the naive strategy of using the source model.

time perform dynamic adaptation based on the intensity of domain shift. AR-TTA outperforms existing approaches on both synthetic and more real-world benchmarks, and shows robustness across a variety of TTA scenarios.

4 Real-life use cases

While the most obvious application of zero-waste machine learning methods is large language models, another promising field of application is Autonomous Mobile Robots (AMRs). AMRs should independently understand and explore their environment, which can currently be obtained only with state-of-the-art computer vision methods. Those methods, however, require significant energy consumption due to the limited battery capacity of AMRs. That is why it is essential to introduce methods dedicated to autonomous machines rather than adapting existing approaches, as robotic teams currently do.

In our research, we introduce methods dedicated to AMRs in two research tasks. The first of them is Active Visual Explorations (AVE), which involves dynamically selecting observations (glimpses) to understand and navigate environments effectively. While current AVE techniques show impressive results, they are limited to fixed-scale glimpses from predetermined grids. In contrast, modern mobile platforms with optical zoom capabilities allow capturing glimpses at various positions and scales. We aim to fill this gap in our research.

Active Visual Exploration Based on Attention-Map Entropy [24]. This paper introduces a transformer-based approach to Active Visual Exploration (AVE) that uses internal model uncertainty to guide the selection of glimpses for better scene understanding and navigation. Unlike previous methods that require separate sampling decision modules, the proposed method leverages entropy from the transformer's attention maps to choose the next observation efficiently, simplifying the overall system. The model, built upon a masked autoencoder architecture, integrates glimpse selection directly into its framework, allowing it to gather partial observations sequentially without additional training streams. Experimental evaluations demonstrate that this approach outperforms existing state-of-the-art methods in tasks like image reconstruction, classification, and segmentation. The method is validated across various patch configurations, including retina-like glimpses, and consistently shows superior performance and applicability to realistic AVE scenarios.

Beyond Grids: Exploring Elastic Input Sampling for Vision Transformers [23]. This paper addresses the limitations of Vision Transformers (ViT) in handling input tokens from irregular grids, particularly in Active Visual Exploration (AVE) scenarios, where agents must navigate environments using observations of varying scales and positions. The main aim is to investigate and enhance the resilience of ViT architectures to non-standard input sampling strategies, which are crucial for real-life applications like drone navigation. For this purpose, we introduce an evaluation protocol to measure three types of transformer elasticity: scale, missing data, and positional elasticity. We analyze how standard ViT architectures respond to changes in input sampling, and then we propose architectural and training modifications, collectively termed ElasticViT, to improve this resilience. The experimental results demonstrate that the proposed ElasticViT modifications significantly boost performance in scenarios with limited and varied input data, outperforming traditional ViT models. By formalizing input elasticity and validating

it through rigorous testing, this work highlights the importance of flexible input sampling strategies, suggesting that accommodating alternative patch resolutions can enhance the practical applicability of vision transformers in real-world tasks.

AdaGlimpse: Active Visual Exploration with Arbitrary Glimpse Position and Scale [25]. This paper introduces AdaGlimpse, a novel method for Active Visual Exploration (AVE) that overcomes the limitations of current approaches by allowing agents to select glimpses of arbitrary scale and position, enabling more effective and efficient visual exploration for embodied agents. The proposed method leverages an input-elastic vision transformer that, in each exploration step, predicts the optimal position and scale for the next glimpse in a continuous space. This is achieved using a reinforcement learning framework, specifically the Soft Actor-Critic algorithm, which excels in exploration tasks. Unlike traditional methods that rely on fixed grids, AdaGlimpse dynamically selects visual samples, mimicking the natural way humans explore their surroundings by adjusting their focus based on prior observations. Experimental results demonstrate that we significantly outperform state-of-the-art methods in tasks such as image reconstruction, classification, and segmentation. The ability to select glimpses of varying scales and positions allows for faster and more effective environmental awareness, making it highly applicable to real-world scenarios and fully exploiting the capabilities of modern hardware in embodied platforms.

The second research task we consider in our AMR-related research is Self-Supervised Learning (SSL), which is a powerful technique for learning robust representations from unlabeled data by remaining invariant to applied data augmentations. We want to introduce robotic-centric SSL methods.

A deep cut into Split Federated Self-supervised Learning [28]. This paper introduces Momentum-Aligned Contrastive Split Federated Learning (MonAcoSFL), a novel approach to improving communication efficiency and accuracy in federated learning setups, particularly focusing on models split at deeper layers, more optimal from the point of privacy. The main aim is to address the limitations of current federated learning frameworks, especially the Split Federated Learning (SFL) and Momentum Contrast (MoCo)-based MocoSFL, which suffer from increased communication overhead and reduced client data protection at deeper splits. MonAcoSFL these issues by synchronizing both online and momentum models during training, unlike MocoSFL, which only synchronizes the online models. This alignment prevents the divergence between the models, reducing confusion and enhancing training consistency. By optimizing the synchronization process, we ensure that deeper model splits, which are more communication-efficient, do not compromise accuracy or increase computational overhead significantly. Experimental results demonstrate that we significantly outperform baseline methods, maintaining high performance and reducing communication overhead. This way, we offer a robust solution for real-world federated learning applications, making it more suitable for scenarios with mobile and resource-constrained devices.

Augmentation-aware Self-Supervised Learning with Conditioned Projector [27]. This paper introduces Conditional Augmentation-aware Self-supervised Learning (CASSLE). This novel method addresses the limitations of traditional contrastive SSL methods that suffer from augmentation-induced invariance in representation spaces, which can hinder performance on downstream tasks requiring sensitivity to specific data changes. CASSLE improves upon existing SSL frameworks by introducing a conditioned projector network that incorporates augmentation information during training. This approach ensures that the feature extractor network preserves more meaningful details about augmented images, enhancing the sensitivity of learned representations without major modifications to network architectures. Experimental validation demonstrates that CASSLE outperforms previous augmentation-aware methods thanks to its ability to seamlessly integrate into existing SSL approaches, making it a practical solution for improving model transferability and robustness. By maintaining sensitivity to changes induced by augmentations, CASSLE enables robots to adapt more effectively to varying environmental conditions and sensor inputs, enhancing their overall performance in dynamic real-world scenarios.

5 Conclusions and future outlook

In this paper, we have introduced and explored the concept of zero-waste machine learning, a novel approach aimed at enhancing the efficiency and sustainability of machine learning models. Drawing inspiration from the principles of a green sustainable economy, we focused on reusing existing computational resources, partial information accessible at runtime, and knowledge from previous training sessions rather than merely constraining computational budgets or compressing models.

We have demonstrated that conditional computation techniques such as Mixture of Experts and Early-Exit strategies can significantly reduce computational costs without compromising performance. Our proposed methods, like Zero Time Waste and D2DMoE, have shown promising results in optimizing the execution time and resource usage of neural networks. Addressing the challenge of catastrophic forgetting, our research in continual learning has emphasized the importance of efficient knowledge accumulation. We presented various methods such as SEED and ADC that enhance the stability and plasticity of models, ensuring sustainable performance in dynamic real-world environments.

Our exploration of zero-waste machine learning in autonomous mobile robots has illustrated the practical benefits of our approach. By introducing methods like Active Visual Exploration and self-supervised learning tailored for robotics, we have showcased how these techniques can improve efficiency and adaptability in resource-constrained settings.

Future research should focus on scaling zero-waste machine learning techniques to more extensive and diverse datasets. Ensuring that these methods generalize well across different applications and domains will be crucial for their widespread adoption. Integrating zero-waste machine learning with emerging technologies such as edge computing, Internet of Things, and federated learning could further enhance the efficiency and applicability of our methods. Exploring these synergies will open new avenues for research and development.

Further refinement of continual learning models is necessary to address the limitations related to catastrophic forgetting and data drift. Developing more robust techniques for knowledge distillation and feature alignment will be key areas of focus. As machine learning models continue to grow in complexity, it is essential to consider their ethical and environmental impacts. Future research should include comprehensive evaluations of the carbon footprint and energy consumption of machine learning models, promoting the development of more sustainable AI technologies.

In conclusion, zero-waste machine learning represents a significant step towards more sustainable and efficient AI. By continuing to refine these techniques and exploring their integration with other technologies, we can pave the way for a future where machine learning models are both powerful and environmentally conscious.

Acknowledgements

This research was supported by National Science Centre, Poland grant no 2020/39/B/ST6/01511, 2022/45/B/ST6/02817, 2023/50/E/ST6/00469, and 2023/51/D/ST6/02846. Bartłomiej Twardowski acknowledges the grant RYC2021-032765-I. This paper has been supported by the Horizon Europe Programme (HORIZON-CL4-2022-HUMAN-02) under the project "ELIAS: European Lighthouse of AI for Sustainability", GA no. 101120237.

References

[1] H. Ahn, J. Kwak, S. Lim, H. Bang, H. Kim, and T. Moon. Ss-il: Separated softmax for incremental learning. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 844–853, 2021.

[2] R. Aljundi, P. Chakravarty, and T. Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Conference on Computer Vision and Pattern Recognition*, CVPR, 2017.

[3] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 233–248, 2018.

[4] M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[5] A. Douillard, M. Cord, C. Ollion, T. Robert, and E. Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 86–102. Springer, 2020.

[6] W. Fedus, J. Dean, and B. Zoph. A review of sparse expert models in deep learning. *arXiv preprint arXiv:2209.01667*, 2022.

[7] E. Fini, V. G. T. da Costa, X. Alameda-Pineda, E. Ricci, K. Alahari, and J. Mairal. Self-supervised models are continual learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[8] A. Gomez-Villa, B. Twardowski, L. Yu, A. D. Bagdanov, and J. van de Weijer. Continually learning self-supervised representations with projected functional regularization. *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021.

[9] D. Goswami, Y. Liu, B. Twardowski, and J. van de Weijer. Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] D. Goswami, A. Soutif-Cormerais, Y. Liu, S. Kamath, B. Twardowski, J. van de Weijer, et al. Resurrecting old classes with new data for exemplar-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28525–28534, 2024.

[11] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 831–839, 2019.

[12] V. Khan, S. Cygert, K. Deja, T. Trzciński, and B. Twardowski. Looking through the past: better knowledge retention for generative replay in continual learning. *IEEE Access*, 12:45309–45317, 2024.

[13] S. Kim, L. Noci, A. Orvieto, and T. Hofmann. Achieving a better stability-plasticity trade-off via auxiliary networks in continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11930–11939, 2023.

[14] J. Krajewski, J. Ludziejewski, K. Adamczewski, M. Pióro, M. Krutul, S. Antoniak, K. Ciebiera, K. Król, T. Odrzygóźdź, P. Sankowski, et al. Scaling laws for fine-grained mixture of experts. *arXiv preprint arXiv:2402.07871*, 2024.

[15] J. Lee, S. Kim, S. Kim, W. Jo, and H.-J. Yoo. Gst: Group-sparse training for accelerating deep reinforcement learning, 2021.

[16] J. Li, Z. Ji, G. Wang, Q. Wang, and F. Gao. Learning from students: Online contrastive distillation network for general continual learning. In *Proc. 31st Int. Joint Conf. Artif. Intell.*, pages 3215–3221, 2022.

[17] M. Li, E. Yumer, and D. Ramanan. Budgeted training: Rethinking deep neural network training under resource constraints. *CoRR*, abs/1905.04753, 2019.

[18] D. Madaan, J. Yoon, Y. Li, Y. Liu, and S. J. Hwang. Representational continuity for unsupervised continual learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[19] D. Marczak, S. Cygert, T. Trzciński, and B. Twardowski. Revisiting supervision for continual representation learning. *arXiv preprint arXiv:2311.13321*, 2023.

[20] D. Marczak, S. Cygert, T. Trzciński, and B. Twardowski. Magmax: Leveraging model merging for seamless continual learning. *arXiv preprint arXiv: 2407.06322*, 2024.

[21] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. van de Weijer. Class-incremental learning: survey and performance evaluation. *arXiv preprint arXiv:2010.15277*, 2020.

[22] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. Psychology of Learning and Motivation.

[23] A. Pardyl, G. Kurzejamski, J. Olszewski, T. Trzciński, and B. Zieliński. Beyond grids: Exploring elastic input sampling for vision transformers. *arXiv preprint arXiv:2309.13353*, 2023.

[24] A. Pardyl, G. Rypeść, G. Kurzejamski, B. Zieliński, and T. Trzciński. Active visual exploration based on attention-map entropy. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1303–1311, 2023.

[25] A. Pardyl, M. Wronka, M. Wołczyk, K. Adamczewski, T. Trzciński, and B. Zieliński. Adaglimpse: Active visual exploration with arbitrary glimpse position and scale. In *Proceedings of the European Conference on Computer Vision*, 2024.

[26] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2019.01.012. URL https://www.sciencedirect.com/science/article/pii/S0893608019300231.

[27] M. Przewięźlikowski, M. Pyla, B. Zieliński, B. Twardowski, J. Tabor, and M. Śmieja. Augmentation-aware self-supervised learning with conditioned projector. *arXiv preprint arXiv:2306.06082*, 2023.

[28] M. Przewięźlikowski, M. Osial, B. Zieliński, and M. Śmieja. A deep cut into split federated self-supervised learning. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Database*, 2024.

[29] S. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. iCaRL: Incremental Classifier and Representation Learning. In *CVPR*, 2017.

[30] G. Rypeść, S. Cygert, V. Khan, T. Trzcinski, B. M. Zieliński, and B. Twardowski. Divide and not forget: Ensemble of selectively trained experts in continual learning. In *The Twelfth International Conference on Learning Representations*, 2023.

[31] S. Scardapane, M. Scarpiniti, E. Baccarelli, and A. Uncini. Why should we add early exits to neural networks? *Cognitive Computation*, 12(5): 954–966, 2020.

[32] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni. Green ai. 63(12), 2020.

[33] D. Sójka, S. Cygert, B. Twardowski, and T. Trzciński. Ar-tta: A simple method for real-world continual test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3491–3495, 2023.

[34] F. Szatkowski, B. Wójcik, M. Piórczyński, and K. Adamczewski. Sadmoe: Exploiting activation sparsity with dynamic-k gating. *arXiv e-prints*, pages arXiv–2310, 2023.

[35] F. Szatkowski, M. Pyla, M. Przewięźlikowski, S. Cygert, B. Twardowski, and T. Trzciński. Adapt your teacher: Improving knowledge distillation for exemplar-free continual learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1977–1987, 2024.

[36] L. Wang, X. Zhang, H. Su, and J. Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024. doi: 10.1109/TPAMI.2024.3367329.

[37] B. Wójcik, M. Przewiźlikowski, F. Szatkowski, M. Wołczyk, K. Bałazy, B. Krzepkowski, I. Podolak, J. Tabor, M. Śmieja, and T. Trzciński. Zero time waste in pre-trained early exit neural networks. *Neural Networks*, 168:580–601, 2023.

[38] S. Yan, J. Xie, and X. He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3014–3023, 2021.

[39] L. Yu, B. Twardowski, X. Liu, L. Herranz, K. Wang, Y. Cheng, S. Jui, and J. v. d. Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6982–6991, 2020.

[40] Z. Zhang, Y. Lin, Z. Liu, P. Li, M. Sun, and J. Zhou. Moefication: Transformer feed-forward layers are mixtures of experts. *arXiv preprint arXiv:2110.01786*, 2021.