AFRL-RY-WP-TR-2022-0230

AN INTEGRATED NONPARAMETRIC BAYESIAN AND DEEP NEURAL NETWORK FRAMEWORK FOR BIOLOGICALLY-INSPIRED LIFELONG LEARNING

Yoshua Bengio, Matthew Botvinick, Lawrence Carin, Junya Chen, Yoojin Choi, Jonathan Cohen, Neal J. Cohen, Yulai Cong, Aaron Courvill, Ishita Dasgupta, Nathaniel Daw, Mostafa El-Khamy, Zoubin Ghahramani, Ian Goodfellow, Thomas Griffiths, Dennis Hassabis, Ricardo Henao, Christopher Kanan, Sungchul Kim, Ronald Kemker, Sreejan Kumar, Dharshan Kumaran, Jianqiao Li, Kevin Liang, Michael McCloskey, Nikhil Mehtra, Mehdi Mirza, Subrata Mitra, Sherjil Ozair, Jose L. Part, Jean, Pouget-Abadie, German I. Parisi, Piyush Rai, Roger Ratclif, PK Srijith, Christopher Summerfield, Lakshi Varshney, Vinay Kumar Verma, Rui Wang, Sijia Wang, David Warde-Farley, Stefan Wermter, Bing Xu, Tong Yu, Ruiyi Zhang, Handong Zhao, and Miaoyun Zhao

Duke University

SEPTEMBER 2022
Final Report

AIR FORCE RESEARCH LABORATORY
SENSORS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7320
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with The Under Secretary of Defense memorandum dated 24 May 2010 and AFRL/DSO policy clarification email dated 13 January 2020. This report is available to the general public, including foreign nationals.

Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RY-WP-TR-2022-0230 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//Signature//

ASHLEY DEMANGE BRANDEWIE
Program Manager
Sensors Subsystems Branch
Aerospace Components & Subsystems Division

//Signature//

TIMOTHY R. JOHNSON, Chief
Sensors Subsystems Branch
Aerospace Components & Subsystems Division

//Signature//

GENE M. WILKINS, Lt Col, USAF
Deputy Chief, Aerospace Components &
Subsystems Technology Division
Sensors Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

*Disseminated copies will show "//Signature//" stamped or typed above the signature blocks.

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
September 2022	Final	20 February 2018	31 March 2022

4. TITLE AND SUBTITLE

AN INTEGRATED NONPARAMETRIC BAYESIAN AND DEEP NEURAL NETWORK FRAMEWORK FOR BIOLOGICALLY-INSPIRED LIFELONG LEARNING

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8650-18-2-7832	N/A	61101E
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
N/A	N/A	Y1SA

6. AUTHOR(S)

Yoshua Bengio, Matthew Botvinick, Lawrence Carin, Junya Chen, Yoojin Choi, Jonathan Cohen, Neal J. Cohen, Yulai Cong, Aaron Courvill, Ishita Dasgupta, Nathaniel Daw, Mostafa El-Khamy, Zoubin Ghahramani, Ian Goodfellow, Thomas Griffiths, Dennis Hassabis, Ricardo Henao, Christopher Kanan, Sungchul Kim, Ronald Kemker, Sreejan Kumar, Dharshan Kumaran, Jianqiao Li, Kevin Liang, Michael McCloskey, Nikhil Mehtra, Mehdi Mirza, Subrata Mitra, Sherjil Ozair, Jose L. Part, Jean, Pouget-Abadie, German I. Parisi, Piyush Rai, Roger Ratclif, PK Srijith, Christopher Summerfield, Lakshi Varshney, Vinay Kumar Verma, Rui Wang, Sijia Wang, David Warde-Farley, Stefan Wermter, Bing Xu, Tong Yu, Ruiyi Zhang, Handong Zhao, and Miaoyun Zhao

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Duke University
Office of Research Support
2200 West Main St. Suite 710
Durham NC 27705-4677

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'SREPORT NUMBER(S)
Air Force Research Laboratory, Sensors Directorate Wright-Patterson Air Force Base, OH 45433-7320 Air Force Materiel Command, United States Air Forces	Defense Advanced Research Projects Agency (DARPA/MTO) 675 North Randolph Street Arlington, VA 22203	AFRL/RYDR	AFRL-RY-WP-TP-2022-0230

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with The Under Secretary of Defense memorandum dated 24 May 2010 and AFRL/DSO policy clarification email dated 13 January 2020. This report is available to the general public, including foreign nationals. This material is based on research sponsored by the Air Force Research Lab (AFRL) and the Defense Advanced Research Projects Agency (DARPA) under agreement number FA8650-18-2-7832.The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Labs (AFRL), the Defense Advanced Research Projects Agency (DARPA) or the U.S. Government. Report contains color.

14. ABSTRACT

Deep learning, trained primarily on a single task under the assumption of independent and identically dis- tributed (i.i.d.) data, has made enormous progress in recent years. However, when naively trained sequentially on multiple tasks, without revisiting previous tasks, neural networks are known to suffer catastrophic forgetting (McCloskey and Cohen, 1989; Ratcliff, 1990), namely, the ability to perform old tasks is often lost while learning new ones. In contrast, biological life is capable of learning many tasks throughout a lifetime from decidedly non-i.i.d. experiences, acquiring new skills and reusing old ones to learn fresh abilities, all while retaining important previous knowledge. As we strive to make artificial systems increasingly more intelligent, natural life's ability to learn continually is an important capability to emulate. Continual learning (Parisi et al., 2019) has attracted considerable attention recently in machine learning research, and a number of desiderata have emerged. Models should be able to learn multiple tasks sequentially, with the eventual number and complexity of tasks unknown. Importantly, new tasks should be learned without catastrophically forgetting previous ones, ideally without having to keep any data from previous tasks to re-train on. Models should also be capable of positive transfer: previously learned tasks should help with the learning of new tasks. Knowledge transfer between tasks maximizes sample efficiency, with this particularly important when data are scarce. A number of methods address continual learning through expansion, i.e., the model is grown with each additional task. By diverting learning to new network components for each task, these approaches mitigate catastrophic forgetting by design, as previously learned parameters are left undisturbed. A key challenge for these strategies is deciding when and how much to expand the network. While it is typically claimed that this can be tailored to the incoming task, doing so requires human estimation of how much expansion is needed, which is not a straightforward process. Instead, a preset, constant expansion is commonly employed for each new task. Alternatively, we could consider either a dynamic, data driven, expansion of the model or a modular approach to model growth that enables the development of a framework to build compact models for continual learning in which the size of the models efficiently scales with the (ideally ever) increasing number of tasks while mitigating catastrophic forgetting. Moreover, we seek to develop a framework in a way that is generalizable to different continual learning tasks, e.g., classification, generative processes for images and natural language processing sequence labeling, i.e., named entity recognition. In a continual learning setting, we are presented with a sequence of tasks with a predefined purpose, but such that each task consists of a distinct dataset. The main objective is to build model(s) that perform as consistently as possible across distinct tasks while i) reusing information from previous tasks, and ii) preventing the model to grow uncontrollably (in size). Our approach consists of building one model per task, however, such that most of the components of the model are shared across tasks (global) and the remaining few are task specific (local), thus allowing information sharing and controlled growth. Consequently, we seek to develop task specific (deep learning) models with global and task-specific parameters to enable effective and efficient continual learning.

15. SUBJECT TERMS

lifelong learning, nonparametric, Bayesian, continuous learning, deep neural networks, biologically-inspired

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
Unclassified	Unclassified	Unclassified	SAR	20

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Ashley Demange	

PREVIOUS EDITION IS OBSOLETE.

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Table of Contents

Section **Page**

List of Tables

Table **Page**

1 INTRODUCTION

1.1 Background

Deep learning, trained primarily on a single task under the assumption of independent and identically distributed (i.i.d.) data, has made enormous progress in recent years. However, when naively trained sequentially on multiple tasks, without revisiting previous tasks, neural networks are known to suffer catastrophic forgetting (McCloskey and Cohen, 1989; Ratcliff, 1990), namely, the ability to perform old tasks is often lost while learning new ones. In contrast, biological life is capable of learning many tasks throughout a lifetime from decidedly non-i.i.d. experiences, acquiring new skills and reusing old ones to learn fresh abilities, all while retaining important previous knowledge. As we strive to make artificial systems increasingly more intelligent, natural life's ability to learn continually is an important capability to emulate.

Continual learning (Parisi et al., 2019) has attracted considerable attention recently in machine learning research, and a number of desiderata have emerged. Models should be able to learn multiple tasks sequentially, with the eventual number and complexity of tasks unknown. Importantly, new tasks should be learned without catastrophically forgetting previous ones, ideally without having to keep any data from previous tasks to re-train on. Models should also be capable of positive transfer: previously learned tasks should help with the learning of new tasks. Knowledge transfer between tasks maximizes sample efficiency, with this particularly important when data are scarce.

A number of methods address continual learning through expansion, i.e., the model is grown with each additional task. By diverting learning to new network components for each task, these approaches mitigate catastrophic forgetting by design, as previously learned parameters are left undisturbed. A key challenge for these strategies is deciding when and how much to expand the network. While it is typically claimed that this can be tailored to the incoming task, doing so requires human estimation of how much expansion is needed, which is not a straightforward process. Instead, a preset, constant expansion is commonly employed for each new task. Alternatively, we could consider either a dynamic, data driven, expansion of the model or a modular approach to model growth that enables the development of a framework to build compact models for continual learning in which the size of the models efficiently scales with the (ideally ever) increasing number of tasks while mitigating catastrophic forgetting. Moreover, we seek to develop a framework in a way that is generalizable to different continual learning tasks, e.g., classification, generative processes for images and natural language processing sequence labeling, i.e., named entity recognition.

In a continual learning setting, we are presented with a sequence of tasks with a predefined purpose, but such that each task consists of a distinct dataset. The main objective is to build model(s) that perform as consistently as possible across distinct tasks while i) reusing information from previous tasks, and ii) preventing the model to grow uncontrollably (in size). Our approach consists of building one model per task, however, such that most of the components of the model are shared across tasks (global) and the remaining few are task specific (local), thus allowing information sharing and controlled growth. Consequently, we seek to develop task specific (deep learning) models with global and task-specific parameters to enable effective and efficient continual learning.

The contributions from our work to the DARPA Lifelong Learning Machines (L2M) program and to the continual learning community are listed below.

- Developed continual adaptation models for generative adversarial networks (Cong et al., 2020).
    - Impact: The proposed framework outperforms state-of-the-art methods with significantly less parameters and computational cost.
- Developed continual learning using Bayesian nonparametric dictionary weight factors (Mehta et al., 2021).
    - Impact: The first heuristic-free approach for adaptive (data driven) model expansion in continual learning.
- Developed a meta-learning framework for structured task distributions in humans and machines (Kumar et al., 2020).
    - Impact: We discovered a double dissociation in which humans do better in structured (compositional) tasks whereas agents (machines) do better in statistical (non-compositional) tasks despite having comparable complexity. •
- Developed efficient feature transformations (EFTs) for discriminative and generative continual learning (Verma et al., 2021).
    - Impact: EFTS minimize parameter count for new tasks (< 5% than base model) while allowing for task prediction in class incremental settings.
- Developed a continual learning approach for generative adversarial networks (GANs), by designing and leveraging parameter-efficient feature map transformations.
    - Impact: The proposed approach provides a memory-efficient way to perform effective continual data generation and we show that the feature-map-transformation approach outperforms state-of-the-art methods for continually-learned GANs, with substantially fewer parameters (Varshney et al., 2021).
- Developed the first work of few-shot class-incremental learning for NER (Wang et al., 2022a).
    - Impact: The proposed frameworks learns to recognize new entity classes with minimal labeled data.
- Developed Structured Sparse Convolution (SSC), that leverages the inherent structure in images to reduce the parameters in the convolutional filters (Verma et al., 2022).
    - Impact: SSC filters, unlike existing approaches, require no additional pruning during or after training.

- Developed a new task continual learning framework that does not assume that the sequence of tasks are distinct or unique, thus requiring a task similarity identification module (Wang et al., 2022b).
    - Impact: We identify similar tasks without the need for training a new model, by leveraging a task similarity metric, which in practice results in high task similarity identification accuracy.
- Produced 9 academic scientific contributions in the machine learning field, 7 of which have been published and 2 currently under review.
    - Impact: Our papers have appeared in top machine learning venues such as NeurIPS, ICLR, ICML, AISTATS and ACL. Moreover, all of our published methodologies have publicly available source code. See Table 1 for details.
- Contributed to the M21 Evaluation Classification benchmark with our EFT framework (Verma et al., 2021).
    - Impact: Demonstrated competitive performance across different metrics, namely, 96.6 ± 0.23 Top-1 accuracy, 0.86 ± 0.23 sample efficiency, 0.93 ± 0.01 performance relative to a single task expert, 1.21± forward transfer ratio, and 0.99 ± 0.00 backward transfer ratio.

Below, we provide a brief summary for each of the contribution followed by a list of the publicly available software packages resulting from our project and a complete list references to which we refer the reader for full methodological details, experimental settings and empirical results.

2 CONTRIBUTIONS AND PRODUCTS

2.1 Leveraging Pretrained GANs for Generation with Limited DataConceptual Design Methodology

Recent research has demonstrated the increasing power of generative adversarial networks (GANs) (Goodfellow et al., 2014) to generate high-quality samples, that are often indistinguishable from real data This demonstrates the capability of GANs to exploit the valuable information within the underlying data distribution.

Though many powerful GAN models pretrained on large scale datasets have been released, few efforts have been made to take advantage of the valuable information within those models to facilitate downstream tasks This shows a clear contrast with the popularity of transfer learning for discriminative tasks, e.g., to reuse the feature extractor of a pretrained classifier, and transfer learning in natural language processing, e.g., to reuse the expensively-pretrained BERT model.

Motivated by the significant value of released pretrained GAN models, we propose to leverage the information therein to facilitate downstream tasks in a target domain with limited training data. This situation arises frequently due to expensive data collection or privacy issues that may arise in medical or biological applications. We concentrate on the challenging scenario of GAN model development when limited training data are available. One key observation motivating our method is that a well-trained GAN can generate realistic images not observed in the training dataset, demonstrating the generalization ability of GANs to capture the training data manifold. Likely arising from novel combinations of information, attributes and styles (see stunning illustrations in StyleGAN), this generalization of GANs is extremely appealing for scenarios in which there are limited data. For example, GANs can be used to augment the training set via realistic data generation, to alleviate overfitting or provide regularization for classification, segmentation, or detection.

Leveraging insights from the aforementioned transfer learning on discriminative tasks, we posit that the low-level filters of a GAN discriminator pretrained on a large-scale source dataset are likely to be generalizable and hence transferable to various target domains. For a pretrained GAN generator, it is shown that the low-level layers (those close to output observations) capture properties of generally-applicable local patterns like materials, edges, and colors, while the high-level layers (those distant from observations) are associated with more domain-specific semantic aspects of data. We therefore consider transferring/freezing the lowlevel filters from both the generator and discriminator of a pretrained GAN model to facilitate generation in perceptually-distinct target domains with limited training data. As an illustrative example, we consider the widely studied GAN scenario of natural-image generation, although the proposed techniques are general and may be applicable to other domains, such as in medicine or biology.

Products On Leveraging Pretrained GANs for Generation with Limited Data (Zhao et al., 2020) and https://github.com/MiaoyunZhao/GANTransferLimitedData.

2.2 GAN Memory with No Forgetting

Lifelong learning (or continual learning) is a long-standing challenge for machine learning and artificial intelligence systems, concerning the ability of a model to continually learn new knowledge without forgetting previously learned experiences. An important issue associated

with lifelong learning is the notorious catastrophic forgetting of deep neural networks, i.e., training a model with new information severely interferes with previously learned knowledge.

To alleviate catastrophic forgetting, many methods have been proposed, with most focusing on discriminative/classification tasks. Existing methods revealed that generative replay (or pseudo-rehearsal) is an effective and general strategy for lifelong learning. That revelation is anticipated, for if the characteristics of previous data are remembered perfectly, e.g., via realistic generative replay), no forgetting should be expected for lifelong learning. Compared with the Core Set idea, that saves representative samples of previous data, generative replay has advantages in addressing privacy concerns and remembering potentially more complete data information (via the generative process). However, most existing generative replay methods either deliver blurry generated samples or only work well on simple datasets like MNIST. Besides, they often do not scale well to practical situations with high resolution or a long sequence, sometimes even with negative backward transfer. Therefore, it is challenging to continually learn a well-behaved generative replay model, even for moderately complex datasets like CIFAR10.

We seek a realistic generative replay framework to alleviate catastrophic forgetting; going further, we consider developing a realistic generative memory with growing (expressive) power, believed to be a fundamental building block toward general lifelong learning systems. We leverage the popular GAN setup (Goodfellow et al., 2014) as the key component of that generative memory, which we term GAN memory (Cong et al., 2020), because i) GANs have shown remarkable power in synthesizing realistic high-dimensional samples; ii) by modeling the generative process of training data, GANs summarize the data statistical information in the model parameters, consequently also protecting privacy (the original data need not be saved); and iii) a GAN often generates realistic samples not observed in training data, delivering a synthetic data augmentation that potentially benefits better performance of downstream tasks. Distinct from existing methods, our GAN memory leverages transfer learning and (image) style transfer. Its key foundation is a discovery that one can leverage the modified variants of style-transfer techniques to modulate a source generator/discriminator into a powerful generator/discriminator for perceptually-distant target domains, with a limited amount of style parameters. Exploiting that discovery, our GAN memory sequentially modulates (and also transfers knowledge from) a well-behaved base/source GAN model to realistically remember a sequence of (target) generative processes with no forgetting. Note by "well-behaved" we mean the shape of source kernels is well trained. Eempirically, this requirement can be readily satisfied if i) the source model is pretrained on a (moderately) large dataset (e.g., CelebA; often a dense dataset is preferred) and ii) it is sufficiently trained and shows relatively high generation quality. Therefore, many pretrained GANs can be "well-behaved", showing great flexibility in selecting the base/source model. Our experiments show that flexibility roughly means source and target data should have the same data type (e.g., images).

Products GAN Memory with No Forgetting (Cong et al., 2020) and https://github.com/MiaoyunZhao/ GANmemory_LifelongLearning.

2.3 Continual Learning using Bayesian Nonparametric Weight Factors

Naively trained neural networks tend to experience catastrophic forgetting in sequential task settings, where data from previous tasks are unavailable. A number of methods, using various model expansion strategies, have been proposed recently as possible solutions. However,

determining how much to expand the model is left to the practitioner, and often a constant schedule is chosen for simplicity, regardless of how complex the incoming task is.

A number of methods address continual learning through expansion. By diverting learning to new network components for each task, these approaches mitigate catastrophic forgetting by design, as previously learned parameters are left undisturbed. A key challenge for these strategies is deciding when and how much to expand the network. While it is typically claimed that this can be tailored to the incoming task, doing so requires human estimation of how much expansion is needed, which is not a straightforward process. Instead, a preset, constant expansion is commonly employed for each new task. Rather than relying on engineered heuristics, we choose to let the data dictate the model-expansion rate, employing a Bayesian nonparametric approach. Specifically, we couple rank-1 weight factor (WF) dictionary learning with the Indian Buffet Process (IBP) (Ghahramani and Griffiths, 2005), creating a framework we call IBP-WF (Mehta et al., 2021). An IBP-based formulation allows automatic scaling of the network, but only as needed, even if the number or complexity of future tasks is unknown initially. An IBP prior also naturally encourages recycling of previously learned skills, enabling positive transfer between tasks, which other expansion methods tend to either ignore or deal with in a more ad hoc manner. Moreover, Bayesian modeling enables model sampling, allowing for both ensembling models for increased accuracy and uncertainty estimation, which are important but rarely discussed topics in continual learning. The effectiveness of IBP-WF is demonstrated on a number of continual learning tasks, outperforming other methods. We also visualize the weight factor usage across tasks, confirming both sparsity and reuse of these factors.

Products Continual Learning using a Bayesian Nonparametric Dictionary of Weight Factors (Mehta et al., 2021) and https://github.com/nikhil-dce/IBP-WF.

2.4 Meta-Learning of Structured Tasks in Humans and Machines

While machine learning has supported tremendous progress in artificial intelligence, a major weakness, especially in comparison to humans, has been its relative inability to learn structured representations, such as compositional grammar rules, causal graphs, discrete symbolic objects, etc. One way that humans acquire these structured forms of reasoning is via “learning-to-learn”, in which we improve our learning strategies over time to give rise to better reasoning strategies. Inspired by this, researchers have renewed investigations into meta-learning. Under this approach, a model is trained on a family of learning tasks based on structured representations such that they achieve better performance across the task distribution. This approach has demonstrated the acquisition of sophisticated abilities including model-based learning, causal reasoning, compositional generalization, linguistic structure, and theory of mind, all in relatively simple neural network models. The meta-learning approach, along with interaction with designed environments, has also been suggested as a general way to automatically generate artificial intelligence. These approaches have made great strides, and have great promise, toward closing the gap between human and machine learning.

However, we argue that significant challenges remain in how we evaluate whether structured forms of reasoning have indeed been acquired. There are often multiple strategies that can result in good metatest performance, and there is no guarantee a priori that meta-learners will learn the strategies we intend when generating the training distribution. Previous work on meta learning structured representations do partially acknowledge this. In our work, we highlight these

challenges more generally. At the end of the day, meta-learning is simply another learning problem. And similar to any vanilla learning algorithm, metalearners themselves have inductive biases (which we term meta-inductive bias). Note that meta-learning is a way to learn inductive biases for vanilla learning algorithms. We consider the fact the meta-learners themselves have inductive biases that impact the kinds of strategies (and inductive biases) they prefer to learn. A key contribution of our work is to also develop control task environments that are not generated using the same simple recursively applied rules, but are comparable in statistical complexity. We provide a rigorous comparison between human and meta-learning agent behavior in tasks performed in distributions of environments of each type. We show through three different analyses that human behavior is consistent with having learned the structure that results from our compositional rules in the structured environments. In contrast, despite training on distributions that contain this structure, standard meta-learning agents instead prefer (i.e., have a meta-inductive bias toward) more global statistical patterns that are a downstream consequence of these low-dimensional rules. Our results show that simply doing well at meta-test on a tasks in a distribution of structured environments does not necessarily indicate meta-learning of that structure. We therefore argue that architectural inductive biases still play a crucial role in the kinds of structure acquired by meta-learners, and simply embedding the requisite structure in a training task distribution may not be adequate.

Products Meta-Learning of Structured Task Distributions in Humans and Machines (Kumar et al., 2020) and https://github.com/sreejank/Compositional_MetaRL.

2.5 Efficient Feature Transformations for Continual Learning

While deep learning has led to impressive advances in many fields, neural networks still tend to struggle in sequential learning settings, largely due to catastrophic forgetting, i.e., when the training distribution of a model shifts over time, neural networks overwrite previously learned knowledge if not repeatedly revisited during training. Pragmatically, this typically means that data collection must be completed before training a neural network, which can be problematic in settings like reinforcement learning or the real world, which is constantly evolving. Otherwise, the model must constantly be re-trained as new data arrives. This limitation significantly hampers building and deploying intelligent systems in changing environments.

A variety of continual learning methods have been proposed to address this shortcoming. Regularizationbased methods prevent forgetting by constraining model parameters from drifting too far away from previous solutions, but they can also restrict the model's ability to adapt to new tasks, often resulting in suboptimal solutions. Additionally, regularization methods commonly make the assumption that each weight's importance for a task is independent, which may explain why they have difficulty scaling to more complex networks and tasks. Replay methods retain knowledge by rehearsing on data saved from previous tasks. While effective at preventing forgetting, the performance of replay-based approaches is highly dependent on the size and contents of the memory buffer, and in certain strict settings, saving any data at all may not be an option. The nature of this replay buffer also tends to highly bias the model toward recently learned tasks. As the number of tasks grows, performance degrades quickly, especially in large-scale settings.

To overcome these limitations, we propose a compact, task-specific feature map transformation for large scale continual learning, which we call Efficient Feature Transformation (EFT) (Verma

et al., 2021). In particular, we partition the model into global parameters (θ) and task-specific local parameters (τ), with the pair ($\{\theta,\tau\}$) as the optimal parameters for a particular task t. In constructing these local transformations, we leverage efficient convolution operations, maintaining expressivity, while keeping model growth small. We also minimize the impact on the global base architecture, allowing us to use pre-existing architectures, which can be critical for achieving strong performance in large-scale settings. This compact nature of the added transformations also makes EFTs faster to train than comparable methods because we have to update only task-specific parameters. Finally, we propose a strategy for maximizing feature distance to improve task prediction, a critical component for continual learning methods operating in class incremental settings. To show the efficacy and efficiency of the proposed approach, we extensively evaluate our model on a variety of datasets and architectures. In class incremental and task incremental sequential classification settings, EFTs achieve significant performance gains on CIFAR100 and ImageNet with only a minor growth in parameter count and computation. We also evaluate our approach for continual generative modeling, demonstrating a 22.7% relative improvement in FID on the LSUN, CUB-200, and ImageNet datasets compared to recent state-of-the-art models.

Products Efficient Feature Transformations for Discriminative and Generative Continual Learning (Verma et al., 2021) and https://github.com/vkverma01/EFT.

2.6 Continual Adaptation Modules for GANs

Lifelong learning is an innate human characteristic; we continuously acquire knowledge and upgrade our skills. We also accumulate our previous experiences to learn a new task efficiently. However, learning new tasks rarely affects our performance on already-learned tasks. For example, after learning one programming language, if we learn a new such language it is rare that our skill on the first deteriorates. In fact, the knowledge of the previous task(s) often speeds up learning of the subsequent task(s). On the other hand, attaining lifelong learning in deep neural networks is still a challenge. Naive implementation of continual learning with deep learning models suffers from catastrophic forgetting, which makes lifelong or continual learning (CL) difficult. Catastrophic forgetting refers to the situation when a model exhibits a decline in its performance on previously learned tasks, after it learns a new task.

Recently, expansion-based CL models have shown promising results for discriminative (supervised) continual learning settings. These approaches are dynamic in nature and allow the number of network parameters to grow to accommodate new tasks. Moreover, these approaches can also be regularized appropriately by partitioning the previous and new task parameters. Therefore, unlike regularization-based approaches, the model is free to adapt to novel tasks. However, despite their promising performance, the excessive growth in the number of parameters and floating-point (FLOP) requirements are critical concerns.

We propose a simple expansion-based approach for GANs, which continually adapts to novel tasks without forgetting the previously-learned knowledge. The proposed approach considers a base model with global parameters and corresponding global feature maps. While learning each novel task, the base model is expanded to consider a task-specific feature transformation which efficiently adapts a global feature map to a task-specific feature map in each layer. Though the total number of parameters increases due to the additional local/task-specific parameters, this feature-map transformation approach allows the leveraging of efficient architecture design

choices, e.g., group-wise and point-wise convolution) to obtain compactsized task-specific parameters, which controls the growth and keeps the proposed model compact. To show the efficacy of the proposed model, we conduct extensive experiments in various settings on real-world datasets. We show that the proposed approach can sequentially learn a large number of tasks without catastrophic forgetting, while incurring much smaller parameter and FLOP growth compared to the existing continual-learning GAN models. Further, we show that our approach is also applicable to the generativereplay-based discriminative continual learning, e.g., for classification problems. We empirically show that the pseudo-rehearsal provided by the proposed approach shows promising results for generative-replay-based discriminative models. Also, we conduct experiments to demonstrate the effectiveness of considering the task similarity in continual image generation, which we believe can lead to a promising direction for continual learning

Products CAM-GAN: Continual Adaptation Modules for Generative Adversarial Networks (Varshney et al., 2021) and https://github.com/sakshivarshney/CAM-GAN.

2.7 Few-Shot Class-Incremental Learning for NER

Existing models of Named Entity Recognition (NER) are usually trained on a large scale dataset with predefined entity classes, then deployed for entity extraction on the test data without further adaptation or refinement. In practice, data of new entity classes that the NER model has not seen during training arrives constantly, thus it is desirable that the NER model can be incrementally updated over time with knowledge of data for these new classes. In this case, one challenge is that the training data of old entity classes may not be available due to privacy concerns or memory limitations. Then, the model can easily degrade in terms of the performance on old classes when being fine-tuned with only annotations of new entity classes, i.e., catastrophic forgetting. In addressing this problem, previous work in class-incremental learning for NER regularizes the current model by distilling from the previous model trained on old (existing) classes, using text from the training dataset of new classes. However, this requires abundance of data in the new training dataset being used for distillation. Such an assumption is usually unrealistic since the token-level annotations required by NER training are labor-consuming and scarce, especially for the new unseen classes. We study a more realistic setting, i.e., few-shot class incremental learning for NER, where the model i) incrementally learns on new classes with few annotations, and ii) without requiring access to training data for old classes. We propose a framework to enable few-shot class-incremental learning for NER (Wang et al., 2022a). Since the few-shot dataset may not contain enough entities of old classes as replay data for distilling from the previous model, which leads to catastrophic forgetting, we consider generating synthetic data of the old entity classes for distillation. Such data is termed as synthetic replay. Specifically, we generate synthetic data samples of old classes by inverting the NER model. Given the previous model trained on the old classes, we optimize the token embeddings of the synthetic data, so that predictions from the previous model can contain old entity classes, given the synthetic data as input. In this way, the synthetic data is likely to contain entities of old classes, and distilling from the previous model with such data will thus encourage knowledge preservation of old classes. Additionally, to ensure the synthetic (reconstructed) data to be realistic, we propose to leverage the readily available real text data for new classes, via adversarially matching the hidden features of tokens from the synthetic data and those from the real data. Note that the synthetic data generated from such adversarial match with real data will contain semantics that are close to the real text data for new classes. Consequently, compared

with training with only the few samples of new classes, the synthetic data will provide more diverse context that are close to the samples of the few-shot dataset, augmenting the few-shot training for the new classes. Further, with the generated synthetic data, we propose a framework that trains the NER model with annotations of the new classes, while distilling from the previous model with both the synthetic data and real text from the new training data.

Products Few-Shot Class-Incremental Learning for Named Entity Recognition (Wang et al., 2022a).

2.8 Pushing the Efficiency Limit using Structured Sparse Convolutions

Overparameterized deep neural networks (DNNs) are known to generalize well on the test data. However, over-parameterization increases the network size, making DNNs resource hungry and leading to extended training and inference time. This hinders the training and deployment of DNNs on low-end devices and limits the application of DNNs in systems with strict latency requirements. Several efforts have been made to reduce the storage and computational complexity of DNNs using model compression. Network pruning is the most popular approach for model compression. In network pruning, we compress a large neural network by pruning redundant parameters while maintaining the model performance. The pruning approaches can be divided into two categories: unstructured and structured pruning. Unstructured pruning removes redundant connections in the kernel, leading to sparse tensors. Unstructured sparsity produces random connectivity in the neural architecture, causing irregular memory access that adversely impacts the acceleration in hardware platforms.

An open research question concerns how to design a subnetwork without undergoing the expensive multistage process of training, pruning and finetuning. There have been recent attempts to alleviate this issue, involving a one-time neural network pruning at initialization by solving an optimization problem for detecting and removing unimportant connections. Once the sub-network is identified, the model is trained without carrying out further pruning. This procedure of pruning only once is referred to as pruning at initialization or foresight pruning. While these methods can find an approximation to the winning ticket, they have the following limitations hindering their practical applicability: i) The initial optimization procedure still requires large memory, since the optimization process is carried out over the original overparameterized model. ii) The obtained winning ticket is specific to a particular dataset on which they are approximated, i.e., a network pruned using a particular dataset may not perform optimally on a different dataset. iii) These pruning based methods lead to unstructured sparsity in the model. Due to common hardware limitations, it is very difficult to get a practical speedup from unstructured compression.

We design a novel structured sparse convolution (SSC) filter for convolutional layers, requiring significantly fewer parameters compared to standard convolution (Verma et al., 2022). The proposed filter leverages the inherent spatial properties in the images. The commonly used deep convolutional architectures, when coupled with SSC, outperform other state-of-the-art methods that do pruning at initialization. Unlike typical pruning approaches, the proposed architecture is sparse by design and does not require multiple stages of pruning. The sparsity of the architecture is dataset agnostic and leads to better transfer ability of the model when compared to existing state-of-the-art methods that do pruning at initialization. We also show that the proposed filter has implicit orthogonality that ensures minimum filter redundancy at each layer. Additionally,

we show that the proposed filter can be viewed as a generalization of existing efficient convolutional filters used in group-wise convolution (GWC), point-wise convolution (PWC), and depth-wise convolution (DWC). Extensive experiments and ablations on standard benchmarks depict the efficacy of the proposed filter. Moreover, we further compress existing efficient models such as MobileNetv2 and ShuffleNetv2 while achieving performance comparable to the original models.

Products Pushing the Efficiency Limit using Structured Sparse Convolutions (under review at NeurIPS 2022) (Verma et al., 2022).

2.9 Toward Sustainable Continual Learning

Human intelligence is distinguished by the ability to learn new tasks over time while remembering how to perform previously experienced tasks. Continual learning (CL), an advanced machine learning paradigm requiring intelligent agents to continuously learn new knowledge while trying not to forget past knowledge, has a pivotal role in machines imitating human-level intelligence (Hassabis et al., 2017). The main problem in continual learning is catastrophic forgetting (CF) of previous knowledge when new tasks, observed over time, are incorporated to the model via training or (parameter) expansion.

This work discusses CL under a task continual learning (TCL) setting, i.e., that in which data arrives sequentially in groups of tasks. For works under this scenario, the assumption is usually that once a new task is presented, all of its data becomes readily available for batch (offline) training. In this setting, a task is defined as an individual training phase with a new collection of data that belongs to a new (never seen) group of classes, or in general, a new domain. Further, TCL also (implicitly) requires a task identifier during training. However, in practice, when the model has seen enough tasks, a newly arriving batch of data becomes increasingly likely to belong to the same group of classes or domain of a previously seen task. Importantly, most existing works on TCL fail to acknowledge this possibility. Moreover and in general, the task definition or identifier may not be available during training, e.g., the model does not have access to the task description due to (user) privacy concerns. In such case mostly concerning dynamic models, the system has to treat every task as new, thus constantly learning new sets of parameters regardless of task similarity or overlap. This clearly constitutes a suboptimal use of resources (predominantly memory), especially as the number of tasks experienced by the CL system grows.

Our study investigates the aforementioned scenario and makes it an endeavor to create a memory efficient CL system which though focused on image classification tasks, is general and in principle can be readily used toward other applications or data modality settings. We provide a solution for dynamic models to identify similar tasks when no task identifier is provided during the training phase. To the best of our knowledge, the only work that also discusses the learning of a continual learning system with mixed similar and dissimilar tasks is, which proposes a task similarity function to identify previously seen similar tasks, which requires training a reference model every time a new task becomes available. Alternatively, in this work, we identify similar tasks without the need for training a new model, by leveraging a task similarity metric, which in practice results in high task similarity identification accuracy. We also discuss memory usage under challenging scenarios where longer, more realistic, sequences of more than 20 tasks are used. Moreover, our task similarity detection module shows remarkable performance on widely

used computer vision benchmarks, such as CIFAR10, CIFAR100, EMNIST, from which we create sequences of 10 to 100 tasks.

Products Toward Sustainable Continual Learning: Detection and Knowledge Repurposing of Similar Tasks (under review at NeurIPS 2022) (Wang et al., 2022b).

3 SOFTWARE

Table 1 presents a complete list of the software packages made publicly available as version-controlled repositories to encourage reproducibility and further research developments.

Table 1. Software Packages and Source Code Repositories Developed, Tested, and Made Available as Products or Open Source

Title	Link to Repository
On Leveraging Pretrained GANs for Generation with Limited Data (Zhao et al., 2020)	https://github.com/MiaoyunZhao/GANTransferLimitedData
GAN Memory with No Forgetting (Cong et al., 2020)	https://github.com/MiaoyunZhao/GANmemory_LifelongLearning
Efficient Feature Transformations for Discriminative and Generative Continual Learning (Verma et al., 2021)	https://github.com/vkverma01/EFT
Continual Learning using a Bayesian Nonparametric Dictionary of Weight Factors (Mehta et al., 2021)	https://github.com/nikhil-dce/IBP-WF
CAM-GAN: Continual Adaptation Modules for Generative Adversarial Networks (Varshney et al., 2021)	https://github.com/sakshivarshney/CAM-GAN
Meta-Learning of Compositional Task Distributions in Humans and Machines (Kumar et al., 2020)	https://github.com/sreejank/Compositional_MetaRL

4 REFERENCES

1. Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.
2. Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.
3. German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.
4. Yulai Cong, Miaoyun Zhao, Jianqiao Li, Sijia Wang, and Lawrence Carin. Gan memory with no forgetting. *Advances in Neural Information Processing Systems*, 33:16481–16494, 2020.
5. Nikhil Mehta, Kevin Liang, Vinay Kumar Verma, and Lawrence Carin. Continual learning using a bayesian nonparametric dictionary of weight factors. In *International Conference on Artificial Intelligence and Statistics*, pages 100–108. PMLR, 2021.
6. Sreejan Kumar, Ishita Dasgupta, Jonathan Cohen, Nathaniel Daw, and Thomas Griffiths. Meta-learning of structured task distributions in humans and machines. In *International Conference on Learning Representations*, 2020.
7. Vinay Kumar Verma, Kevin J Liang, Nikhil Mehta, Piyush Rai, and Lawrence Carin. Efficient feature transformations for discriminative and generative continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13865–13875, 2021.
8. Sakshi Varshney, Vinay Kumar Verma, PK Srijith, Lawrence Carin, and Piyush Rai. Cam-gan: Continual adaptation modules for generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:15175–15187, 2021.
9. Rui Wang, Tong Yu, Handong Zhao, Sungchul Kim, Subrata Mitra, Ruiyi Zhang, and Ricardo Henao. Fewshot class-incremental learning for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 571–582, 2022a.
10. Vinay Kumar Verma, Ricardo Henao, and Lawrence Carin. Pushing the efficiency limit using structured sparse convolutions, 2022.
11. Sijia Wang, Yoojin Choi, Junya Chen, Mostafa El-Khamy, and Ricardo Henao. Pushing the efficiency limit using structured sparse convolutions, 2022b.
12. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
13. Miaoyun Zhao, Yulai Cong, and Lawrence Carin. On leveraging pretrained gans for generation with limited data. In *International Conference on Machine Learning*, pages 11340–11351. PMLR, 2020.
14. Zoubin Ghahramani and Thomas Griffiths. Infinite latent feature models and the indian buffet process. *Advances in neural information processing systems*, 18, 2005.
15. Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscienceinspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

ACRONYM	DESCRIPTION
CF	Catastrophic Forgetting
CL	Continuous Learning
DARPA	Defense Advanced Research Projects Agency
DNN	Deep Neural Network
DWC	Depth-Wise Convolution
EFT	Efficient Feature Transformations
FLOP	Floating-Point Operations
GAN	Generative Adversarial Network
GWC	Group-Wise Convolution
IBP	Indian Buffet Process
L2M	Lifelong Learning Machines
M21	Milestone 21 for DARPA L2M effort
NER	Named Entity Recognition
PWC	Point-Wise Convolution
SSC	Structured Sparse Convolution
TCL	Task Continual Learning
WF	Weight Factor