

Deep Continual Learning

Tinne Tuytelaars^{*1}, Bing Liu^{*2}, Vincenzo Lomonaco^{*3},
Gido van de Ven^{*4}, and Andrea Cossu^{†5}

- 1 KU Leuven, BE. tinne.tuytelaars@esat.kuleuven.be
- 2 University of Illinois – Chicago, US. liub@uic.edu
- 3 University of Pisa, IT. vincenzo.lomonaco@unipi.it
- 4 KU Leuven, BE. gido.vandeven@kuleuven.be
- 5 University of Pisa, IT. andrea.cossu@sns.it

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 23122 “Deep Continual Learning”. This seminar brought together 26 researchers to discuss open problems and future directions of Continual Learning. The discussion revolved around key properties and the definition of Continual Learning itself, on the way Continual Learning should be evaluated, and on its real-world applications beyond academic research.

Seminar March 19–24, 2023 – <https://www.dagstuhl.de/23122>

2012 ACM Subject Classification Computing methodologies → Learning settings; Computing methodologies → Neural networks

Keywords and phrases continual learning, incremental learning

Digital Object Identifier 10.4230/DagRep.13.3.74


1 Executive Summary

Bing Liu (University of Illinois – Chicago, US)

Vincenzo Lomonaco (University of Pisa, IT)

Tinne Tuytelaars (KU Leuven, BE)

Gido van de Ven (KU Leuven, BE)

License  Creative Commons BY 4.0 International license
© Bing Liu, Vincenzo Lomonaco, Tinne Tuytelaars, and Gido van de Ven

Continual learning, also referred to as lifelong learning, is a sub-field of machine learning that focuses on the challenging problem of incrementally training models for sequentially arriving tasks and/or when data distributions vary over time. Such non-stationarity calls for learning algorithms that can acquire new knowledge over time with minimal forgetting of what they have learned previously, transfer knowledge across tasks, and smoothly adapt to new circumstances as needed. This is in contrast with the traditional setting of machine learning, which typically builds on the premise that all data, both for training and testing, are sampled i.i.d. from a single, stationary data distribution.

Deep learning models in particular are in need of continual learning capabilities. A first reason for this is the strong data-dependence of these models. When trained on a stream of data whose underlying distribution changes over time, deep learning models tend to almost fully adapt to the most recently seen data, thereby “catastrophically” forgetting the skills that have been learned earlier. Second, continual learning capabilities can be especially

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Deep Continual Learning, *Dagstuhl Reports*, Vol. 13, Issue 3, pp. 74–91

Editors: T. Tuytelaars, B. Liu, V. Lomonaco, G. van de Ven, and A. Cossu



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

beneficial for deep learning models as they can help deal with the very long training time of these models. The current practice in industry is to re-train on a regular basis to add new skills and to prevent the knowledge learned previously from being outdated. Re-training is time inefficient, unsustainable and sub-optimal. Freezing the feature extraction layers is often not an option, as the power of deep learning in many challenging applications, be it in computer vision, natural language processing or audio processing, hinges on the learned representations.

The objective of the seminar was to bring together world-class researchers in the field of deep continual learning, as well as in the related fields of online learning, meta-learning, Bayesian deep learning, robotics and neuroscience, to discuss and to brainstorm, and to set the research agenda for years to come.

During the seminar, participants presented new ideas and recent findings from their research in plenary sessions that triggered many interesting discussions. There were also several tutorials that helped create a shared understanding of similarities and differences between continual learning and other related fields. Specifically, the relation with online learning and streaming learning was discussed in detail. Furthermore, there were several breakout discussion sessions in which open research questions and points of controversy within the continual learning field were discussed. An important outcome of the seminar is the shared feeling that the scope and potential benefit of the research on deep continual learning should be communicated better to computer scientists outside of our subfield. Following up on this, most of the seminar participants are currently collaborating on writing a perspective article to do so.

2 Table of Contents

Executive Summary

Bing Liu, Vincenzo Lomonaco, Tinne Tuytelaars, and Gido van de Ven 74

Overview of Tutorials

Deep Continual Learning
Gido van de Ven 78

Neuroscience inspired continual learning
Dhireesha Kudithipudi 78

A Light Introduction to Online Algorithms and Concept Drift
Joao Gama 78

Overview of Talks

Replay free representation learning
Rahaf Aljundi 79

Reinventing science as a long-term ensemble learning machine
Matthias Bethge 79

Beyond Forgetting with Continual Pre-Training
Andrea Cossu 80

Explaining Change – Towards Online Explanations on Data Streams
Fabian Fumagalli 80

XPM-Explainable Predictive Maintenance
Joao Gama 80

Replay-based continual learning with constant time complexity
Alexander Geppert 81

Lifelong Learning: Where Do We Go Next?
Tyler Hayes 81

Uncertainty Representation in Continual and Online Learning: Challenges and Opportunities
Eyke Hüllermeier 82

Let's Get Continual Learning Out of the Lab!
Christopher Kanan 82

Continual domain generalization/adaptation
Tatsuya Konishi 82

Continual Learning Theory?
Christoph H. Lampert 83

Class-Incremental Learning and Open-world Continual Learning
Bing Liu 83

Learning Continually from Compressed Knowledge and Skills
Vincenzo Lomonaco 83

Into the Unknown: Premises, Pitfalls, Promises
Martin Mundt 84

Role of CL in large scale learning <i>Razvan Pascanu</i>	84
Transfer-learning-based exemplar-free incremental learning <i>Adrian Popescu</i>	84
Repetition and Reconstruction in Continual Learning <i>James M. Rehg</i>	85
Using Generative Models for Continual Learning <i>Andreas Toliás</i>	85
How we applied Continual Learning for Long-sequence Neural Rendering <i>Tinne Tuytelaars</i>	86
The “Stability Gap” <i>Gido van de Ven</i>	86
Projected Functional Regularization for Continual Learning <i>Joost van de Weijer</i>	87
Knowledge Accumulation in Continually Learned Representations and the Issue of Feature Forgetting <i>Eli Verwimp</i>	87
Prediction Error-based Classification for Class-Incremental Learning <i>Michal Zajac</i>	87
Working groups	
Evaluation (Part 1) <i>Alexander Geppert</i>	88
Evaluation (Part 2) <i>Andrea Cossu</i>	88
Reproducibility <i>Alexander Geppert</i>	89
Online Learning and Continual Learning <i>Andrea Cossu</i>	89
Optimization in continual learning <i>Vincenzo Lomonaco</i>	89
Participants	91

3 Overview of Tutorials

3.1 Deep Continual Learning


Gido van de Ven (KU Leuven, BE)

License  Creative Commons BY 4.0 International license
© Gido van de Ven

Incrementally learning new information from a non-stationary stream of data, referred to as “continual learning”, is a key feature of natural intelligence, but an open challenge for deep learning. For example, standard deep neural networks tend to catastrophically forget previous tasks or data distributions when trained on a new one. Enabling these networks to incrementally learn, and retain, information from different contexts has become a topic of intense research. In the first half of this tutorial I introduce the continual learning problem. After covering some key terminology, I discuss three different types of continual learning, each with their own set of challenges: task-incremental, domain-incremental and class-incremental learning. I also cover the distinction between task-based and task-free continual learning. I end this part of the tutorial with a general framework for continual learning unifying these different aspects. In the second half of the tutorial I review approaches that have been proposed for addressing the continual learning problem. I do this at the level of computational strategies, distinguishing between the following: (1) using context-specific components, (2) parameter regularization, (3) functional regularization, (4) replay, and (5) template-based classification. For each strategy I highlight two representative example methods.

3.2 Neuroscience inspired continual learning

Dhireesha Kudithipudi (University of Texas – San Antonio, US)

License  Creative Commons BY 4.0 International license
© Dhireesha Kudithipudi

Continual learning is commonplace in humans and other mammals, but has proven difficult to achieve in artificial systems. By leveraging findings from neuroscience we can make progress towards designing continual learning AI. In this tutorial, we present the key features desirable in a continual learning system and how brain-inspired mechanisms for regularization, dynamic architectures and replay can be realized in artificial systems. Specific examples of metaplasticity, synaptic consolidation and neurogenesis are delved into closely. A canonical theme in these neuro-inspired approaches is that they can be performed at extreme low energy. We present a case for such framework.

3.3 A Light Introduction to Online Algorithms and Concept Drift

Joao Gama (INESC TEC – Porto, PT)

License  Creative Commons BY 4.0 International license
© Joao Gama

In this tutorial we present the basic concepts about online learning from data streams. In the first part of the tutorial, we present Hoeffding algorithms for learning decision trees, regression trees, decision and regression rules, bagging, boosting and random forests. The

second part covers concept drift topics. We discuss data management, detection methods, adaptation methods and model management methods to deal with non-stationary data. We present few illustrative algorithms for explicitly drift detection. We end the tutorial, presenting open-source software available that implement most of the algorithms we discuss in the tutorial.

4 Overview of Talks

4.1 Replay free representation learning

Rahaf Aljundi (Toyota Motor Europe – Zaventem, BE)

License © Creative Commons BY 4.0 International license
© Rahaf Aljundi

This talk will focus on the effectiveness of representation learning as opposed to directly optimizing a classifier. With that we aim for replay free efficient methods and we explore how and when to adapt pretrained representations.

4.2 Reinventing science as a long-term ensemble learning machine


Matthias Bethge (Universität Tübingen, DE)

License © Creative Commons BY 4.0 International license
© Matthias Bethge

Foundation models such as GPT-4 have demonstrated striking task generality based on massively increasing the amount of training data and model capacity. The quest for unifying models in science as well as the strong grounding in empirical data and evaluation of models raises the question for opportunities and limitations of the current avenue to such foundation models. Despite the widespread scientific impact of models like Alphafold-2 and MedPALM, a large range of scientific questions are still hard to approach within a unified benchmarking approach. The impressive flexibility of recent large language models due to their zero-shot and in-context adaptation capabilities may help overcome this limitation – however, they are only developed by a small group of people and not designed for easy updating. In science we want models that are revisable by anyone, calling for the possibility of continual model evaluation and updating. In order to achieve such an efficient continual model extensibility ($M_{n+1} = f(M_n, U)$, with n arbitrary large), I argue that the key challenge is to modularize continual learning without sacrificing the power and scalability of current LLMs. A large part of current continual learning research aims at developing a better understanding of how stochastic gradient descent (SGD) learning is affected by the curriculum, i.e. by the order at which the data is processed. The focus lies on avoiding catastrophic forgetting rather than achieving modularity. I argue to focus on “Scalable Compositionality Discovery” (SCD) as the key challenge to overcome the limitations of collective continual foundation model building that could (1) make large scale data-driven learning ubiquitously useful for science, and (2) solve the credit assignment problem underlying catastrophic forgetting. I conclude with a super brief sketch of how current model benchmarking can be turned into an integrative ensemble learning approach for collective model building.

4.3 Beyond Forgetting with Continual Pre-Training

Andrea Cossu (University of Pisa, IT)

License  Creative Commons BY 4.0 International license
© Andrea Cossu

Pre-trained models are widely used in continual learning. They allow to leverage general and robust representations that can be then fine-tuned during continual learning. However, the existing continual learning scenarios do not fully exploit the potential of pre-trained models. We will present the Continual Pre-Training scenario, which keeps a pre-trained model updated over time. Under appropriate conditions, Continual Pre-Training proves to be surprisingly resilient to forgetting. We will discuss the relationship between Continual Pre-Training and existing paradigms, as well as its potential impact on both continual learning research and applications.

4.4 Explaining Change – Towards Online Explanations on Data Streams


Fabian Fumagalli (Universität Bielefeld, DE)

License  Creative Commons BY 4.0 International license
© Fabian Fumagalli

Recent advances in deep learning methods have shown impressive improvements in predictive accuracy in many tasks at the cost of interpretability. Explainable Artificial Intelligence (XAI) has emerged to understand the reasoning of such black-box models. However, XAI has mainly considered static learning scenarios, whereas many real-world applications require dynamic models that constantly adapt over time. In extreme cases, models learn incrementally on a data stream, where observations are used only once to update the model and are then discarded. In this talk, we present incremental SAGE, an efficient incremental variant of the well-established model-agnostic global feature importance method SAGE (Covert et al., 2020). We describe a general framework to efficiently compute these feature importance values in a data stream scenario with concept drift and present an open-source implementation of our method. Beyond incremental learning on data streams, we explore and discuss further applications of incremental XAI in other areas of deep continual learning.

4.5 XPM-Explainable Predictive Maintenance

Joao Gama (INESC TEC – Porto, PT)

License  Creative Commons BY 4.0 International license
© Joao Gama

Predictive Maintenance applications are increasingly complex, with interactions between many components. Black-box models, based on deep-learning techniques, are popular approaches due to their predictive accuracy. This talk presents a neural-symbolic architecture that uses an online rule-learning algorithm to explain when the black-box model predicts failures. The proposed system solves two problems in parallel: (i) anomaly detection and (ii) explanation of the anomaly. For the first problem, we use an unsupervised state-of-the-art autoencoder. For the second problem, we train a rule learning system that learns a mapping from the

input features to the reconstruction error of the autoencoder. Both systems run online and in parallel. The autoencoder signals an alarm for the examples with a reconstruction error that exceeds a threshold. The causes of the signal alarm are hard to understand by humans because they are the result of a non-linear combination of the sensor data. The rule that triggers that example describes the relationship between the input features and the autoencoder's reconstruction error. The rule explains the failure signal in that it indicates which sensors contribute to the alarm and allows the identification of the component involved in the failure. The system can present global explanations that model the black-box model and local explanations that describe why the black-box model predicts a failure. We evaluate the proposed system in a real-world case study of Metro do Porto.

4.6 Replay-based continual learning with constant time complexity

Alexander Geppert (Hochschule für Angewandte Wissenschaften Fulda, DE)

License © Creative Commons BY 4.0 International license
© Alexander Geppert

This talk describes a new CL approach based on generative replay (GR). The salient point is that GR time complexity does not increase over time but stays constant, under some mild assumptions.

GR protects existing knowledge by having auxiliary generator networks replay/generate samples from previous sub-tasks. At each sub-task, the union of new and replayed data is then used for training a new model (or scholar). The innovation we propose is to replay only samples that cause conflicts with new data. In contrast, existing GR approaches replay all of the previously acquired knowledge, which leads to an unbounded increase in computation time.

In order to achieve constant time-complexity GR, we propose to use a GMM-based generator/solver structure that allows selective modification of existing knowledge only where it overlaps with new data. The same generator/solver can be queried with new data, selectively replaying samples from overlapping areas only. Thus, we can maintain a constant ratio between new and generated samples, irrespective of the number of sub-tasks already processed.

We tested the proposed strategy on CL problems from visual classification and found that it compares very favorably to VAE-based GR, despite vastly inferior model complexity.

4.7 Lifelong Learning: Where Do We Go Next?

Tyler Hayes (NAVER Labs Europe – Meylan, FR)

License © Creative Commons BY 4.0 International license
© Tyler Hayes

The last few years have seen immense progress in developing lifelong learning models capable of performing tasks such as incremental image classification (e.g., on ImageNet). However, today's lifelong learning models still lack the necessary capabilities to generalize to and discover novel concepts in an open world. In this talk, I outline several future research directions for lifelong learning, what advantages they offer, and initial research questions to be addressed in these areas.

4.8 Uncertainty Representation in Continual and Online Learning: Challenges and Opportunities


Eyke Hüllermeier (LMU München, DE)

License  Creative Commons BY 4.0 International license
© Eyke Hüllermeier

The notion of uncertainty has recently drawn increasing attention in machine learning research due to the field's burgeoning relevance for practical applications, many of which have safety requirements. This talk will elaborate on the representation and adequate handling of (predictive) uncertainty in (supervised) machine learning. In this regard, the usefulness of distinguishing between two important types of uncertainty, often referred to as aleatoric and epistemic, will be elucidated. Finally, some challenges and opportunities of uncertainty handling in the realm of continual learning will be highlighted.

4.9 Let's Get Continual Learning Out of the Lab!

Christopher Kanan (University of Rochester, US)

License  Creative Commons BY 4.0 International license
© Christopher Kanan

Continual learning has been a heavily researched topic over the past six years, with mitigation of catastrophic forgetting being the primary focus. However, I argue that there is a lot more to continual learning than catastrophic forgetting. Moreover, many of the systems being created do not have the characteristics needed for real-world applications. In this talk, I outline four real-world applications for continual learning: 1) efficiently updating large neural network models, 2) learning on embedded devices, 3) enabling more efficient learning algorithms, and 4) facilitating applications such as open world learning. I describe the properties that an ideal continual learning method would need for these problem areas. I then describe a new algorithm from my research group that attempts to meet many of these criteria.

4.10 Continual domain generalization/adaptation

Tatsuya Konishi (KDDI – Saitama, JP)

License  Creative Commons BY 4.0 International license
© Tatsuya Konishi

Many studies have been done for the domain-shift in continual learning. Some papers have tackled this issue by techniques of test-time adaptation, but those methods depend on an already pre-trained model. We believe it would be beneficial to propose a continual pre-training procedure that is aware of possible future domain-shifts from the perspective of both domain generalization and adaptation. We present preliminary results about this problem.

4.11 Continual Learning Theory?

Christoph H. Lampert (IST Austria – Klosterneuburg, AT)

License  Creative Commons BY 4.0 International license
© Christoph H. Lampert

We introduce some of the fundamental concepts and results of statistical learning theory in the PAC-Bayesian setting. Afterwards, we discuss the special case of representation learning from multiple tasks and –time permitting– extensions to the continual learning regime.

4.12 Class-Incremental Learning and Open-world Continual Learning


Bing Liu (University of Illinois – Chicago, US)

License  Creative Commons BY 4.0 International license
© Bing Liu

Continual learning (CL) learns a sequence of tasks incrementally. A challenging setting of CL is class incremental learning (CIL). While it is well known that catastrophic forgetting (CF) is a major difficulty for CIL, we argue that there is also an equally challenging problem of inter-task class separation (ICS). This talk first presents a theoretical investigation on how to solve the CIL problem. The key results are (1) that the necessary and sufficient conditions for good CIL are good within-task prediction and task-id prediction, and (2) that task-id prediction is correlated with out-of-distribution (OOD) detection. The theory thus states that good within-task prediction and OOD detection are necessary and sufficient conditions for good CIL. This theory is also applicable to open-world learning. I will then present a general framework for open world learning, called Self-initiated Open-world continual Learning & Adaptation (SOLA).

4.13 Learning Continually from Compressed Knowledge and Skills

Vincenzo Lomonaco (University of Pisa, IT)

License  Creative Commons BY 4.0 International license
© Vincenzo Lomonaco

Learning continually from non-stationary data streams is a challenging research topic of growing popularity in the last few years. Being able to learn, adapt, and generalize continually in an efficient, effective, and scalable way is fundamental for a sustainable development of Artificial Intelligent systems. However, an agent-centric view of continual learning requires learning directly from raw data (i.e. by trial and error), which limits the efficiency, effectiveness and privacy of current solutions. Instead, we argue that continual learning systems should exploit the availability of compressed knowledge and skills in the form of trained models made globally available from a decentralized network of independent agents. In this talk, we suggest to investigate this new paradigm, also known as “Ex-Model Continual Learning” (ExML), where an agent learns from a sequence of previously trained models instead of raw data.

4.14 Into the Unknown: Premises, Pitfalls, Promises

Martin Mundt (TU Darmstadt, DE)

License  Creative Commons BY 4.0 International license
© Martin Mundt

Deep neural networks excel in many areas seems to be a common conclusion drawn from their success on predefined training and dedicated test set data. When moving beyond this paradigm to learning data sequentially, we seem to draw similar conclusions when we find techniques that transfer knowledge and avoid forgetting over time. However, the real world is full of novel and unknown experiences, its complexity cannot be captured by benchmarking knowledge accumulation alone. In this presentation, I will talk upon design of lifelong learning systems in open worlds. These systems are able to robustly deal with novel situations and incorporate new knowledge from data streams over time as humans do. To this end, I will dive into symbiotic mechanisms for deep models to prevent erratic predictions for unknown concepts, actively query new data, and avoid rapidly forgetting past knowledge when learning on new tasks. I will then finish by revisiting the challenge of evaluation of such complex systems and means to promote reproducibility.

4.15 Role of CL in large scale learning


Razvan Pascanu (DeepMind – London, GB)

License  Creative Commons BY 4.0 International license
© Razvan Pascanu

In this talk I will focus on what could be the goals of Continual Learning, particularly for typical Deep Learning settings. Firstly I will show that deep learning is fundamentally computationally inefficient due to interference or forgetting, which leads to concepts being learnt sequentially even if they are all present at once. This leads to the hypothesis that learning efficiently might require us to figure out how to learn continually, which can be a well formed target for continual learning. Afterwards I will describe some limitations of typical train-test setup, and argue that continual learning can be seen as a change of perspective that can allow rephrasing several concepts and find new ways to address these limitations. For example, it can alter how we think about evaluation at large scale. Finally I will enumerate some research directions for continual learning that I feel are receiving less attention than they should.

4.16 Transfer-learning-based exemplar-free incremental learning

Adrian Popescu (CEA LIST – Nano-INNOV, FR)

License  Creative Commons BY 4.0 International license
© Adrian Popescu

The effect of catastrophic forgetting is strong when storage of exemplars for past classes is impossible. Most existing methods designed for this scenario implement variants of fine tuning with knowledge distillation to reduce forgetting. This presentation discusses transfer-learning-based methods, which use a fixed model learned with the initial classes and

the update only the classification layer during the incremental process. Experiments with different datasets and incremental splits show that transfer-based methods obtain competitive performance, while being much faster to train than mainstream fine-tuning methods. These results resonate with past works which show that simple methods can be highly effective in incremental learning, and question our progress in the exemplar-free scenario.

4.17 Repetition and Reconstruction in Continual Learning


James M. Rehg (Georgia Institute of Technology – Atlanta, US)

License  Creative Commons BY 4.0 International license
© James M. Rehg

This talk describes some recent advances that shed light on the role of forgetting in continual learning (CL). First, we introduce CL with repeated exposures, in which sequentially-presented concepts are allowed to repeat a small number of times. We show that simple memory-based CL methods can converge to accuracy approaching batch learning in this setting. Second, we introduce a class of continual reconstruction tasks which do not suffer from forgetting in either the single or repeated exposure settings. This finding is based on a novel SOTA method for single image shape reconstruction (Thai 20). We further show that shape reconstruction can be used as a proxy task for continual classification, resulting in SOTA performance. We close by developing some links between 3D reconstruction and self-supervised learning.

4.18 Using Generative Models for Continual Learning

Andreas Tolias (Baylor College of Medicine – Houston, US)

License  Creative Commons BY 4.0 International license
© Andreas Tolias

Continual learning is a key feature of natural intelligence, but an unsolved problem in deep learning. Particularly challenging for deep neural networks is “class-incremental learning”, whereby a network must learn to distinguish between classes that are not observed together. In this short talk, I will discuss two ways in which generative models can be used to address the class-incremental learning problem. The first one is “generative replay” (e.g., van de Ven et al., 2020 Nat Commun). With this approach, typically two models are learned: a classifier network and an additional generative model. Then, when learning new classes, samples from the generative model are interleaved – or replayed – along with the training data of the new classes. The second approach is “generative classification” (e.g., van de Ven et al., 2021 CVPR-W). With this approach, rather than using a generative model indirectly for generating samples to train a discriminative classifier on (as is done with generative replay), the generative model is used directly to perform classification using Bayes’ rule.

4.19 How we applied Continual Learning for Long-sequence Neural Rendering


Tinne Tuytelaars (KU Leuven, BE)

License  Creative Commons BY 4.0 International license
© Tinne Tuytelaars

The focus in most literature on Continual Learning lies on image classification problems. In that context, it makes sense to reason about the learning process in terms of the learned representation (penultimate layer of the network), which is the part that is shared over all tasks. It's often argued that a good representation makes it easy to learn new tasks and leads to minimal forgetting. It is not clear though how these observations generalize to continual learning beyond classification tasks. In this work, we apply continual learning in a very different context, that of neural rendering. We argue there is an opportunity for continual learning in this setting if one wants to process long-sequences, as it is impossible to load all views for all timestamps in memory simultaneously, multiple views of the same timestamp are required in the same batch to learn effectively from intersecting rays, and repeatedly decoding and transferring views to/from memory is expensive. The standard architecture used for Neural Radiance Fields is not well suited for continual learning though, as the model itself is basically the representation: all properties of the dynamic scene are stored implicitly in the model parameters. Instead, we show that switching to an image-based rendering pipeline gives much better results, as it allows a good balance between what to store implicitly (the learned part) and what to store explicitly (the training views). This results in better transfer and good results when combined with a ray-based replay scheme. This, for the first time, makes it possible to handle dynamic scenes of 1000+ frames with low storage requirements and good quality.

4.20 The “Stability Gap”

Gido van de Ven (KU Leuven, BE)

License  Creative Commons BY 4.0 International license
© Gido van de Ven

Continually learning from a stream of non-stationary data is challenging for deep neural networks. When these networks are trained on something new, they tend to quickly forget what was learned before. In recent years, considerable progress has been made towards overcoming such catastrophic forgetting, predominantly thanks to an approach called “replay”. With replay, examples of past tasks are stored in a memory buffer and later revisited when the network is trained on new tasks. Strikingly, even with just a handful of stored samples per task, replay still performs very strongly. Replay seems to work so well that it has even been suggested that forgetting is no longer a major issue in continual learning. A recent discovery of us challenges this (De Lange et al., 2023 ICLR). Surprisingly, we found that replay still suffers from substantial forgetting when starting to learn a new task, but that this forgetting is temporary and followed by a phase of performance recovery. We demonstrate empirically that this phenomenon of transient forgetting – which we call the “stability gap” – is consistently observed with replay, even in relatively simple toy problems.

4.21 Projected Functional Regularization for Continual Learning

Joost van de Weijer (Computer Vision Center – Barcelona, ES)

License  Creative Commons BY 4.0 International license
© Joost van de Weijer

Recent self-supervised learning methods are able to learn high-quality image representations and are closing the gap with supervised approaches. However, these methods are mostly used as a pre-training phase over IID data. In this talk, we focus on self-supervised methods for continual learning of visual feature representations. I introduce, Projected Functional Regularization (PFR) where a separate temporal projection network prevents forgetting of previously learned representations without jeopardizing plasticity. The main advantage of the new regularization method over existing methods is that it does not penalize the learning of new knowledge, and as a results can reach a better plasticity-stability trade-off.

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

Eli Verwimp (KU Leuven, BE)

License  Creative Commons BY 4.0 International license
© Eli Verwimp

During this presentation, I will present and discuss how continual learners learn and forget representations. We have observed two phenomena: knowledge accumulation, i.e. the improvement of a representation over time, and feature forgetting, i.e. the loss of task-specific representations. To better understand both phenomena, we introduced a new analysis technique called task exclusion comparison. If a model has seen a task and it has not forgotten all the task-specific features, then its representation for that task should be better than that of a model that was trained on similar tasks, but not that exact one. Our experiments show that most task-specific features are quickly forgotten, in contrast to what has been suggested in the past. Further, we demonstrate how some continual learning methods, like replay, and ideas from representation learning affect a continually learned representation.

4.23 Prediction Error-based Classification for Class-Incremental Learning

Michał Zajac (Jagiellonian University – Kraków, PL)

License  Creative Commons BY 4.0 International license
© Michał Zajac

Class-incremental learning (CIL) is a particularly challenging variant of continual learning, where the objective is to discriminate between all classes presented during the incremental learning process. Existing solutions often suffer from excessive forgetting and imbalance of the scores assigned to classes that have not been seen together during training. In our work, we introduce a novel approach, Prediction Error-based Classification (PEC), which differs from traditional discriminative and generative classification paradigms. PEC determines a class score by measuring the prediction error of a model trained to replicate the outputs of

a frozen random neural network on data from that class. Our empirical results show that PEC performs strongly and is on par or better than all considered rehearsal-free baselines, including those based on discriminative and generative classification, across multiple CIL benchmarks.

5 Working groups

5.1 Evaluation (Part 1)

Alexander Geppert (Hochschule für Angewandte Wissenschaften Fulda, DE)

License  Creative Commons BY 4.0 International license
© Alexander Geppert

Various aspects of evaluation procedures in CL were discussed, such as the proper and improper way of tuning hyper-parameters, the use of simple datasets like MNIST, and what useful evaluation measures for CL could be. It was commonly felt that new evaluation measures should also reflect what CL can contribute in terms for real-world applicability. For example, consistency, speed or compute-time/energy benefits achievable by CL when training large-scale models could be metrics to be used. We raised the issues of CL benefiting data privacy, and the application of CL to other modalities beyond vision. The general difficulty of evaluating models on large-sale data, as well as difficulties with the very concept of dataset were raised.

5.2 Evaluation (Part 2)

Andrea Cossu (University of Pisa, IT)

License  Creative Commons BY 4.0 International license
© Andrea Cossu

State of the art is useful provided that we study hard problems where it is “hard to cheat”. In particular, in continual learning the state of the art should be associated to a precisely specified setup. This is also due to the fact that, in continual learning, it is especially easy to cheat. Toy problems like MNIST can be useful, although some phenomena may only be visible at a certain scale. Surely, MNIST-like problems are useful as sanity checks before proceeding with more complex benchmarks. MNIST may still be relevant in extreme setups (e.g., online, replay-free, single-class learning). Continual learning is sometimes modality-specific. This is especially true for computer vision, where heavy use of augmentations restricts the applicability of continual learning strategies to other modalities.

5.3 Reproducibility

Alexander Geppert (Hochschule für Angewandte Wissenschaften Fulda, DE)

License © Creative Commons BY 4.0 International license
© Alexander Geppert

The session discussed how reproducibility in CL could be improved by, e.g., organizing a special track at a conference. The general goals of such an undertaking, as well as the target population of potential authors were discussed, as well as questions about what papers submitted to such a track could discuss. It was agreed that, despite a focus on reproducing results, papers should contain newness realized by, e.g., supplementary experiments, extended hyper-parameter searches or an application to other datasets. Finally, issues concerning the workflow of the submission and the review process were discussed.

5.4 Online Learning and Continual Learning

Andrea Cossu (University of Pisa, IT)

License © Creative Commons BY 4.0 International license
© Andrea Cossu

In online learning, there is no notion of generalization. Instead, algorithms are evaluated on regret. Online learning algorithms make a decision in each step (or datapoint). While online learning only cares about what happens at the current moment, continual learning cares about what happened during model lifetime. More, continual learning mainly works with neural networks. As such, it usually requires lots of data, making it difficult to relearn something. This requires to mitigate forgetting. Online learning, instead, does not have this requirement because relearning happens quickly. Can we come up with real data streams that have natural distribution? For example, data from Twitter can provide hundreds or thousands of datapoints per second, with gradual drift. Unfortunately, the Twitter API does not allow to extract this data anymore. One other difference is that, continual learning with replay always considers that distribution for a certain task remains stationary (the input-output mapping does not really change). It is still unclear whether or not continual learning and online learning can be integrated together.

5.5 Optimization in continual learning

Vincenzo Lomonaco (University of Pisa, IT)

License © Creative Commons BY 4.0 International license
© Vincenzo Lomonaco

The group discussed whether or not environments with piecewise iid data and environments with constant drift are really different for the optimization process. One possible solution would include the approximation of the static setting (only a patch) to make SGD work, for example by approximating a global static target function or by performing local optimization related to a moving target function. The usage of constraint optimization processes may help in maintaining important properties. A completely different solution would depart from the usual end-to-end training by leveraging separate objectives for different tasks and representations.

The group also discussed the role of bias for optimization in biology. It could be important to put similar bias into the model (like memory consolidation). In this sense, local learning is not similar to back-propagation which is global.

Participants

- Rahaf Aljundi
Toyota Motor Europe –
Zaventem, BE
- Shai Ben-David
University of Waterloo, CA
- Matthias Bethge
Universität Tübingen, DE
- Andrea Cossu
University of Pisa, IT
- Fabian Fumagalli
Universität Bielefeld, DE
- Joao Gama
INESC TEC – Porto, PT
- Alexander Geppert
Hochschule für Angewandte
Wissenschaften Fulda, DE
- Tyler Hayes
NAVER Labs Europe –
Meylan, FR
- Paul Hofman
LMU München, DE
- Eyke Hüllermeier
LMU München, DE
- Christopher Kanan
University of Rochester, US
- Tatsuya Konishi
KDDI – Saitama, JP
- Dhireesha Kudithipudi
University of Texas –
San Antonio, US
- Christoph H. Lampert
IST Austria –
Klosterneuburg, AT
- Bing Liu
University of Illinois –
Chicago, US
- Vincenzo Lomonaco
University of Pisa, IT
- Martin Mundt
TU Darmstadt, DE
- Razvan Pascanu
DeepMind – London, GB
- Adrian Popescu
CEA LIST – Nano-INNOV, FR
- James M. Rehg
Georgia Institute of Technology –
Atlanta, US
- Andreas Tolias
Baylor College of Medicine –
Houston, US
- Tinne Tuytelaars
KU Leuven, BE
- Gido van de Ven
KU Leuven, BE
- Joost van de Weijer
Computer Vision Center –
Barcelona, ES
- Eli Verwimp
KU Leuven, BE
- Michal Zajac
Jagiellonian University –
Kraków, PL

