Topic: Investigating Explainable AI Methods in Continual Learning

Artificial Intelligence (AI) research has made remarkable progress in recent years, excelling human performance on various tasks such as image recognition [1] or playing games like Go [2]. Despite these advances, modern AI algorithms rely on static datasets and fixed environments, contradicting the non-stationary nature of the world, in which tasks evolve over time and data arrives sequentially [3]. However, when these models are exposed to non-stationary data, they tend to forget previously learned knowledge while acquiring new knowledge. This phenomenon is known as catastrophic forgetting, where a model's performance on previous tasks degrades heavily as it learns new ones [4]. Such behavior poses a critical limitation for AI algorithms in real-world applications.

Continual Learning (CL), a subfield of machine learning, aims to address this issue by employing techniques such as replay-based, regularization-based, representation-based, and optimization-based methods to retain knowledge from previous tasks while learning new ones [5]. This not only allows AI algorithms to learn new tasks sequentially from a stream of data [6] but has also proven to be an effective method to prevent catastrophic forgetting [7]. While alleviating catastrophic forgetting is a critical goal, understanding how these models operate and make decisions is equally important to ensure their reliability in real-world applications. Especially in Deep Learning (DL), where models are often considered to be black-boxes due to their lack of interpretability and transparency, this underscores the need for explainability in AI models to ensure accountability and trustworthiness for critical applications [8]. To overcome this issue, Explainable Artificial Intelligence (XAI) offers methods such as Grad-CAM [9], LRP [10], or FovEx [11] to make the decision-making process of AI algorithms more transparent and understandable.

These XAI methods have been widely studied in static learning scenarios, however, only limited research exists on their application in dynamic, evolving settings such as continual learning scenarios [12]. Specifically, the relationship between performance degradation due to forgetting and its impact on the robustness and suitability of different XAI methods remains underexplored. Therefore, this thesis aims to investigate how XAI methods are affected by catastrophic forgetting in continual learning tasks, evaluating their robustness and suitability for enhancing the explainability of AI algorithms in CL scenarios.

By investigating XAI in CL, this thesis aims to address the following key objectives:

- Implement three CL scenarios, i.e., task-incremental, class-incremental, and domain-incremental learning, using the Avalanche library. These scenarios are tested on two backbone architectures: ResNet50 [13] and Vision Transformers (ViT) [14].
- Apply XAI methods such as Grad-CAM, LRP, and FovEx to analyze how model explanations evolve during high-accuracy phases and periods of catastrophic forgetting.
- Quantitative evaluation of explanation quality through metrics like Insertion [15], Deletion [15], % Drop [16], and % Increase [16].
- Quantitative analysis of the correlation between XAI metrics and CL performance metrics (e.g., accuracy deterioration, forgetting rate) to examine the relationship between explanation quality and model performance over time.
- Optional: Investigate and develop more efficient continual learning replay strategies by focusing on regions of the input where explanations exhibit significant changes over time.

This comprehensive analysis will provide valuable insights into how XAI methods are affected by catastrophic forgetting, advancing our understanding of how these methods can improve the robustness and transparency of continual learning systems.

Supervisors:	Dr. Dario Zanca, and Prof. Dr. Björn Eskofier
External advisors:	Dr. Matteo Tiezzi (IIT, Genova), and Prof. Stefano Melacci (University of Siena)
Student:	Oliver Dassinger
Start – End:	01.01.2025 – 30.06.2025

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.

[2] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, Jan 2016.

[3] Raia Hadsell, Dushyant Rao, Andrei Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24:1028–1040, 12 2020.

[4] Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks, 2015.

[5] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application, 2024.

[6] German I. Parisi and Vincenzo Lomonaco. *Online Continual Learning on Sequences*, pages 197–221. Springer International Publishing, 2020.

[7] Jianshu Zhang, Yankai Fu, Ziheng Peng, Dongyu Yao, and Kun He. Core: Mitigating catastrophic forgetting in continual learning through cognitive replay, 2024.

[8] Vanessa Buhrmester, David MÃ¼nch, and Michael Arens. Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction*, pages 966–989, 2021.

[9] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2019.

[10] Alexander Binder, Gregoire Montavon, Sebastian Bach, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers, 2016.

[11] Mahadev Prasad Panda, Matteo Tiezzi, Martina Vilas, Gemma Roig, Bjoern M. Eskofier, and Dario Zanca. Human-inspired explanations for vision transformers and convolutional neural networks, 2024.

[12] Andrea Cossu, Francesco Spinnato, Riccardo Guidotti, and Davide Bacciu. A protocol for continual explanation of shap, 2023.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[15] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018.

[16] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018.