Foundation Models are the cornerstone of most recent breakthroughs in Machine Learning. While achieving impressive capabilities [15], the prohibitive and sharp increases in training and inference costs [44] raise questions about whether this is the right path toward artificial intelligence. Indeed, these models follow specific scaling laws [19, 17], where performance not only improves with model size but also with the volume of training data [25, 31], raising concerns about the potential depletion of data resources [28] and significant implications for data privacy and copyright infringement [50, 36]. Developing such models requires a **centralization of resources and efforts** that only a few privileged teams can afford [41]. This not only grants significant influence over research and development but, more alarmingly, allows only a select few to shape policy decisions that affect end-users. In light of this, there is a pressing need for a change, transitioning towards a more **decentralized**, **collaborative**, and **efficient** paradigm.

In my research, I focus on creating models and algorithms to address these issues for a more open and accessible development of machine learning tools. To tackle these fundamental challenges, I establish my research agenda around three pillars:

1. **Decentralized Learning**, to enable collaboration across users and sharing knowledge while preserving privacy and distributing computational efforts.

2. **Continual Learning**, to enable continuous knowledge integration allowing models to keep improving without forgetting previously acquired skills, unless on purpose.

3. **Modular Learning**, for creating models that are more reusable, efficient, and maintainable.

The synergies between these three interconnected directions are discussed in detail below, as well as previous achievements and future directions.

1 Decentralized and Collaborative Learning

Federated Learning [11] (FL) offers a practical solution for collaborative model training while preserving user privacy. This decentralized technique allows individual clients to train models on their own data independently and share only model updates, circumventing the need to directly share sensitive information. By involving real-world users in the training process [14, 18], FL has the potential to learn from a wider range of data, reflecting real-world complexities. However, the privacy-preserving nature of FL presents challenges as local data generally reflects user preferences, habits, and geographical locations, potentially leading to biased local optimizations and hindering model convergence [20].

> My research addresses the crucial challenges of improving convergence speed and communication efficiency in decentralized settings particularly when faced with statistical heterogeneity across clients.

Statistical heterogeneity in federated learning could take the form of data imbalance [22, 34], domain [21, 24], and label [34, 35] shift. To address these challenges, I develop methods that approach them from an **optimization** and **generalization** perspective. Specifically, I draw connections between training difficulty and the geometry of the loss landscape [22], leverage global information to align local and global objectives [51], and exploit clients' similarity to mitigate interference and maximize knowledge transfer. This is achieved via clustering-based methods [21, 43, 30], normalizations [24], and hierarchical aggregation [43]. To foster research on federated learning, I also contributed to the release of **two benchmarks** [24, 35, 43] to characterize the effect of heterogeneity in real-world vision applications.

2 Continual and Incremental Learning

One of the hallmarks of intelligence is the ability to continually learn and improve by accruing knowledge. While machine learning algorithms have achieved remarkable advancements, they still cannot learn as humans [4], who are efficient, robust, and capable of learning from sequential experience. In real-world ML applications, several scenarios necessitate updating models, including adapting to distribution shifts, incorporating new features, and addressing unwanted behaviors. However, to date, there are no standard approaches for updating models while preserving their previously acquired knowledge [16]. As a result, the prevalent solution remains to re-train from scratch with the updated data – an inefficient practice that leads to time and resource wastage, contributing to increased energy consumption and carbon footprint [39].

> My research aims to surpass the limitations of current model development to enable Continual Learning (CL) and knowledge retention in the face of evolving distributions and requirements.

Drawing on my expertise in computer vision [10, 12], I delved into practical applications of continual learning for semantic segmentation. Recognizing the limitations of expensive dense annotations, even for incremental methods, together with my collaborators, we proposed [23] a novel distillation framework for Weakly Incremental Learning that enables segmentation models to learn from cheap and readily available image-level labels, paving the way for efficient continual learning in the real world.

More recently, I focused on a crucial weakness in existing CL approaches and their evaluation. I demonstrated [29] that the performance of most algorithms fluctuates significantly if the same data is presented in a different order, also referred to as data *schedule*. This sensitivity poses a significant issue for the deployment in real-world applications, where CL methods may be applied to novel data streams with unknown schedules. To address this challenge, I introduced the concept of *"schedule robustness"*, which quantifies the variance of performance across schedules and proposed a novel schedule-robust method with theoretical guarantees.

This line of research also led to a successful extension in federated learning to achieve **invariance to statistical heterogeneity** [34, 45]. Indeed, in a separate work, my research [22] has identified a critical challenge in FL where data heterogeneity can lead to over-specialized local models and consequent *catastrophic forgetting* when clients overwrite each other's knowledge during training - similar to the problem encountered in sequential learning [5]. Our approach [34, 45], not only prevents forgetting but also recovers the same solution that would be achieved in centralized training, while offering a significant convergence speed-up in real-world settings where data exhibit long-tail distributions and imbalances [18], further exacerbating forgetting.

3 Towards Modular, Collaborative and Decentralized Machine Learning

While the success of large-scale deep learning models hinged on the *"bigger is better"* approach – scaling model size and training data – this paradigm is rapidly reaching an inflection point. Beyond the prohibitive cost of training and maintaining gigantic models, this approach exposes and exacerbates inherent flaws in the current design philosophy of machine learning systems.

One of the most glaring contradictions lies in the development life cycle of these models which once deprecated are simply discarded in favor of new ones, generally trained from scratch. This unsustainable practice stems from the fact that models are currently built and trained as *generalist black-box monolithic systems* where functionalities and emerging capabilities are intertwined in their parameters and any attempt to change a specific aspect can have unpredictable and potentially disastrous consequences for the entire model's performance.

In stark contrast, a fundamental principle in software development is the organization of code into **modular components** [1]. This allows developers to import modules and seamlessly integrate new functionalities, leading to improved code reusability and maintainability. Similarly, biological systems provide compelling evidence for the benefits of modularity and functional specialization [2, 3], such as rapid adaptation to new environments and resilience to perturbations [6]. Despite these clear benefits, modular approaches are rarely applied in the development of machine learning models [42] presenting a significant opportunity for innovation.

My research envisions a future where deep learning models are built with **modular design** and **functional specialization** in mind, unlocking two key capabilities:

- **Asynchronous training and incremental updates**: functional specialization enables efficient, continuous improvement of individual modules, reducing interference and enhancing maintainability and interpretability.
- **Composability for cross-task generalization**: The modular design enables *post hoc* composition of different modules to adapt to new tasks and domains, promoting reusability and systematic generalization.

Realizing this vision presents a wealth of open research questions, raising exciting possibilities for future exploration. Below, I describe a few directions I intend to pursue.

Decomposing and Recombining Monolithic Models A major focus of my future research is developing techniques to decompose large, monolithic models into independent, task-specialized modules *after training*. This approach mirrors the **refactoring process** in software development, where a codebase is restructured for modularity and reusability. My goal is to create methods that enable this decomposition for deep neural networks, effectively disentangling knowledge representation from control flow within the model. While training large models remains resource-intensive, there is a growing abundance of open-source alternatives. My research aims to make these models even more accessible and reusable, empowering users to leverage their power without extensive computational resources.

I envision multiple benefits from modularizing large models. First, smaller, conditionally activated modules significantly **boost computational efficiency**. Additionally, recombining modules can improve cross-task generalization, **enabling more efficient transfer learning** to new tasks and domains. Finally, decomposing models into **semantically interpretable components** will offer crucial insights into the decision-making processes of complex neural networks, including a better understanding of how knowledge is stored and used to solve tasks.

I have begun exploring this direction by developing data-driven methods [46] that analyze network activation paths to identify and extract functional modules. This technique has already shown promise when applied to pruning models at initialization, effectively supporting efficient transfer learning from pre-trained models to downstream tasks. My future work focuses on scaling these methods and developing new techniques for large vision and language models, aiming to extract and recombine semantically meaningful modules post hoc. Additionally, I am actively investigating low-rank decomposition of pre-trained models [49] and ordered representations [37] where different dimensions have different degrees of importance and the rank can be selected layer-wise and adaptively based on the tasks and the available resources.

Theoretical and empirical understanding of model merging and knowledge transfer The promise of modular neural network design remains limited by the lack of a principled approach to combining knowledge from independently trained models. Current merging techniques are often ad-hoc, relying on heuristics rather than a theoretical understanding of when and how to achieve optimal knowledge transfer. My research seeks to establish a theoretical framework for effective model merging, along with practical guidelines for module composition in collaborative model development tools [38]. By investigating the interplay between architectures [26, 27, 32], optimization [22, 47], and task relationships [13], I aim to develop techniques that improve flexibility of model development while maximizing knowledge transfer when merging models.

My research seeks to bridge this gap, building upon insights from my work in continual and federated learning with heterogeneous data. Specifically, I have explored the role of optimization landscapes in model merging [22], finding flat minima to be conducive to better merging and reduced interference. Similarly, our work on path-aware pruning [46] could be instrumental in characterizing and reducing catastrophic interference across activation paths of different tasks.

An open question is how to transfer knowledge across models with different architectures and trained on different modalities. The interplay between knowledge distillation [9, 48] or other data-driven techniques [40, 33] and model merging is currently unexplored and could be a complementary approach to share knowledge in the context of decentralized and distributed training on partitioned datasets.

Communication-Efficient Large-Scale Decentralized Learning Training large foundational models demands centralized clusters of tightly interconnected accelerators. This involves substantial costs to build and maintain a single, massive computing infrastructure. Indeed, traditional distributed training relies on frequent synchronization across all devices creating a communication bottleneck as model size and dataset complexity increase.

In contrast, local optimizer methods [7, 8, 11] empower independent training processes on separate computing clusters or devices. Each worker trains on its local data partition, updating its model replica for multiple steps before periodically exchanging gradients for synchronization. This decentralized approach offers several benefits. By avoiding synchronization at every training step, local methods significantly reduce network traffic and associated latency. Additionally, training is more robust to individual device failures, as progress continues independently across workers. Finally, decentralization and delayed communication allow the integration of heterogeneous devices with varying speeds and capabilities.

Despite their advantages, the potential of local optimizers for decentralized training of large-scale vision and language models remains underexplored. My research has focused on developing effective local optimization methods for federated learning to accelerate convergence in heterogeneous data distributions [45, 51, 30] and improve communication efficiency [51] reducing the need for frequent synchronization.

I am currently investigating the extension of these algorithms for training large language and multimodal models and effectively leveraging decentralized compute resources. Furthermore, integrating modular architectures within this framework allows for independent training of task-specific and parameter-efficient modules [26, 27], effectively improving scalability. Future exploration could investigate hierarchical [21, 43] or peer-to-peer topologies [30] to further optimize communication efficiency based on location and data similarity.

Research impact: Modularity has the potential to democratize the development of machine learning models, shifting from the current centralized paradigm towards a collaborative ecosystem. By enabling independent training and the combination of specialized modules, research institutions, and smaller teams could build powerful models by sharing their knowledge and resources. This would also pave the way for more flexible and efficient models, with modularity facilitating updates, customization, and reduced training costs. Finally, the ability to analyze and replace specific modules would dramatically improve model interpretability and trustworthiness, fostering rigorous evaluation and accountability.

References

[1] DL Parnas. “On the criteria to be used in decomposing systems into modules”. In: *Communications of the ACM* 15.12 (1972), pp. 1053–1058.

[2] JA Fodor. *The modularity of mind*. MIT press, 1983.

[3] DH Ballard. “Cortical connections and parallel processing: Structure and function”. In: *Behavioral and brain sciences* 9.1 (1986), pp. 67–90.

[4] M McCloskey and NJ Cohen. “Catastrophic interference in connectionist networks: The sequential learning problem”. In: *Psychology of learning and motivation*. Vol. 24. Elsevier, 1989, pp. 109–165.

[5] RM French. “Catastrophic forgetting in connectionist networks”. In: *Trends in cognitive sciences* 3.4 (1999), pp. 128–135.

[6] GP Wagner, J Mezey, and R Calabretta. “Natural Selection and the Origin of Modules”. In: *Modularity: Understanding the Development and Evolution of Natural Complex Systems*. The MIT Press, May 2005. ISBN: 9780262269698. DOI: 10.7551/mitpress/4734.003.0009. eprint: https://direct.mit.edu/book/chapter-pdf/2302813/c001000_9780262269698.pdf. URL: https://doi.org/10.7551/mitpress/4734.003.0009.

[7] R Mcdonald, M Mohri, N Silberman, D Walker, and G Mann. “Efficient large-scale distributed training of conditional maximum entropy models”. In: *Advances in neural information processing systems* 22 (2009).

[8] M Zinkevich, M Weimer, L Li, and A Smola. “Parallelized stochastic gradient descent”. In: *Advances in neural information processing systems* 23 (2010).

[9] G Hinton, O Vinyals, and J Dean. “Distilling the Knowledge in a Neural Network”. In: *NIPS Deep Learning and Representation Learning Workshop*. 2015. URL: http://arxiv.org/abs/1503.02531.

[10] F Visin, **M Ciccone**, A Romero, K Kastner, C Kyunghyun, Y Bengio, M Matteucci, and A Courville. “ReSeg : A Recurrent Neural Network-based Model for Semantic Segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2016.

[11] B McMahan, E Moore, D Ramage, S Hampson, and BA y Arcas. “Communication-efficient learning of deep networks from decentralized data”. In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

[12] F Lattari*, **M Ciccone***, M Matteucci, J Masci, and F Visin. “ReConvNet: Video Object Segmentation with Spatio-Temporal Features Modulation”. In: *The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops* (2018).

[13] A Achille, M Lam, R Tewari, A Ravichandran, S Maji, CC Fowlkes, S Soatto, and P Perona. “Task2vec: Task embedding for meta-learning”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6430–6439.

[14] K Bonawitz, H Eichner, W Grieskamp, D Huba, A Ingerman, V Ivanov, C Kiddon, J Konečnỳ, S Mazzocchi, B McMahan, et al. “Towards federated learning at scale: System design”. In: *Proceedings of machine learning and systems* 1 (2019), pp. 374–388.

[15] T Brown, B Mann, N Ryder, M Subbiah, JD Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[16] R Hadsell, D Rao, AA Rusu, and R Pascanu. “Embracing change: Continual learning in deep neural networks”. In: *Trends in cognitive sciences* 24.12 (2020), pp. 1028–1040.

[17] T Henighan, J Kaplan, M Katz, M Chen, C Hesse, J Jackson, H Jun, TB Brown, P Dhariwal, S Gray, et al. “Scaling laws for autoregressive generative modeling”. In: *arXiv preprint arXiv:2010.14701* (2020).

[18] TMH Hsu, H Qi, and M Brown. “Federated visual classification with real-world data distribution”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer. 2020, pp. 76–92.

[19] J Kaplan, S McCandlish, T Henighan, TB Brown, B Chess, R Child, S Gray, A Radford, J Wu, and D Amodei. “Scaling laws for neural language models”. In: *arXiv preprint arXiv:2001.08361* (2020).

[20] SP Karimireddy, S Kale, M Mohri, S Reddi, S Stich, and AT Suresh. “Scaffold: Stochastic controlled averaging for federated learning”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5132–5143.

[21] D Caldarola, M Mancini, F Galasso, **M Ciccone**, E Rodola, and B Caputo. “Cluster-Driven Graph Federated Learning Over Multiple Domains”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2021, pp. 2749–2758.

[22] D Caldarola*, B Caputo, and **M Ciccone***. “Improving Generalization in Federated Learning by Seeking Flat Minima”. In: *European Conference on Computer Vision*. 2022.

[23] F Cermelli*, D Fontanel*, A Tavera*, **M Ciccone**, and B Caputo. “Incremental Learning in Semantic Segmentation from Image Labels”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022.

[24] L Fantauzzo, E Fani, D Caldarola, A Tavera, F Cermelli, **M Ciccone**, and B Caputo. “FedDrive: Generalizing Federated Learning to Semantic Segmentation in Autonomous Driving”. In: *Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2022.

[25] J Hoffmann, S Borgeaud, A Mensch, E Buchatskaya, T Cai, E Rutherford, D de las Casas, LA Hendricks, J Welbl, A Clark, T Hennigan, E Noland, K Millican, G van den Driessche, B Damoc, A Guy, S Osindero, K Simonyan, E Elsen, O Vinyals, JW Rae, and L Sifre. “An empirical analysis of compute-optimal large language model training”. In: *Advances in Neural Information Processing Systems*. Ed. by AH Oh, A Agarwal, D Belgrave, and K Cho. 2022. URL: https://openreview.net/forum?id=iBBcRUlOAPR.

[26] EJ Hu, yelong shen, P Wallis, Z Allen-Zhu, Y Li, S Wang, L Wang, and W Chen. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[27] H Liu, D Tam, M Muqeeth, J Mohta, T Huang, M Bansal, and CA Raffel. “Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 1950–1965.

[28] P Villalobos, J Sevilla, L Heim, T Besiroglu, M Hobbhahn, and A Ho. “Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning”. In: *arXiv preprint arXiv:2211.04325* (2022).

[29] R Wang*, **M Ciccone***, G Luise, M Pontil, A Yapp, and C Ciliberto. “Schedule-Robust Online Continual Learning”. In: *submitted to IEEE TPAMI* (2022).

[30] R Zaccone, A Rizzardi, D Caldarola, **M Ciccone**, and B Caputo. “Speeding up Heterogeneous Federated Learning with Sequentially Trained Superclients”. In: *26th International Conference on Pattern Recognition (ICPR)*. IEEE. 2022.

[31] I Alabdulmohsin, X Zhai, A Kolesnikov, and L Beyer. “Getting ViT in Shape: Scaling Laws for Compute-Optimal Model Design”. In: *Advances in Neural Information Processing Systems*. 2023. URL: https://openreview.net/forum?id=en4LGxpd9E.

[32] L Caccia, E Ponti, Z Su, M Pereira, NL Roux, and A Sordoni. “Multi-Head Adapter Routing for Cross-Task Generalization”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023. URL: https://openreview.net/forum?id=qcQhBli5Ho.

[33] D Crisostomi, I Cannistraci, L Moschella, P Barbiero, **M Ciccone**, P Lio, and E Rodolà. “From Charts to Atlas: Merging Latent Spaces into One”. In: *NeurIPS Workshop on Symmetry and Geometry in Neural Representations*. 2023.

[34] E Fanì, R Camoriano, B Caputo, and **M Ciccone**. “Fed3R: Recursive Ridge Regression for Federated Learning with strong pre-trained models”. In: *NeurIPS Workshop on Federated Learning in the Age of Foundation Models* (2023).

[35] E Fanì, M Ciccone, and B Caputo. “FedDrive v2: an Analysis of the Impact of Label Skewness in Federated Semantic Segmentation for Autonomous Driving”. In: 2023.

[36] MM Grynbaum and TNYT Ryan Mac. *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work*. 2023. URL: https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html (visited on 02/22/2024).

[37] S Horvath, S Laskaridis, S Rajput, and H Wang. “Maestro: Uncovering Low-Rank Structures via Trainable Decomposition”. In: *arXiv preprint arXiv:2308.14929* (2023).

[38] N Kandpal, B Lester, M Muqeeth, A Mascarenhas, M Evans, V Baskaran, T Huang, H Liu, and C Raffel. “Git-Theta: A Git Extension for Collaborative Development of Machine Learning Models”. In: 2023. URL: https://arxiv.org/abs/2306.04529.

[39] AS Luccioni and A Hernandez-Garcia. "Counting carbon: A survey of factors influencing the emissions of machine learning". In: *arXiv preprint arXiv:2302.08476* (2023).

[40] L Moschella, V Maiorca, M Fumero, A Norelli, F Locatello, and E Rodolà. "Relative representations enable zero-shot latent space communication". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=SrC-nwieGJ.

[41] D Patel and D Nishball. *Google Gemini Eats The World – Gemini Smashes GPT-4 By 5X, The GPU-Poors*. 2023. URL: https://www.semianalysis.com/p/google-gemini-eats-the-world-gemini (visited on 02/22/2024).

[42] J Pfeiffer, S Ruder, I Vulić, and E Ponti. "Modular Deep Learning". In: *Transactions on Machine Learning Research* (2023). Survey Certification. ISSN: 2835-8856. URL: https://openreview.net/forum?id=z9EkXfvxta.

[43] D Shenaj*, E Fanì*, M Toldo, D Caldarola, A Tavera, U Michieli†, **M Ciccone**†, P Zanuttigh†, and B Caputo†. "Learning Across Domains and Devices: Style-Driven Source-Free Domain Adaptation in Clustered Federated Learning". In: *2023 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2023.

[44] Wired. *OpenAI's CEO Says the Age of Giant AI Models Is Already Over*. 2023. URL: https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/ (visited on 02/22/2024).

[45] E Fanì, R Camoriano, B Caputo, and **M Ciccone**. "Accelerating Heterogeneous Federated Learning with Closed-form Classifiers". In: *submitted to ICML* (2024).

[46] L Iurada, **M Ciccone**, and T Tommasi. "Finding Lottery Tickets in Vision Models via Data-driven Spectral Foresight Pruning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).

[47] G Ortiz-Jimenez, A Favero, and P Frossard. "Task arithmetic in the tangent space: Improved editing of pre-trained models". In: *Advances in Neural Information Processing Systems* 36 (2024).

[48] K Roth, L Thede, AS Koepke, O Vinyals, OJ Henaff, and Z Akata. "Fantastic Gains and Where to Find Them: On the Existence and Prospect of General Knowledge Transfer between Any Pretrained Model". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=m50eKHCttz.

[49] P Sharma, JT Ash, and D Misra. "The Truth Is In There: Improving Reasoning with Layer-Selective Rank Reduction". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=ozX92bu8VA.

[50] TNYT Stuart A. Thompson. *We Asked A.I. to Create the Joker. It Generated a Copyrighted Image*. 2024. URL: https://www.nytimes.com/interactive/2024/01/25/business/ai-image-generators-openai-microsoft-midjourney-copyright.html (visited on 02/22/2024).

[51] R Zaccone, C Masone, and **M Ciccone**. "Communication-Efficient Heterogeneous Federated Learning with Generalized Heavy-Ball Momentum". In: *submitted to ICML* (2024).