FORGING ADAPTABILITY: SYSTEMATIC GENERALIZATION AND CONTINUAL LEARNING IN BIOMIMETIC AI

GIACOMO CAPPELLETTO[1]

CONTENTS

[1] *Computer Engineering, Boston University, Boston MA*

1 OVERVIEW AND SCOPE

Despite the power of Artificial Neural Networks (ANNs), designing effective architectures remains a significant hurdle, often involving painstaking trial-and-error or computationally costly neural architecture searches (NAS). This difficulty highlights a gap between current AI and the fluid adaptability of biological human intelligence. A compelling avenue for progress lies in biomimetic ANNs, drawing inspiration from the structural and functional principles of the mammalian brain. Specifically, incorporating the mammalian brain's modular architecture and intricate hierarchical feedback mechanisms into ANNs shows the possibility of achieving unforeseen efficiency and neural plasticity. Researchers hypothesize that these neurally inspired designs could significantly enhance capabilities where ANNs currently lag: fostering greater creativity, improving systematic generalisation capabilities, enabling true adaptive plasticity in response to new information, and achieving continuous learning without catastrophic forgetting.

Yet, the brain is a product of physical biological constraints vastly different from artificial hardware. Directly translating its complex, evolved and highly energetically efficient mechanisms presents inherent challenges and limitations. This raises a critical question for the future of AI, which this paper aims to investigate. To what extent can incorporating the modular architecture and hierarchical feedback mechanisms of the mammalian brain into artificial neural networks enhance their adaptive plasticity, and continuous learning, and what intrinsic limitations of this biomimetic approach prevents them from achieving human-like intelligence? By analyzing recent research, we will explore both the potential advancements offered by neuro-inspiration and the fundamental obstacles impeding the path toward artificial general intelligence.

2 HUMAN-LIKE SYSTEMATIC COMPOSITIONALITY

2.1 The Human Benchmark

A cornerstone of human language and thought, arguably setting it apart from other forms of intelligence, is its systematic compositionality. This refers to the remarkable, almost algebraic capacity to understand and generate entirely novel combinations—sentences never before heard, thoughts never explicitly conceived—from a finite set of previously known components, such as words or concepts (Fodor and Pylyshyn 1988; Lake and Baroni 2023). For instance, once a person understands the constituents 'walk' and 'backwards', they can immediately comprehend or produce 'walk backwards', even without prior specific exposure to that new concept.

This generative power seems to be a fundamental trait of human cognitive flexibility. It is precisely this capability that became the focal point of a foundational critique against early artificial neural networks (ANNs). In their seminal 1988 paper, Jerry Fodor and Zenon Pylyshyn argued forcefully that connectionist architectures, the precursors to modern ANNs, lacked the necessary internal structure to support such capabilities. They contended that the distributed, associative nature of these models precluded the kind of rule-based, constituent-sensitive processing inherent in human cognition, thereby rendering ANNs fundamentally unsuitable as models of the mind (Fodor and Pylyshyn 1988).

This "systematicity challenge" has echoed through decades of AI research; while ANNs have advanced exponentially in capability, particularly with the advent of deep learning and architectures like the transformer (Vaswani et al. 2017), demonstrating true, human-like systematic generalization has remained an elusive benchmark, a persistent hurdle suggesting an intrinsic limitation in bridging the gap towards artificial general intelligence (Marcus 2001; Lake and Baroni 2018). Addressing this very challenge, and investigating the extent to which ANNs can embody this crucial aspect of human intelligence, is central to understanding both their potential and their current constraints.

2.2 The MLC Approach

Recent work by Brenden Lake and Marco Baroni (2023) directly confronts this enduring systematicity challenge, providing evidence that ANNs, when appropriately optimized, can indeed achieve human-like, and in some cases even superior, systematic generalization. Their research pivots away from solely modifying the base architecture of the network towards refining the learning process itself. They introduce an approach termed Meta-Learning for Compositionality (MLC), an optimization procedure specifically designed to cultivate compositional skills within neural networks.

MLC does not rely on incorporating explicit symbolic machinery or handcrafting internal representations or inductive biases, which might be seen as circumventing the original spirit of Fodor and Pylyshyn's critique of purely connectionist systems. Instead, MLC operates on standard, widely-used ANN architectures, specifically the transformer, which has proven highly effective in domains like natural language processing (Lake and Baroni 2023). The core idea behind MLC is to train the network not just on a static dataset, but through a dynamic stream of related, few-shot compositional tasks, or "episodes." Each episode presents the network with a small set of "study" examples (input/output pairs demonstrating a specific compositional grammar) and then requires it to generalize to a novel "query" instruction that involves compositional use of elements seen only in isolation during study.

By optimizing the network's parameters across these dynamically changing episodes, MLC encourages the network to learn how to learn compositional rules and apply them systematically, rather than merely memorizing specific input-output patterns (Lake and Baroni 2023; Hospedales et al. 2022).

2.3 Experimental Validation

To rigorously evaluate MLC's capabilities against human performance, Lake and Baroni (2023) employed a carefully designed few-shot instruction-learning paradigm. Human participants and MLC-optimized transformers were tasked with interpreting instructions in a pseudo-language and generating corresponding sequences of abstract outputs. The instructions required understanding primitive word meanings and functional words that combined these primitives in specific ways such as repeating an output, alternating between two outputs, reversing sequences. Participants and models had to infer these meanings and rules from only a handful of examples and then apply them to 10 novel query instructions, some requiring longer output sequences or more complex compositions than seen during study. The results were promising for the MLC approach: human participants demonstrated strong systematic generalization, correctly producing the algebraically expected output in 80.7

2.4 Implications

These findings significantly challenge the long-held assertion that ANNs are inherently non-systematic. Lake and Baroni's (2023) work demonstrates that a standard transformer architecture, when optimized via their MLC strategy, can achieve, and even exceed, human levels of systematic generalization on few-shot instruction tasks. By focusing the learning process on acquiring compositional skills, MLC enables networks to infer and apply rules from sparse data, directly addressing the core limitation identified by Fodor and Pylyshyn (1988). Notably, the MLC model also replicates human-like error patterns and inductive biases, suggesting its potential as a nuanced model of human compositional behaviour beyond rigid symbolic approaches (Lake and Baroni 2023).

This research establishes that enhancing ANNs with sophisticated, neurally plausible learning strategies is a viable path towards achieving critical facets of human intelligence like systematicity. However, while systematic generalization addresses the ability to handle novel combinations of known elements, another vital dimension of cognitive flexibility is adaptive plasticity—the capacity to learn entirely new information over time without catastrophically disrupting prior knowledge. How ANNs fare in this domain of continual learning requires further discussion.

3 CONTINUAL LEARNING AND ADAPTIVE PLASTICITY

3.1 The Continual Learning Challenge

While the capacity for systematic generalization addresses how knowledge components are combined, another equally fundamental aspect of human cognitive adaptability is the ability to learn new information and skills sequentially over a lifetime while maintaining prior knowledge. Humans exhibit remarkable continual learning, integrating new experiences while largely preserving existing competences (Flesch et al. 2023). Standard ANNs, however, face a significant obstacle in this regard, known as catastrophic forgetting. When trained sequentially on distinct tasks using typical gradient descent methods, the network's weights are adjusted to optimize performance on the current task, often drastically altering representations crucial for previously learned tasks. This leads to a rapid and ultimately complete loss of performance on earlier tasks, a stark contrast to the more graceful learning trajectory observed in biological systems (French 1999; Kirkpatrick et al. 2017; Flesch et al. 2023).

This limitation completely prevents the applicability of ANNs in dynamic environments where information arrives sequentially, highlighting a need for mechanisms that display adaptive plasticity ability and enable lifelong learning. Exploring whether principles derived from the mammalian brain's own solutions for managing sequential information can endow ANNs with similar capabilities is thus central to this paper.

3.2 PFC-Inspired Mechanisms

Addressing this challenge, Timo Flesch and colleagues (2023) developed a computational model explicitly designed to capture human-like continual learning dynamics by incorporating mechanisms inspired by cognitive control processes within the primate prefrontal cortex (PFC). Their work moves beyond standard ANN training paradigms by introducing two specific, biologically plausible algorithmic motifs.

The first involves "sluggish" task units, designed to mimic the intrinsic temporal integration constants found in biological neural circuits. These units carry over contextual information from previous trials via an exponential moving average, reflecting the influence of recent history on current processing, a common feature of human decision-making that often leads to switch costs when tasks change rapidly (Flesch et al. 2023; Monsell 2003). The second, and perhaps more crucial, mechanism is a form of Hebbian context gating. Following each standard supervised learning update, an additional Hebbian learning step strengthens the connections between input units signaling the current task context and those hidden units found to be encoding task-relevant information. This process, inspired by theories of PFC function where context signals gate or modulate processing pathways (Miller and Cohen 2001; Rougier et al. 2005), effectively learns to orthogonalize the representations of different tasks within the hidden layer.

3.3 Experimental Validation

The integration of these two mechanisms created a network capable of overcoming catastrophic forgetting and replicating key human behavioral patterns in multi-task learning scenarios. When trained sequentially on distinct categorization tasks, the model successfully learned the second task without overwriting the first –a feat which is virtually unachievable with regular stochastic gradient descent methods– demonstrating robust continual learning where standard ANNs typically fail (Flesch et al. 2023). Furthermore, the model also captured the perhaps counterintuitive human tendency to perform less accurately when tasks are randomly switched during training compared to when they are blocked. The "sluggish" units introduce interference or switch costs in the interleaved condition, mirroring human behavioral data, whereas standard ANNs, lacking this temporal dependence and preferring identically distributed data, perform best under interleaving (Flesch et al. 2023).

This success demonstrates that incorporating specific principles of brain architecture and control – namely, temporal integration dynamics and a learned, mammalian PFC-inspired gating mechanism that functions as a form of hierarchical feedback controlling task representations – can significantly enhance the adaptive plasticity of ANNs. It allows them not only to learn sequentially but also to replicate the nuanced performance characteristics of human continual learning.

3.4 Implications

This work, alongside the findings on systematic generalization, suggests that targeted biomimicry, focusing on specific computational principles observed in the brain, can indeed enhance ANN capabilities in critical areas like adaptive plasticity and continuous learning. However, these successes rely on introducing specific modifications –meta-learning strategies or brain-inspired gating mechanisms– to the standard deep learning framework. This raises a crucial question: what are the inherent properties of standard deep learning models themselves regarding plasticity, especially during prolonged exposure to information? Do they possess innate adaptability, or do they suffer from intrinsic limitations that necessitate such specialized, biologically inspired interventions?

4 PLASTICITY LOSS IN STANDARD DEEP LEARNING

4.1 The Problem of Plasticity Loss

While the previous examples demonstrate that carefully designed neurally inspired modifications can create artificial neural networks with specific human-like learning capabilities, they implicitly highlight a potential weakness in the foundational methods themselves. Does standard deep learning, even in its sophisticated modern forms, possess inherent, sustained adaptability, or is its plasticity fundamentally fragile over time? Recent extensive research by Shibhansh Dohare and colleagues (2024) provides a compelling, albeit sobering, answer: standard deep-learning methods, across a wide range of architectures and tasks, systematically and progressively lose plasticity during extended training, particularly in continual learning scenarios. Their work demonstrates that this is not an isolated issue but a pervasive phenomenon, occurring whether using feed-forward convolutional networks on large-scale image classification (Continual ImageNet) or advanced residual networks on class-incremental learning (CIFAR-100), and even extending into the domain of reinforcement learning (simulated ant locomotion). Over time, these networks, trained with conventional backpropagation and related optimization algorithms, become increasingly less effective at learning new information, eventually performing no better, or even worse, than much simpler linear models (Dohare et al. 2024).

4.2 Mechanisms of Decline

Dohare et al. (2024) meticulously document several key factors that correlate with, and likely contribute to, this degradation of learning ability. A primary issue is the emergence of dormant or "dead" units within the network. As training progresses, a significant fraction of neurons cease to activate meaningfully across the training input distribution (like ReLU layers always outputting zero)(Lu et al. 2020) or become saturated in sigmoidal (layers with logistic regression characteristics) layers. These inactive units effectively reduce the network's functional capacity, efficiency, and lose their ability to adapt their incoming weights via gradient descent, as gradients cannot flow back through them 'dead' neurons (Dohare et al. 2024).

This implies that the diversity of the network's internal representations diminishes. This is quantified by a drop in the "effective rank" of the hidden layers, indicating that fewer independent dimensions are contributing to the network's transformations. The network settles into lower-dimensional solutions optimized for recent data, making it harder to find new representations suitable for novel information. Lastly, the magnitude of the network's weights often shows problematic trends; without intervention, weights can grow steadily, potentially leading to unoptimized landscapes and slower learning, while regularization aimed at controlling weight size can sometimes overly constrain the network, preventing it from committing to high quality regressions (Dohare et al. 2024).

4.3 Biological Contrast and Required Interventions

This observed loss of plasticity is in contrast with the enduring adaptability characteristic of biological learning systems, which generally maintain the capacity to acquire new knowledge throughout their lifespan. The findings of Dohare et al. (2024) suggest that the decline in ANN plasticity is not merely a consequence of specific parameter choices or architectures but represents a more fundamental limitation tied to the standard gradient-descent based learning paradigm when applied continuously. The very process of optimizing for current data gradually erodes the beneficial properties of the initial randomly generated weight distribution, leading the network into states from which learning new regressions becomes more and more unlikely.

This perspective gives further insight on the successes achieved by methods like MLC (Lake and Baroni 2023) and Hebbian gating (Flesch et al. 2023). They are not simply enhancements but are, in effect, necessary interventions (or rather emulation of mammalian brain architecture) that actively counteract this intrinsic tendency towards plasticity loss. Indeed, Dohare et al.'s own proposed solution, "continual backpropagation," explicitly addresses this by continually injecting variability back into the network (reinitializing less-used units), demonstrating that maintaining plasticity requires mechanisms that go beyond standard gradient descent (Dohare et al. 2024).

4.4 Implications for Standard Deep Learning

Therefore, while ANNs draw initial inspiration from the brain, their standard learning dynamics exhibit fragility in maintaining long-term adaptability. This crucial limitation underscores why specialized biomimetic approaches are often necessary to achieve robust continual learning or other facets of flexible intelligence. It establishes a clear boundary on the capabilities of standard deep learning. Having acknowledged both the potential enhancements offered by specific brain-inspired strategies and this significant intrinsic limitation of standard models, we can now consider the broader picture: how substantial is the overall gap between current brain-inspired AI and the multifaceted intelligence of the mammalian brain, and what fundamental obstacles remain?

5 CONCLUSION

In conclusion, while neuroscience serves as a rich source of inspiration rather than a strict benchmark for artificial intelligence, incorporating specific brain-like principles demonstrably enhances ANN capabilities. Strategies like meta-learning refined for compositionality have yielded human-level systematic generalization, and biologically plausible gating mechanisms inspired by the PFC significantly improve continual learning by mitigating catastrophic forgetting.

Nevertheless, these targeted advancements often highlight or counteract fundamental weaknesses, such as the progressive loss of plasticity observed in standard deep learning models during prolonged training. Significant hurdles also remain in translating the intricate, evolved, and energetically efficient mechanisms of the biological brain, constrained by vastly different physical realities than artificial hardware. Consequently, despite strides made through biomimicry, substantial gaps persist concerning energy efficiency, true understanding, and creativity, indicating that achieving adaptable, human-like general intelligence requires overcoming both the complexities of neurobiology and the inherent limitations of current AI paradigms.

BIBLIOGRAPHY

A. Fodor, J. and Z. W. Pylyshyn (1988). Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition 28*(1-2).

Dohare, S., J. F. Hernandez-Garcia, Q. Lan, P. Rahman, A. R. Mahmood, and R. S. Sutton (2024, 8). Loss of plasticity in deep continual learning. *Nature 632*(8026), 768–774.

Faramarzi, F., F. Azad, M. Amiri, and B. Linares-Barranco (2019, 10). A Neuromorphic Digital Circuit for Neuronal Information Encoding Using Astrocytic Calcium Oscillations. *Frontiers in Neuroscience 13*.

Flesch, T., D. G. Nagy, A. Saxe, and C. Summerfield (2023, 1). Modelling Continual Learning in Humans With Hebbian Context Gating and Exponentially Decaying Task Signals. *PLoS Computational Biology 19*(1), e1010808.

French, R. (1999, 4). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences 3*(4), 128–135.

Hospedales, T. M., A. Antoniou, P. Micaelli, and A. J. Storkey (2021, 1). Meta-Learning in Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1.

Kirkpatrick, J., R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell (2017, 3). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences 114*(13), 3521–3526.

Lake, B. M. and M. Baroni (2018, 7). Generalization Without Systematicity: On the Compositional Skills of Sequence-to-sequence Recurrent Networks. *International Conference on Machine Learning*, 4487–4499.

Lake, B. M. and M. Baroni (2023, 10). Human-like Systematic Generalization Through a Meta-learning Neural Network. *Nature 623*(7985), 115–121.

Lu, L., Y. S. Y. Shin, Y. S. Y. Su, and G. E. K. G. E. Karniadakis (2020, 1). Dying ReLU and Initialization: Theory and Numerical Examples. *Communications in Computational Physics 28*(5), 1671–1706.

Marcus, G. F. (2001, 1). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*.

Miller, E. K. and J. D. Cohen (2001, 3). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience 24*(1), 167–202.

Monsell, S. (2003, 3). Task switching. *Trends in Cognitive Sciences 7*(3), 134–140.

Rougier, N. P., D. C. Noelle, T. S. Braver, J. D. Cohen, and R. C. O'Reilly (2005, 5). Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences 102*(20), 7338–7343.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017, 6). Attention Is All You Need. *Google Brain Research 30*, 5998–6008.