Nick Alonso

Machine Learning Researcher

nalonso2@uci.edu | www.neuralnetnick.com | https://github.com/nalonso2 | @Nick__Alonso

Research Interests

I am interested in improving memory, learning, and reasoning in current AI systems, especially in the context of online-continual learning. Previous research projects include developing bio-inspired learning algorithms for neural networks, neural network based memory models, and retrieval augmented language models designed for long-context tasks. My current research focuses on developing novel sequence mixing layers in language models to improve long-context processing.

Education

University of California, Irvine *Irvine, CA*
PhD Cognitive Science *2023*
- Advisors: Jeff Krichmar and Emre Neftci
- Dissertation: 'An Energy-based Approach to Learning and Memory in Artificial Neural Networks'
- GPA: 4.0

Georgia State University *Atlanta, GA*
MA Neurophilosophy *2019*

University of Michigan, Ann Arbor *Ann Arbor, MI*
BS Cognitive Science (focus Computation), BS Philosophy *2017*

Selected Professional Experience

Machine Learning Researcher - Zyphra (2023-now)

- Designed and implemented a neural-symbolic, LLM based system for tabular database retrieval that utilized a custom DSL library of retrieval functions and prompt-based agent for query rewriting and function calling.
- Designed and implemented a novel graph-based retrieval algorithm for performing long-context tasks based on the PageRank algorithm, which achieved SOTA results on multiple long-context reasoning benchmarks.
- I am currently developing novel sequence mixing layers for LLMs that aim to be both efficient and effective at long-context processing. My latest project Online Vector Quantized Attention is an example of this effort.

Graduate Student Researcher - UC, Irvine (2019-2023)

As a PhD student, I worked on several projects related to bio-inspired energy based models, including predictive coding models, with results published in journals such as Neurips and Nature Communications.

- For my first major project, I developed a novel theoretical interpretation of the learning algorithm utilized by predictive coding models.
- A subsequent project utilized this theoretical interpretation to develop a novel optimizer for predictive coding models.
- Finally, I lead a project that developed a novel kind of Hopfield network that utilized sparse coding and hierarchical vector quantization to achieve SOTA results on auto-associative memory tasks in online-continual

settings.

Engineering Intern - Cognitive and Emerging Computing, Sandia National Labs (Summer 2022)

As an intern I helped develop a novel mathematical connection between STDP and backpropagation through time, and used this interpretation to tweak the standard STDP model.

Skills

Python - Pytorch - Pytorch DistributedDataParall - Transformers - FAISS - NetworkX - SciPy - Latex

Papers

Alonso, N., Figliolia, T., & Millidge, B. (2026). Online Vector Quantized Attention arXiv preprint arXiv:2602.03922.

Figliolia, T., **Alonso, N.**, Iyer, R., Anthony, Q., & Millidge, B. (2025). Compressed Convolutional Attention: Efficient Attention in a Compressed Latent Space. arXiv preprint arXiv:2510.04476.

Alonso, N., & Millidge, B. (2024). Mixture-of-PageRanks: Replacing Long-Context with Real-Time, Sparse GraphRAG. arXiv preprint arXiv:2412.06078.

Alonso, N., Figliolia, T., Ndirango, A., & Millidge, B. (2024). Toward Conversational Agents with Context and Time Sensitive Long-term Memory. arXiv preprint arXiv:2406.00057.

Alonso, N., & Krichmar, J. L. (2024). A sparse quantized hopfield network for online-continual memory. Nature Communications, 15(1), 3722.

Alonso, N., Krichmar, J., & Neftci, E. (2024). Understanding and improving optimization in predictive coding networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 10, pp. 10812-10820). (Selected for oral presentation)

Alonso, N., Millidge, B., Krichmar, J., & Neftci, E. O. (2022). A Theoretical Framework for Inference Learning. Advances in Neural Information Processing Systems, 35, 37335-37348.

Wang, F., **Alonso, N.**, & Teeter, C. (2022). Combining Spike Time Dependent Plasticity (STDP) and Backpropagation (BP) for Robust and Data Efficient Spiking Neural Networks (SNN) (No. SAND2022-16962). Sandia National Lab.(SNL-NM), Albuquerque, NM (United States). (Technical Report)

Alonso, N., & Neftci, E. (2021). Tightening the Biological Constraints on Gradient-Based Predictive Coding. In International Conference on Neuromorphic Systems 2021 (pp. 1-9).

Awards, Fellowships, & Grants

2022	**Neurips Scholar Award**, Neurips Conference, Travel Funding
2017-2019	**Brain and Behavior Fellowship**, Georgia State Neuroscience Instit., Merit-based, full funding for Master's
2016	**James B. Angell Scholar Award**, Univ. of Mich., Full tuition scholarship for achieving all A's 2+ semesters
2014	**Excellence in the Discipline Award - Philosophy**, William Rainey Harper College, Merit-based award given to one student per school year

Presentations

2024, 'Understanding and Improving Optimization in Predictive Coding Networks' (Talk) - AAAI

2022, 'A Theoretical Framework for Inference Learning' (Poster) - NeurIPS

2022, 'Deriving STDP from Backpropagation through Time' (Poster) - Spiking Neural Networks as Universal Function Approximators (SNUFA)

2021, 'Tightening the Biological Constraints on Gradient-Based Predictive Coding' (Talk) - International Conference on Neuromorphic Computing

2019, ’Can Cyborgs Tell Us Whether AI are Conscious?’ (Talk) - Minds and Machines Conference, Cambridge University