Bruno Magalhaes

Machine Learning, High Performance Computing and Big Data

brunomaga@gmail.com +41 77 487 89 92 citizenships: Portuguese, Swiss
brunomaga brunomaga https://brunomaga.github.io Lausanne, Switzerland
languages: Portuguese, English, French, Spanish waterpolo, skiing, cooking 13/04/2023

Work Experience

present – Sep 2019 | AI Resident » Researcher » Senior Researcher, Microsoft Research, Cambridge, UK

- as Sr Researcher, 2022-: porting Transformer-based models to confidential optical hardware (ongoing). Large ML models scaling via model parallelism, sharding, pipelining, gradient accumulation, checkpointing, IO offloading, shared memory, mixed precision, model compression, and distillation. Likelihood estimators and Gaussian Processes for error quantification and fine-tuning of physical experimental systems. Information encoding (Gray, Huffman) and error correction (LDPC). Mentoring and technical support of junior members and supervision of PhD interns.
- as Researcher, 2021: at Project Silica: computer vision models for million-object classification on 3D glass. Distributed data and shared-memory parallelism. Full-stack MLOps, DevOps and CI/CD. Development and finetuning of ML systems, pipelines and models for local, cluster and cloud compute environments, taking into account hardware and business requirements such as limited memory, low network bandwidth, cost per run and time to solution. Presenter of talks on the fields of *single-node CPU/GPU optimization*, *distributed algorithms* and *AI SuperComputing*.
- as AI Resident, 2019-20: end-to-end development of ML models (pytorch) and pipelines for: (1) ML models for time-series (RNNs, GRUs, Encoder-Decoders, and Bayesian Optimization), to improve load balancing of Exchange email servers with logs on distributed exabyte-scale COSMOS databases; and (2) development of a recommendation system using Graph Neural Nets and DNNs on a trillion-edge graph of meetings, documents, emails and users, stored on a distributed spark databases.

Aug 2019 – Mar 2015 | PhD candidate » postdoctoral researcher, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

- Research, development (C, C++) and publication of improved methods for asynchronous variable-order variable-step optimization and simulation of detailed spiking neural networks on Cray and SGI supercomputers with 10K+ compute nodes.
- Technologies: C, C++; distributed asynchronous runtime systems (HPX-5) for communication (InfiniBand) and computation (concurrency, scheduling) on global memory addressing; dynamic load-balancing; vectorization and cache optimization.
- Teaching assistant for Unsupervised and reinforcement learning, Project in neuroinformatics and *In silico* neuroscience;
- Scientific reviewer for SuperComputing, IPDPS, and ISC conferences. As postdoc: supervision of PhD students and engineers;

Feb 2015 – Mar 2011 | Research Engineer for High Performance Computing, Blue Brain Project, EPFL, Lausanne, Switzerland

- Research, development (C, C++, MPI, OpenMP) and publication of methods for parallel/distributed volumetric spatial decomposition, load balancing, spatial indexing, sorting, I/O, sparse matrix transpose, and graph navigation, that underlie an efficient storage and processing of neural networks on SGI and IBM BlueGene supercomputers with 16K+ compute nodes.

Feb 2011 – Sep 2009 | Junior Architect for IT infrastructures, Noble Group, London, New York, & São Paulo

- Design and configuration of Linux servers and CISCO networks for trading, physical storage and shipping of commodities.

Oct 2008 – Mar 2007 | Analyst programmer, Investment Property Databank (now MSCi Real estate), London, UK

- Development of a search engine and web/windows app (C#, C++) for efficient storage and analytics of financial data.

Education

Jun 2019 – Mar 2015 | PhD Computational Neuroscience, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

- Summary: distributed-parallel optimization & simulation of large neural networks using asynchronous runtime systems;
- PhD thesis nominated for the EPFL PhD excellency award and IBM Research award for best thesis in computational science;
- Visiting researcher at the Center for Research in Extreme Scale Technologies at Indiana University (US), Summers 2015-17

Sep 2009 – Oct 2008 | MSc Advanced Computing, Imperial College London, UK

- Final project *GPU-enabled steady-state solution of large Markov models* based on distributed, multi-core CPU and GPU (CUDA) computation of large Markov models awarded distinction and published at NSMC'10. Finished degree with Merit.

Jul 2007 – Oct 2002 | Licenciatura (5-year BSc) Systems Engineering and Computer Science, University of Minho, Portugal

- Exchange student at the University of Maribor, Slovenia, 2005/06. Intern at IBM and CERN. Finished degree with A, top 10%

Five Selected Publications

full list on scholar.google.com/citations?user=pirWLLgAAAAJ

ongoing Towards finetuning and error quantification of non-linear physical experimental systems via Machine Learning

2020 Fully-Asynchronous Fully-Implicit Variable-Order Variable-Timestep Simulation of Neural Networks, Proc. International Conference on Computational Science (ICCS 2020), Amsterdam, Holland

2019 Asynchronous SIMD-Enabled Branch-Parallelism of Morphologically-Detailed Neuron Models, Frontiers in Neuroinformatics

2019 Fully-Asynchronous Cache-Efficient Simulation of Detailed Neural Networks, Proc. International Conference on Computational Science (ICCS 2019), Faro, Portugal

2019 Exploiting Implicit Flow Graph of System of ODEs to Accelerate the Simulation of Neural Networks, Proc. International Parallel & Distributed Processing Symposium (IPDPS 2019), Rio de Janeiro, Brazil

2015 Reconstruction and Simulation of Neocortical Microcircuitry, Cell 163, 456–492.