Vedant Puri

vedantpuri@cmu.edu | https://vpuri3.github.io/

Summary

Designs transformer architectures with explicit attention to scaling, memory, and communication structure. Developed FLARE, enabling million-token regimes on a single GPU. Implements new architectures in PyTorch and Triton. Background spans high-performance computing, numerical analysis, and computational fluid dynamics.

Research Interests: Efficient attention, language modeling, scientific machine learning.

Education

Carnegie Mellon University (CMU) Jan 2022–Present
Ph.D Mechanical Engineering. Advisors: Levent Burak Kara, Yongjie Jessica Zhang

University of Illinois Urbana-Champaign (UIUC) 2015–2019
B.S. Engineering Mechanics, *B.S. Mathematics.*

Selected Research Contributions

FLARE: Fast Low-Rank Attention Routing Engine | *Efficient attention architectures* 2025
- Derived a flexible low-rank reformulation of self-attention via latent routing
- Reduced quadratic complexity of self-attention to linear complexity while preserving global communication.
- Demonstrated scaling to 1M tokens on a single H100 GPU, attaining over 200× speedup over vanilla self-attention.

FLARE for Language Modeling (Ongoing) | *Decoder-only architectures* 2025–Present
- Extending FLARE to decoder-only next-token prediction with causal attention.
- Enabling adaptive attention state size to control memory and compute during training and inference.
- Implementing fused Triton kernels for causal training, prefill, and decode.

Equation-based PDE modeling with neural fields | *Hybrid data + physics methods* 2025
- Introduced smooth neural fields as nonlinear spatial ansatz functions in equation-based reduced-order modeling.
- Retained physics-based Galerkin time evolution while learning expressive low-dimensional representations.
- Attained 200× speedup over full-order simulations in transport-dominated regimes.

Experience

Carnegie Mellon University | *Research Assistant* Jan 2022–Present
- Phase field simulations of lithium dendritic growth in solid-state batteries.
- Turbulence closure modeling with differentiable physics.

Julia Computing | *Intern Engineer* Apr 2021–Nov 2021
- Built numerical solvers for scientific machine learning ecosystem in Julia.

Carnegie Mellon University | *Research Assistant* Sep 2020–Jan 2021
- Developed differentiable geometry representations and meshing algorithms.

Argonne National Laboratory | *Research Assistant* Mar 2020–Sep 2020
- Executed large-scale simulations of turbulent airflow over urban landscapes on supercomputers.

Argonne National Laboratory | *Research Assistant* May 2018–Jul 2018
- Executed high-fidelity fluid dynamics simulations and analyzed turbulent statistics for closure modeling.

National Center for Supercomputing Applications | *Intern* Sep 2017–May 2018
- Numerical simulation of spacetime metric for gravitational wave simulations.

Publications

Puri, V. et al., *FLARE: Fast Low-Rank Attention Routing Engine.* arXiv 2025. (Under review)
Puri, V. et al., *SNF-ROM: Projection-based nonlinear reduced order modeling with smooth neural fields.* JCP 2025.
Shankar, V., **Puri, V.**, et al., *Differentiable physics closure modeling for Burgers' turbulence.* MLST 2023.

Awards

World Conference on Computational Mechanics | *Best poster in fluid dynamics* 2024
University of Illinois | *Theoretical and Applied Mechanics Merit Award* 2019

Technical Skills

Machine Learning Systems: PyTorch, Triton, mixed-precision/distributed training, causal language modeling
Numerical Computing: Numerical analysis, scientific computing, linear algebra, finite elements, spectral methods
Programming Languages: Python, Julia, C, Fortran 77, MATLAB, UNIX, LaTeX
Modeling Domains: Transformer architectures, neural operators, reduced-order modeling