Università degli Studi di Napoli Federico II

Narendra Patwardhan

Edge-efficient Neural Networks for NLP and Health Monitoring

Tutor: Prof. Carlo Sansone

Cycle: XXXVIII Year: First

My background

- MSc degree:
  Mechanical Engineering @ Michigan Technological University

  Thesis - Proximal Reliability Optimization for Reinforcement Learning
- Research Group: **PICUS Lab**
- PhD start date: 01/11/2022
- Scholarship type: PNRR
- Partner company: SIMAR GROUP s.r.l., Monte Urano (FM)

Research field of interest

Enhancing efficiency of Neural Networks for **Edge Deployment**

Sustainability for Large Language Models

Overcoming Resource Scarcity in Computer Vision

Non-Invasive **Health Monitoring**

Summary of study activities

- Ad hoc PhD courses
  - Statistical Data Analysis for Science and Engineering Research
  - Introduction to Deep Learning
  - Academic Entrepreneurship

- PhD Schools
  - Spring School on Transferable Skills
  - International Summer School on Machine Vision (VISMAC-23)

- Conferences attended
  - 22nd International Conference on Image Analysis and Processing (ICIAP-23)

Research activity I

Problem

How to reduce the computational footprint of large language models?

Objective

To make large language models sustainable, accessible, and fair

Methodology

Survey real-world applications

- Scraped PapersWithCode
- Filtered papers based on Open-Access and relevance
- Created a classification schema
- Read & presented best approaches among each usecase

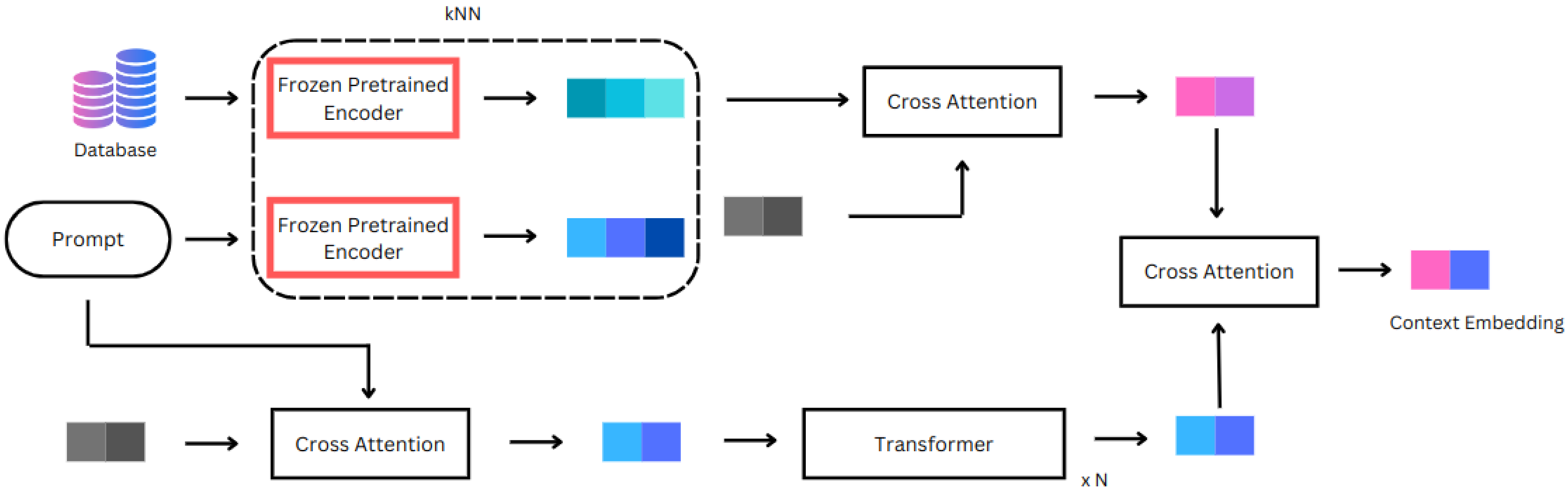
Identify avenues for improvement

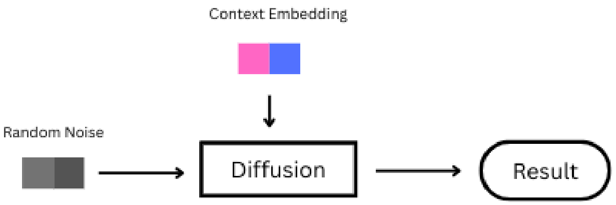
- Identify bottlenecks within the Transformer block
- Measure the impact of data-cleaning methods

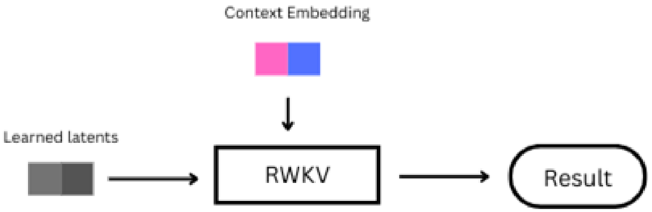
Propose Architecture level Changes

- Propose sustainable alternatives to each component to (I) minimize training time / (II) improve forward pass latency.
- Replace the decoder with different mechanisms.

Proposed Hominis Architecture







Research activity II

Problem How to utilize deep learning techniques in resource-constrained visual domains?

Objectives To obtain high-quality training data with minimal expert involvement.

To provide an architecture with high modeling efficiency that is adaptable to novel visual domains.

Methodology

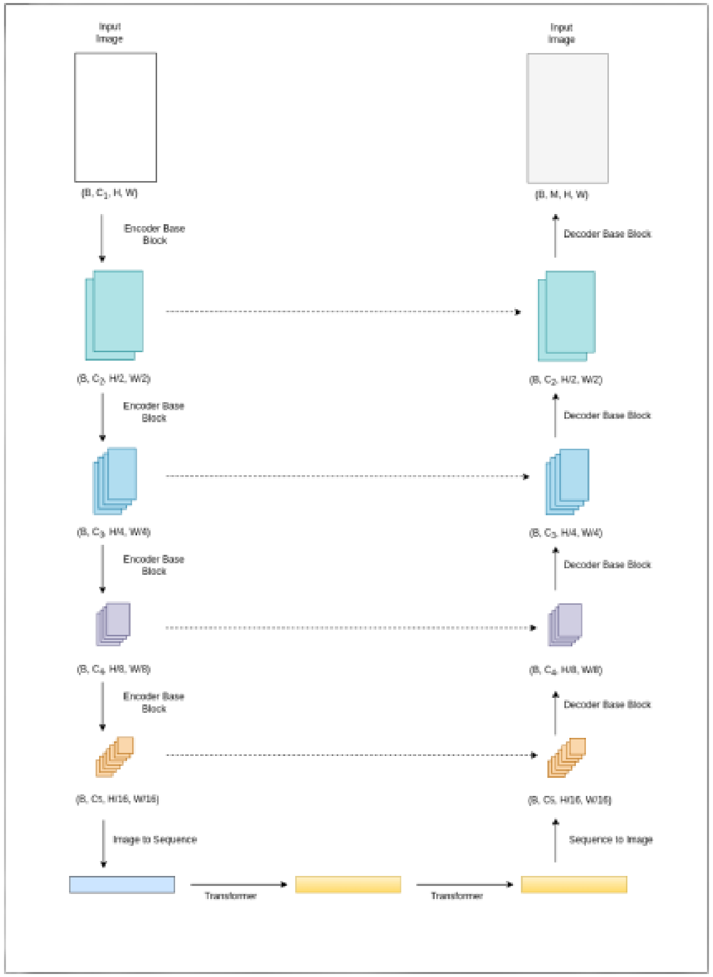
Design a simulator

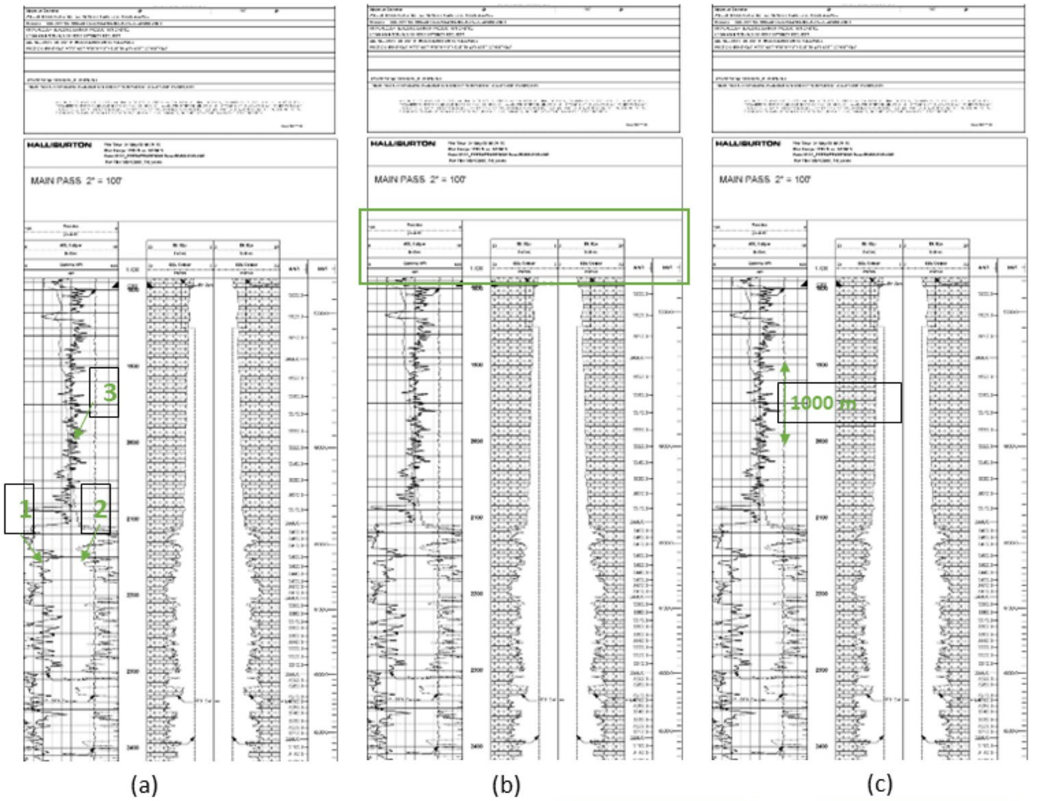
- Obtain a low amount of annotated data.
- Identify/Create a generator
- Bias the output based on domain knowledge and statistics
- Identify and add equivariant transforms for randomness

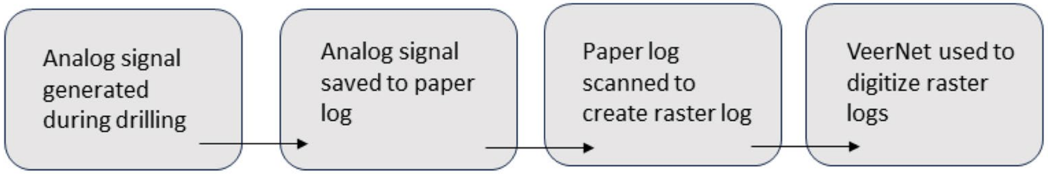
Balance the inductive bias within a model

- Identify a task-invariant architecture -> U-NET
- Retain useful biases such as translation equivariance.
- Minimize other inductive biases through the use of Attention (permutation equivariance)

Veernet Architecture







Analog signal generated during drilling

Analog signal saved to paper log

Paper log scanned to create raster log

VeerNet used to digitize raster logs

Research activity III

Objective

To design a smart-seat for non-invasive health monitoring

Methodology

Perform Feasibility Study

- Identify which proposed features can be solved considering the existing techniques and potential bottlenecks
- Identify avenues for innovation

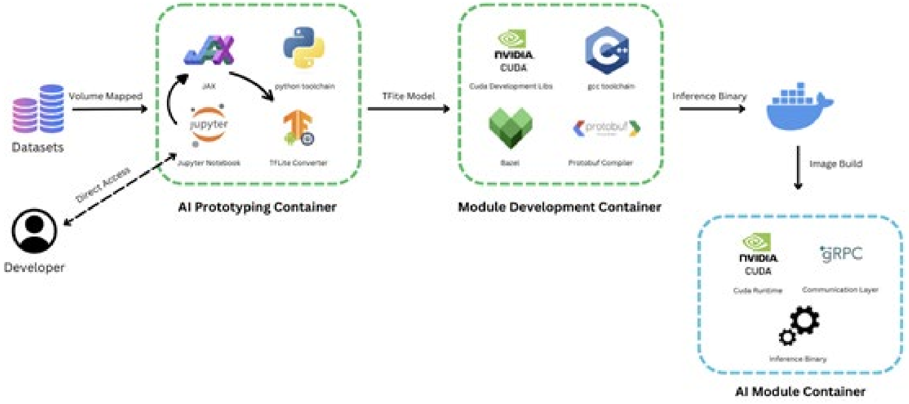
Propose Hardware & Software Components

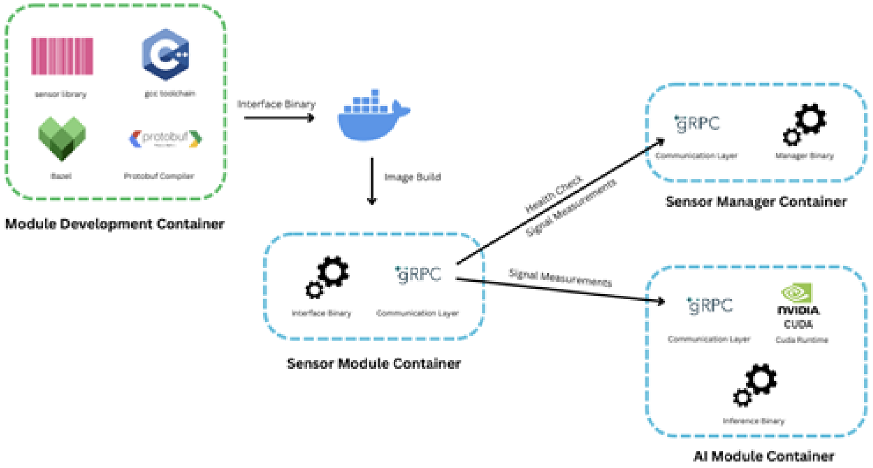
- Identify a common software stack with competency in sensor data collection & and deep learning (C++, TFLite, Bazel, Docker)
- Handle asynchronous sensor readings (Pseudosensor network, serialization schema)
- *Currently underway*

>> Natural Language Processing capabilities for user interaction

>> Camera as the main pseudo-sensor, useful learnings from data-constrained domain

Software Stack for SIMAR Smart Chair





Products

[P1]	**Patwardhan, N.**, Marrone, S., & Sansone, C. (2023). "Transformers in the Real World: A Survey on NLP Applications." Information, 14(4), 242. (Published)
[C1]	**Patwardhan, N.**, Marassi, L., Gravina, M., Galli A., Zuccarini, M., Maiti, T., Singh, T., Marrone, S., & Sansone C. (2023) "Responsible and Reliable AI at PICUS Lab." Convegno Nazionale CINI sull'Intelligenza Artificiale, Ital-IA 2023. (Published)
[P2]	Nasim, M. Q., **Patwardhan, N.**, Maiti, T., Marrone, S., & Singh, T. (2023). “VeerNet: Using Deep Neural Networks for Curve Classification and Digitization of Raster Well-Log Images” Journal of Imaging, 9(7), 136. (Published)

Products

[C2]	Nasim, M. Q., **Patwardhan, N.**, Ali, J., Maiti, T., Marrone, S., Singh, T., & Sansone, C. (2023). “Digitizer: A Synthetic Dataset for Well-Log Analysis” 22nd International Conference on Image Analysis and Processing, ICIAP-23. (Published)
[C3]	Marassi, L., **Patwardhan, N.**, & Gargiulo F. (2023). “Can Justice Be a Measurable Value for AI? Proposed Evaluation of the Relationship Between NLP Models and Principles of Justice“ The First Workshop on User Perspectives in Human-Centred Artificial Intelligence, HCAI4U (Published)
[C4]	**Patwardhan, N.**, Shetye, S., Marassi, L., Zuccarini, M., Maiti, T., & Singh, T. (2023). “Designing Human-Centric Foundation Models” The First Workshop on User Perspectives in Human-Centred Artificial Intelligence, HCAI4U (Published)

Next Year

- Working towards the SIMAR smart-chair prototype
- Training the Hominis language model (with ISCRA-B grant or other avenues)
- Explore if hardware-accelerator specific methods such as flash-attention could be extended for Hominis & SIMAR.