Human Activity Recognition

Hardware Acceleration of Deep Learning Algorithms

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Louie Cai
louie@ucsd.edu

Jeffery Lai
j8lai@ucsd.edu

Lian Gan
lgan@ucsd.edu

Christian DerManuelian
cdermanuelian@ucsd.edu

Rajesh Gupta (Mentor)
rgupta@ucsd.edu

Background & Research Question

- Our project upgrades computer systems to identify human activities quickly and cost-effectively.
- HAR: Systems designed to automatically identify and classify activities or actions performed by humans based on sensor data or other input sources.
- We test how different computers perform in recognizing activities, aiming for speed and affordability.
- Results will guide choosing the best tech for activity recognition, useful across health, security, and more.

Timeseries	Feature1	...	Feature 52	Activity
0.01s				
...				Run
3.00s				

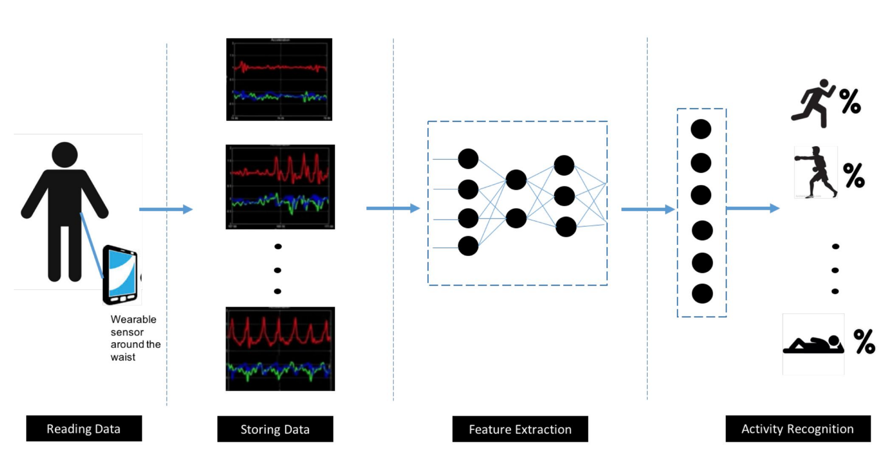
Data Collection

- We test how different computers handle activity recognition, measuring speed and efficiency.
- Real-world activity data ensures our tests are practical and relevant.
- Tests are run in a controlled environment on cloud platforms for consistent results.
- Our aim is to find the fastest and most cost-effective computer setups for recognizing activities.

Methods Overview

- Our project examines performance differences across various wearable and single-core CPU hardware for Human Activity Recognition (HAR) neural networks, focusing on efficiency and scalability.
- A diverse range of HAR neural network architectures, including, CNNs, DNNs, RNNs(LSTM), and Transformers, are selected to understand how computational requirements affect performance across hardware spectra

SCAN ME

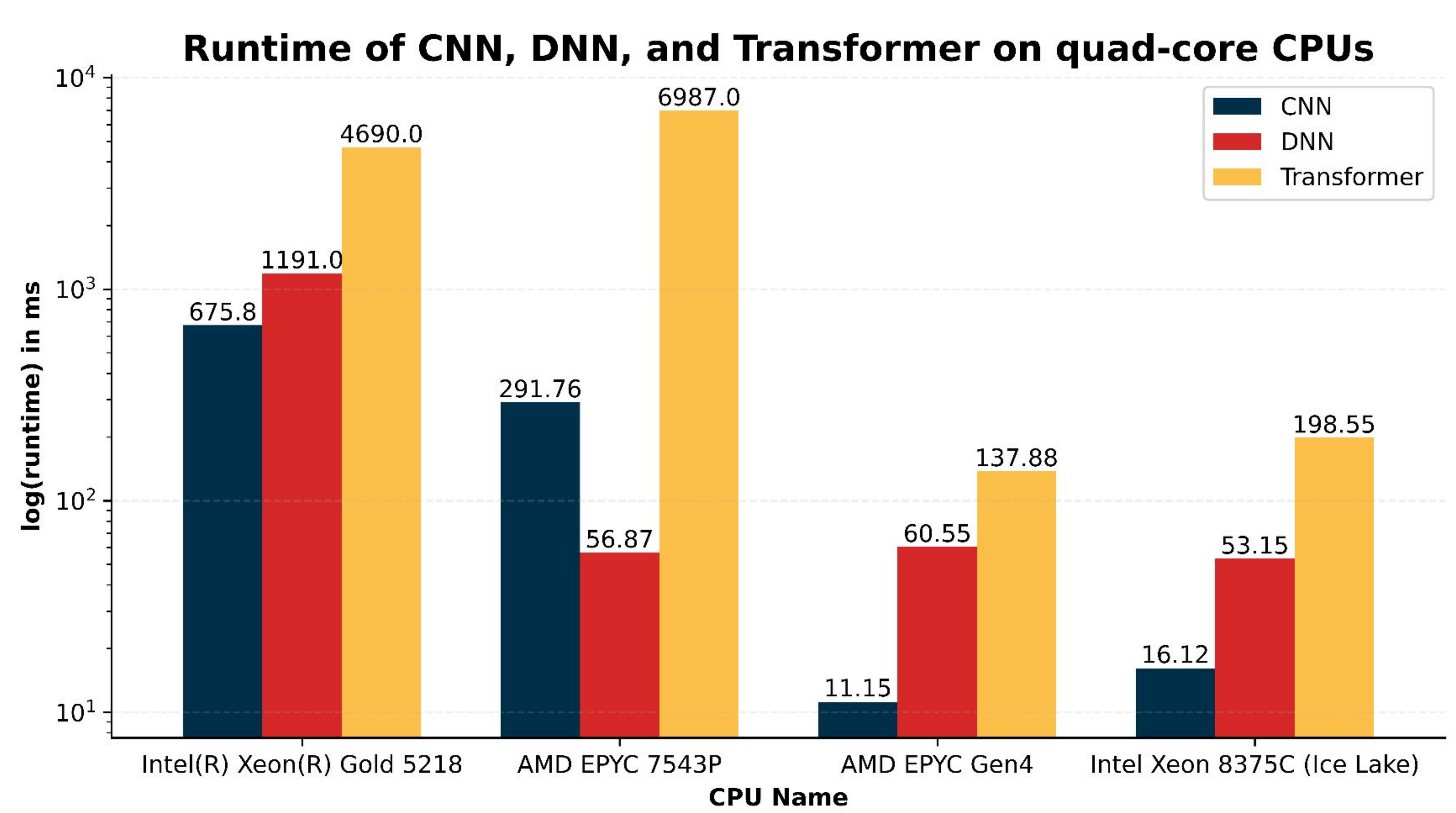


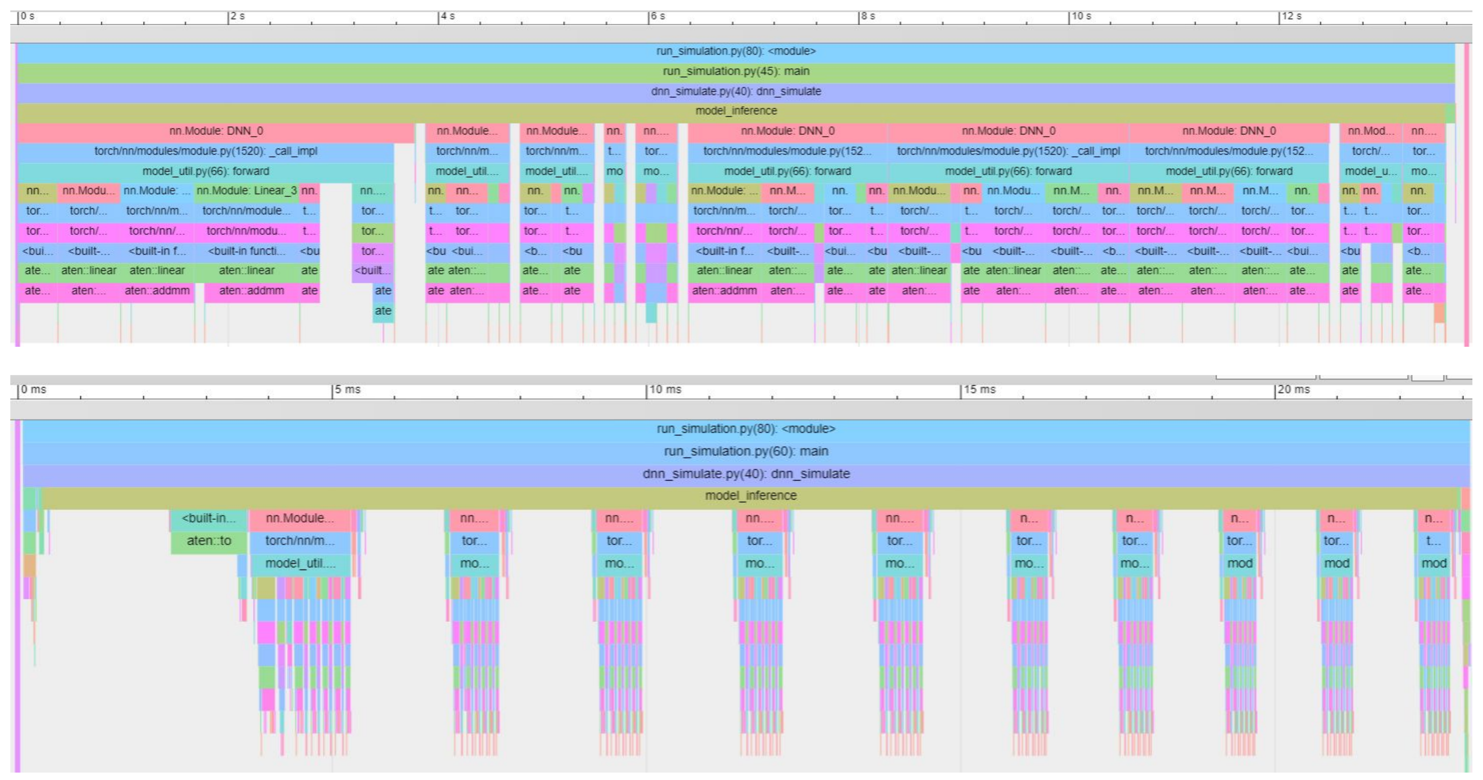
Experimental Design

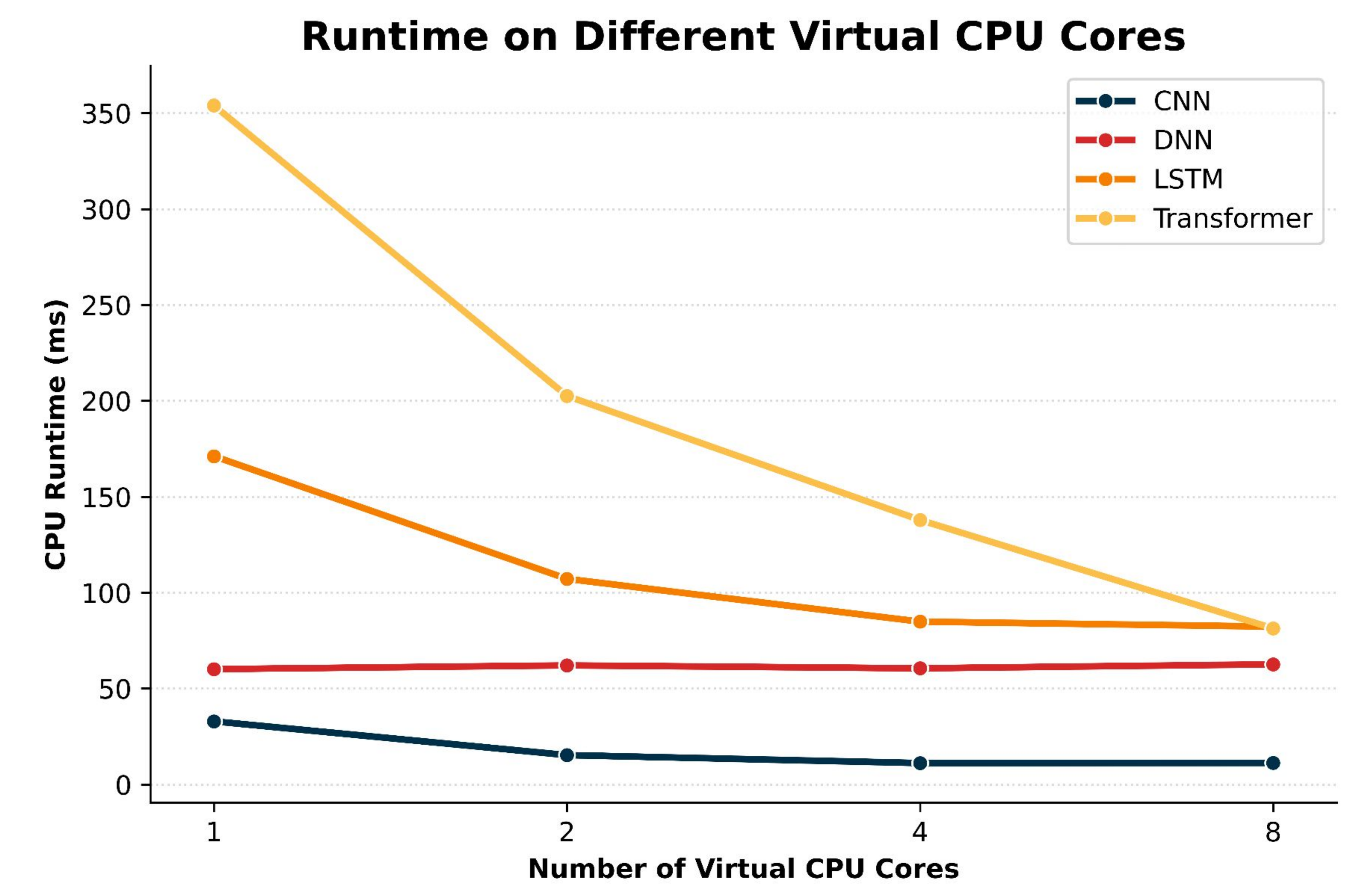
- We are investigating performance trade-offs in Human Activity Recognition (HAR) on various hardware and examines computational efficiency and scalability of different deep learning models.
- A range of hardware configurations, from smartwatches to 8-core CPUs, and a broad spectrum of data handling capabilities, from 4GB to 32GB RAM, are tested for their impact on HAR deep learning model performance.
- A Docker container environment is used to maintain a controlled and consistent experimental setup, utilizing TensorFlow and standard versions of Python for model training and evaluation, ensuring reproducibility across tests.

Performance Metrics

- The primary focus is on measuring speedups across different hardware configurations and how computational power and memory affect model training and inference times.
- The approach isolates hardware's impact on speedup, providing clear insights into performance variations.
- Detailed metrics such as training and inference times, among others, are part of ongoing research and will be defined and analyzed thoroughly.







Summary of Findings

- Increase number of cores, exponentially accelerate speed
- CPU varys runtime a lot, even if some of cpu selling for the same price, due to cpu architecture.
- Using GPU significantly decrease run time
- Some high performance CPU can beat GPU, but when we increase batch size(number of samples), GPU's runtime is thousands time better than cpu due to its ability of paralleling computing

What's Next

- 60-90% runtime are spent on convolution or matrix multiplication.
- We can use FPGAs to optimize bottlenecks in the code through parallel processing, outperforming general purpose CPUs.