Evaluation of Deep Predictive Learning for AI-driven Robots on an Edge-GPU

Keiji Kimura, Zhu Yunkai, Dan Umeda,
Hiroshi Ito, Tetsuya Ogata
Waseda University

AIREC (AI-driven Robot for Embrace and Care) Project in Waseda University

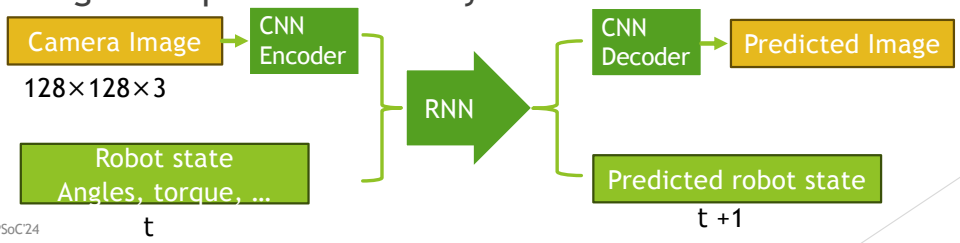
- **GOAL: Realizing AI-driven Robots helping our daily life.**
- **Soft Robotics**
  - Combining flexible machine hardware and *AI innovation* for advanced environment adaptability
- **Physical Intelligence and Mutually Induced Communication Intelligence**
  - *Enabling flexible response* to and interaction with real space

Predictive Deep Learning and Low-Power AI Accelerator are the key technologies

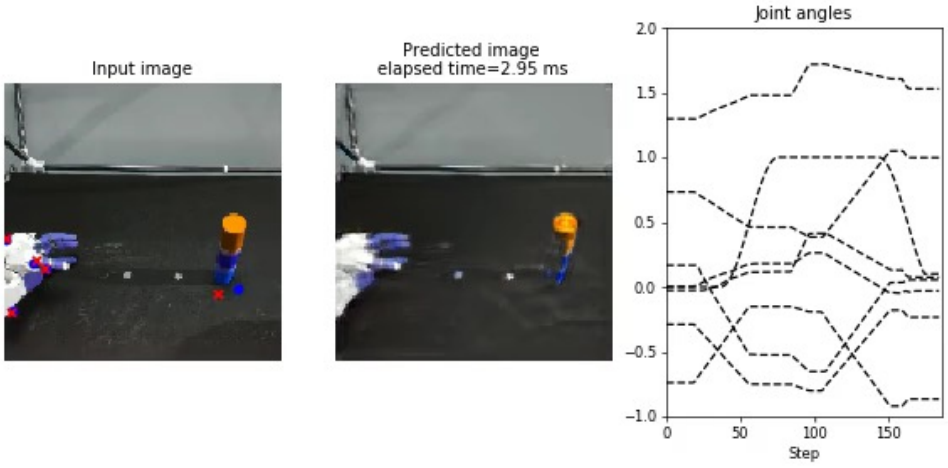
EIPL: Embodied Intelligence with Deep Predictive Learning

https://github.com/ogata-lab/eipl

- Robot motion generation library using deep predictive learning
  - Developed by Ogata-lab, Waseda Univ.
- Image features (CNN) and robot state (ex. motor angles) are given to RNN to predict the robot's next state.
- Original implementation: Pytorch

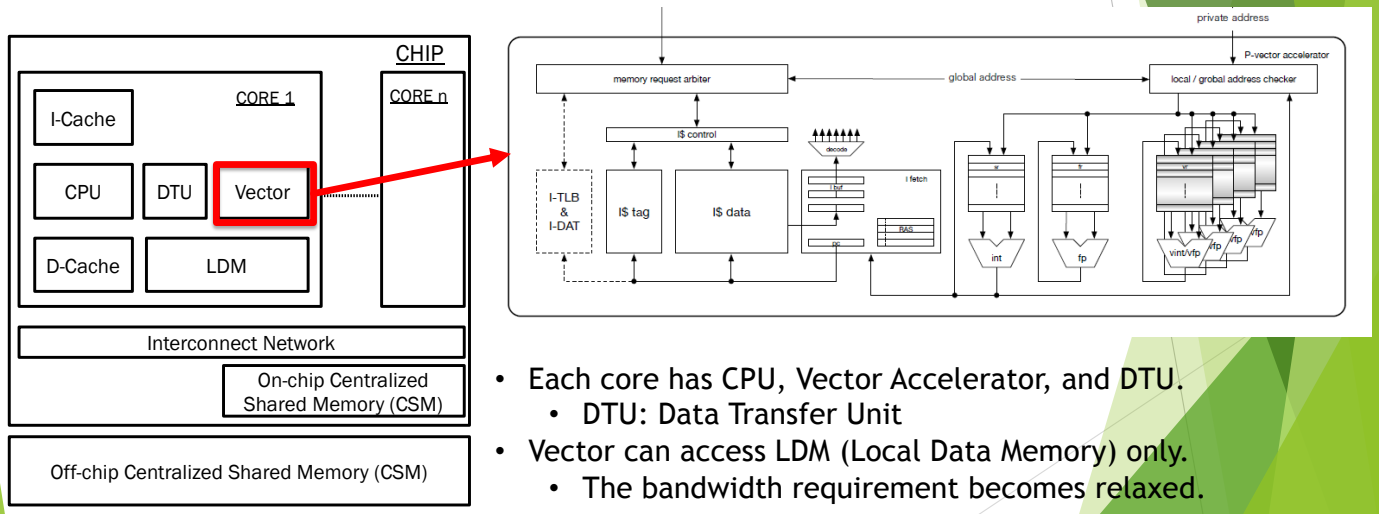


EIPL Execution Image



Dotted lines: Actual angles
Solid lines: Predicted angles

AI-Accelerator Developed in AIREC Project: OSCAR Compiler Cooperative Vector Multicore

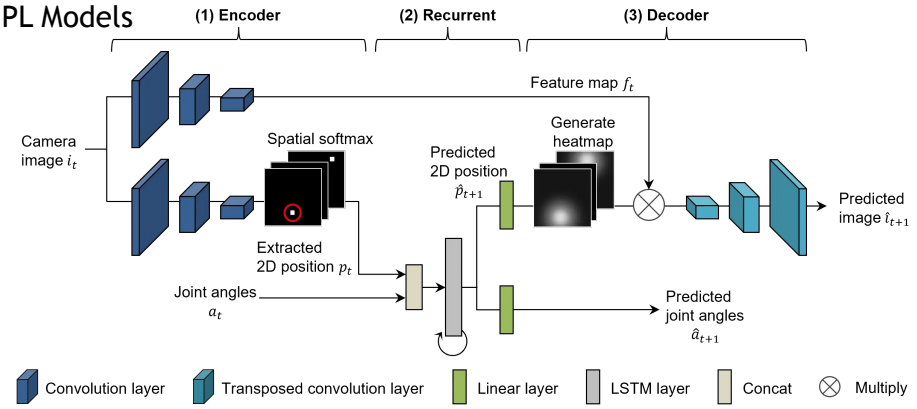


- Each core has CPU, Vector Accelerator, and DTU.
  - DTU: Data Transfer Unit
- Vector can access LDM (Local Data Memory) only.
  - The bandwidth requirement becomes relaxed.

What are the best architecture and required function units for EIPL?

EIPL Model: SARNN
Spatial Attention with Recurrent Neural Network

One of EIPL Models



Other Models:

- CNNRNN (simplified version)
- CNNRNNLN (CNNRNN with Layer Normalization)

Exploring Optimization

- Introducing TensorRT
  - Replacing Pytorch
- Quantization
  - Very popular for image processing
  - Original: FP32
  - Evaluated: FP16, INT8
  - Expecting more computational throughput, less energy consumption, and less memory bandwidth

Evaluation Platform

- Jetson Orin Nano 4GB
  - Ampere architecture GPU
    - **512** CUDA core
    - **16** Tensor Cores
    - 306 - **625**MHz
  - 4GB 64-bit LPDDR5 Memory
    34GB/s
  - 6-core Arm® Cortex®-A78AE v8.2 64-bit CPU
    1.5MB L2 + 4MB L3
    1.5 GHz

Evaluation Result: Inference Time

- Compared to FP32…
  - SARNN
    - FP16: 1.79x, INT8: 1.53x
  - CNNRNN
    - FP16: 1.42x, INT8: 1.47x

Data conversion overhead
FP32 <> FP16/INT8
- Utilizing images and angles
- Utilizing several math functions

Inference time on Jetson Orin Nano 4GB [ms]

	PyTorch	TensorRT		
Model	FP32	FP32	FP16	INT8
SARNN	5.13	4.42	2.47	2.88
CNNRNN	3.11	1.87	1.32	1.27
CNNRNNLN	4.13	3.55	2.77	2.88

Evaluation Result: Energy

- Compared to FP32…
  - SARNN
    - FP16: 61.9%, INT8: 66.7%
  - CNNRNN
    - FP16: 85.0%, INT8: 80.0%

Data conversion overhead is a significant factor.

Model	Data Type	Power [mw]	Energy [mJ]	Energy / Frame [mJ]
SARNN	FP32	7,219	397,395	42
	FP16	6,689	234,353	26
	INT8	6,490	259,854	28
CNNRNN	FP32	6,218	180,513	20
	FP16	6,350	152,544	17
	INT8	5,934	136,619	16
CNNRNNLN	FP32	6,601	297,334	32
	FP16	6,622	265,167	29
	INT8	6,429	250,989	28

Summary

- EIPL: Enabling flexible AI-Robot control in the real-world
- OSCAR Compiler Cooperative Vector Multicore
  - For Power-Efficient AI Acceleration in Robots
  - We need to explore the required architecture and function units!
- Evaluation of EIPL on Jetson for exploring the expected architecture
  - We need to revise the appropriate data precision.
  - Of course, we need to tune the software more.

- This work is supported by JST [Moonshot R&D][Grant Number JPMJMS2031].