CSE3000

Scheduling Multi-inference with Constrained Memory

Jeroen Galjaard - Professor: Dr. Lydia Chen - Supervisors: Amirmasoud Ghiassi, Bart Cox

1 Multi-inference

Executing multiple deep neural networks (DNN) on low-powered devices. Splitting the networks into layers for a layer-by-layer fashion. Networks contain tasks that are mostly either IO or CPU bound in execution.

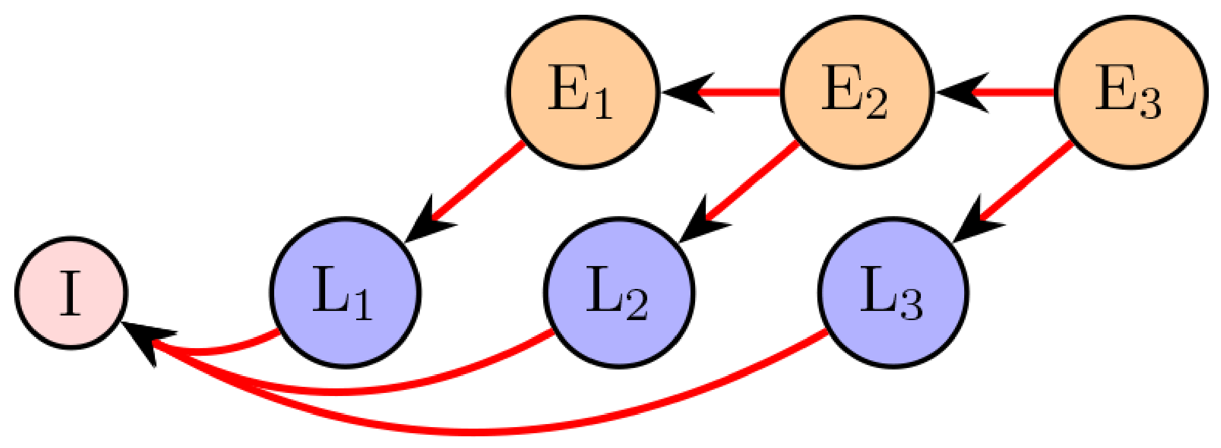

Figure 1: Partial loading, DNN layers can be arbitrarily loaded.

4 Conclusion

Limited effects on small jobs, layer loading policies do the ‘heavy lifting’ (not depicted). However, scheduling policies significantly affect large jobs with stringent memory. MEMA shows a considerable gain over baseline performance.

2 Topic

Effect of scheduling policies affect on multi-inference jobs?

- FCFS, SJF, LJF, and a novel MEMory Aware (MEMA) scheduling policy
- Bulk, linear, DeepEye [1], and partial loading (Figure 1)

3 Results

Minimal impact when job sizes are small (not depicted), or high memory availability (see Figure 2). However, when more stringent, differences become more pronounced (see Figure 3).

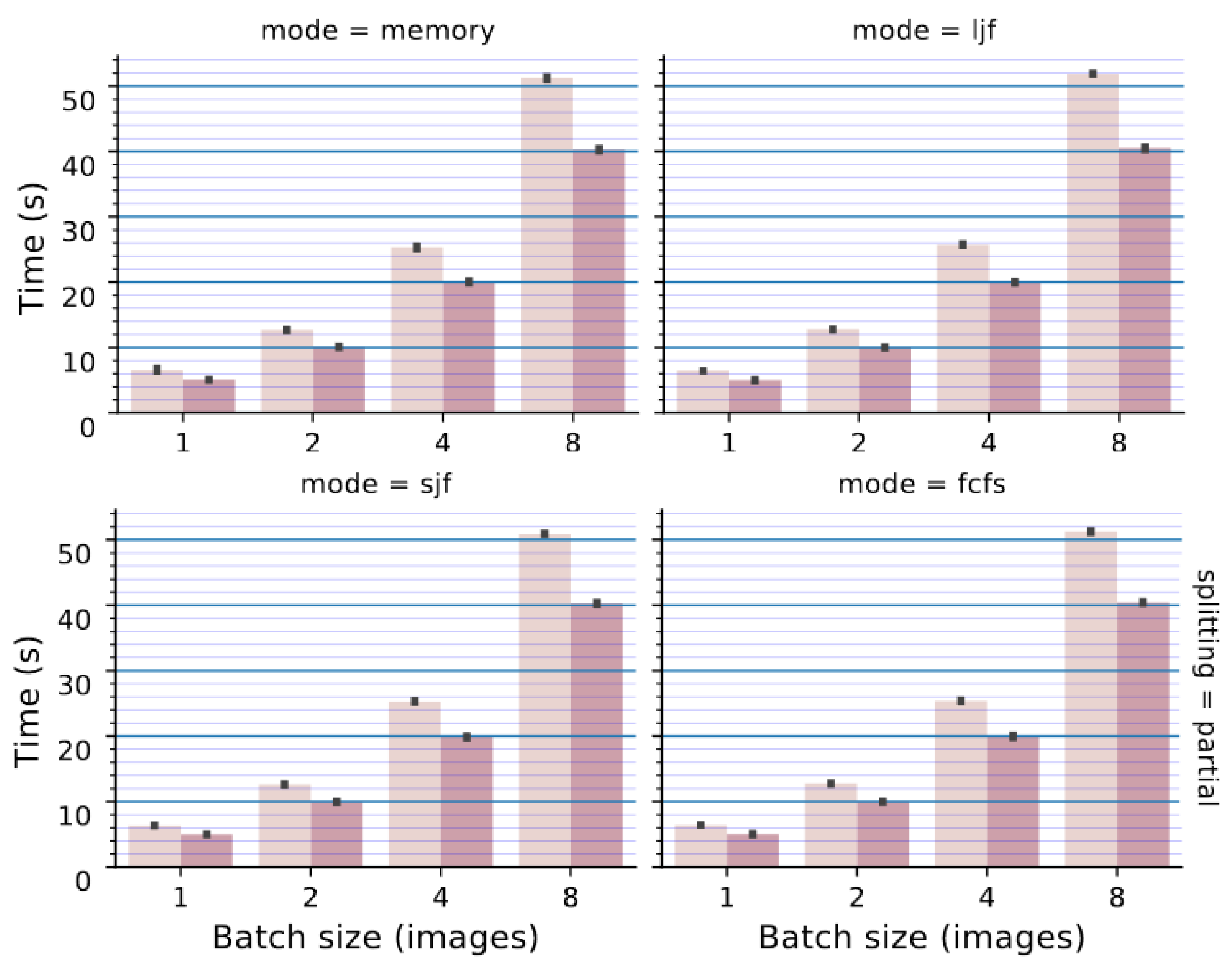

Figure 2: Unconstrained performance (2G RAM). Inference speed is not affected by scheduling policies.

Cont'd 3

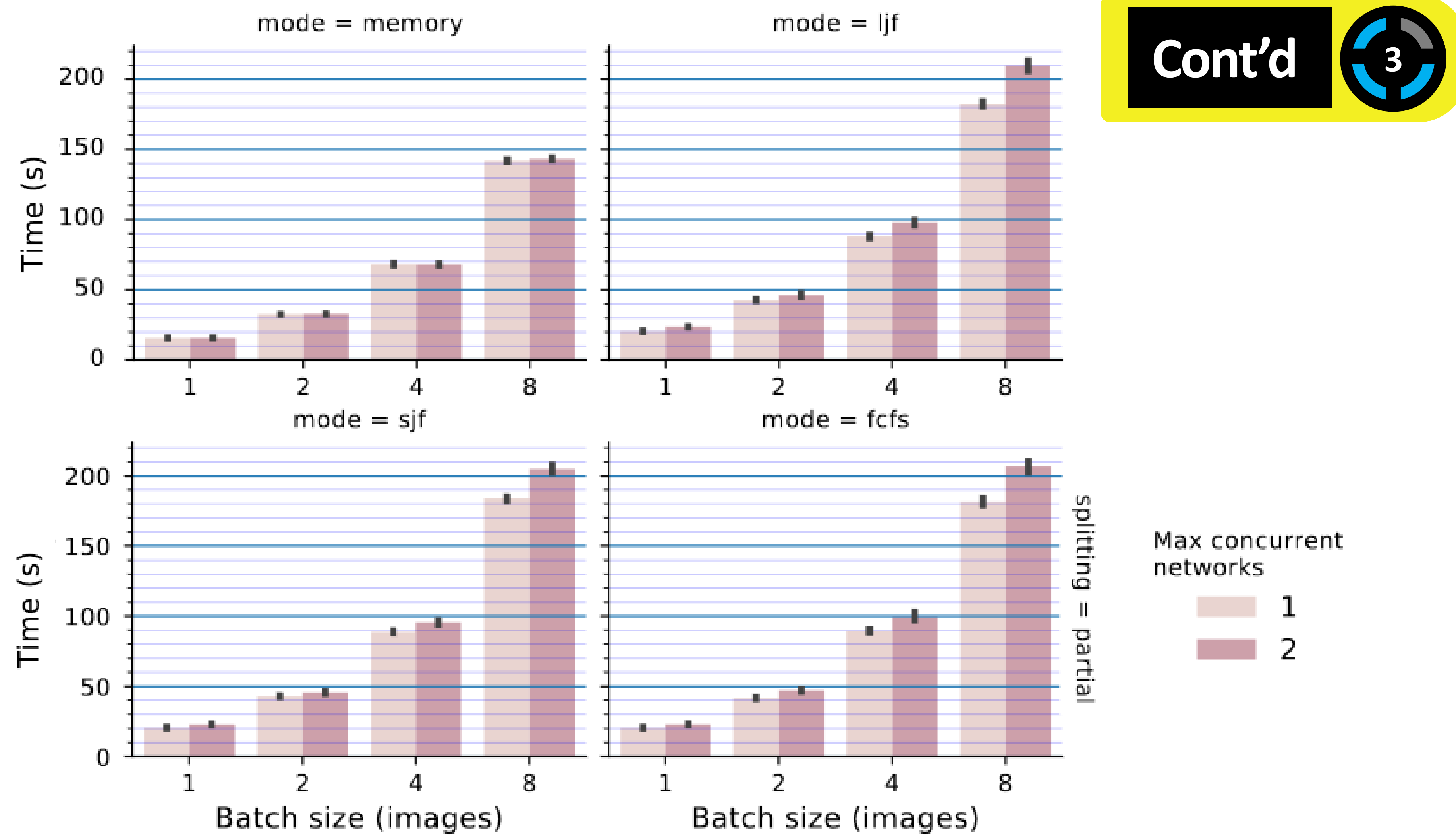

Figure 3: Performance under stringent memory (256MB RAM). MEMA (left) shows a significant improvement over other policies.

✉ j.m.galjaard-1@student.tudelft.nl
✉ l.chen-10@tudelft.nl
✉ m.ghiassi@tudelft.nl
✉ bacox@student.tudelft.nl

TU Delft

[1] Mathur, Akhil et al. "DeepEye: Resource Efficient Local Execution of Multiple Deep Vision Models Using Wearable Commodity Hardware." Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services. Association for Computing Machinery,