DjiNN and Tonic: DNN as a Service and Its Implications for Future Warehouse Scale Computers

Johann Hauswald, Yiping Kang, Michael A. Laurenzano, Quan Chen, Cheng Li, Trevor Mudge, Ronald G. Dreslinski, Jason Mars, Lingjia Tang

University of Michigan — Ann Arbor, MI
Deep Neural Networks (DNN) as a Service

1. DNN Extraction
   - Image Classification
   - Digit Recognition
   - Facial Recognition
   - Speech Recognition
   - Natural Language Processing

2. High Throughput

3. Server Design
   - CPU-Only
   - Integrated GPU
   - Disaggregated GPU

djinn.clarity-lab.org
Real System

133x

771x

4-20x

TCO Improvement

First session right now

djinn.clarity-lab.org