Novel paradigm reduces **communication** within distributed neural network **as much as 8x**.

RingTMP: A Communication Optimised Distributed Neural Network

Chris Hopkins

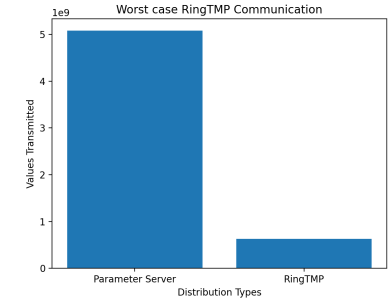
1 Intro

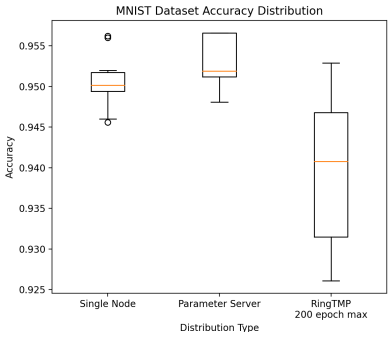
- Current distributed machine learning (DML) models are reaching the limits of scalability due to increase communication leading to network saturation, ultimately limiting training speed.
- Using a new paradigm to distribute nodes could result in less communication therefore more scalability & reduced training times.

2 Problem Description

- Neural networks function by using using parameters in each layer to transform the input, and pass the activation to then next layer.
- The parameter server (a popular DML paradigm) functions by broadcasting all parameters to its workers, which reply with a modified version of those parameters.
- Can we decrease communication by passing activations rather than parameters, would this affect training?

3 Results

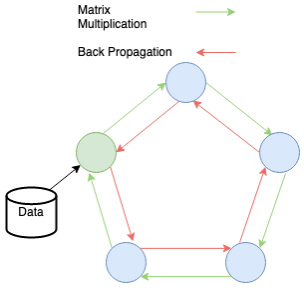




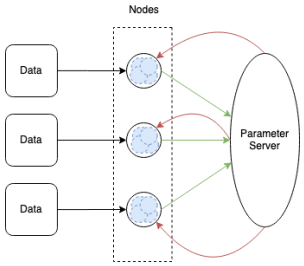
- We trained a single node, a parameter server and a RingTMP implementation on the Iris and MNIST numbers dataset (10, 5 times respectively)
- We found while RingTMP produced slightly less accurate models, but reduces communication 8x per unit time even in the worst case.

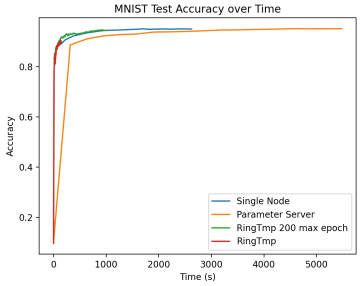
Extra material

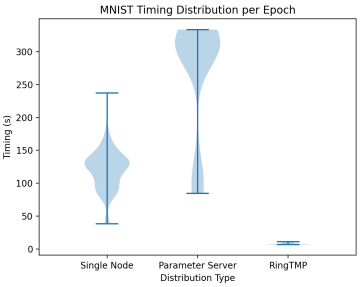
RingTMP



Parameter Server







SCAN ME

← Download the paper

Swansea University
Prifysgol Abertawe