Efficient Training of Giant Neural Networks
Zeyuan Tan, Man-Kit Sit, Luo Mai
School of Informatics, University of Edinburgh

Many strategies are available to parallelise the training of giant neural networks

Giant neural networks: Trillions parameters (e.g., GPT-3, SimCLR)
Parallel training: 100s parallel CPU/GPU devices must be coordinated to overcome memory limits

Research question: How to find best parallelism strategies that can minimise model training time?

State-of-the-art parallel training systems rely on manual & static configuration

Manual configuration (GSPMD [1], MindSpore [2])

Issues:
- Manual configuration requires expert knowledge about parallel training, which is not available on ML developers
- Static parallelism configuration suffer from inaccurate static cost model & model diversity

Designing an efficient dynamic parallelism system for giant neural networks

#1: Programming giant DNNs as on a single device
#2: Augmenting cost models with runtime parallelism metrics
#3: Dynamic compilation for parallelism plans

---