Distributed Training of Deep Neural Networks

Abhinav Bhatele, Siddharth Singh, Daniel Nichols
Department of Computer Science

PSSG PARALLEL SOFTWARE
AND SYSTEMS GROUP

UNIVERSITY OF MARYLAND

Welcome!

Are you interested in training deep learning models on a networked cluster of GPUs?

- If the answer is yes, you have come to the right tutorial.

Tutorial attendees will learn …

- Different forms of parallelism for deep learning
- Setting up a training job on GPGPUs
- Different parallel deep learning frameworks and how to use them

Tutorial organizers

Abhinav Bhatele
Associate Professor, CS@UMD

Siddharth Singh
Doctoral Student, CS@UMD

Daniel Nichols
Doctoral Student, CS@UMD

Deep learning

- An area of machine learning that uses artificial neural networks to learn complex functions
    - Often from high-dimensional data: text, images, audio, …
- Widespread use in computer vision, natural language processing, etc.

Parallel / distributed training

- Many opportunities for exploiting parallelism

- Iterative process of training (epochs)

- Many iterations per epoch (mini-batches)

- Many layers in DNNs

Increase in size of neural networks
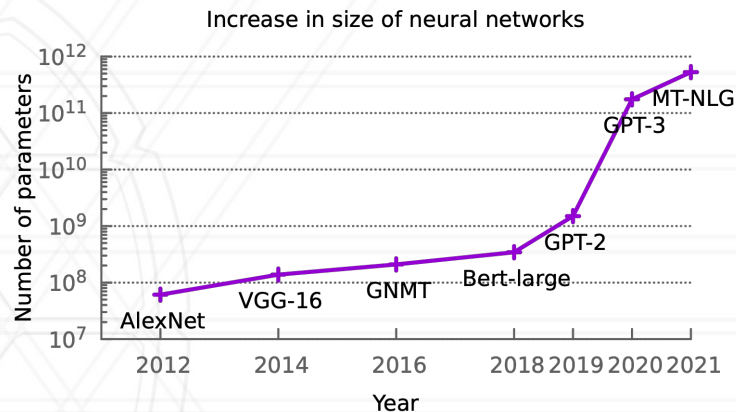
Framework	Type of Parallelism	Largest Accelerator Count	Largest Trained Network (No. of Parameters)
FlexFlow	Hybrid	64 GPUs	24M*
PipeDream**	Inter-Layer	16 GPUs	138M
DDP**	Data	256 GPUs	345M
GPipe	Inter-Layer	8 GPUs	557M
MeshTensorFlow	Intra-Layer	512-core TPUv2	4.9B
Megatron**	Intra-Layer	512 GPUs	8.3B
TorchGPipe**	Inter-Layer	8 GPUs	15.8B
KARMA	Data	2048 GPUs	17B
LBANN**	Data	3072 CPUs	78.6B
ZeRO**	Data	400 GPUs	100B
ZeRO-Infinity	Data	512 GPUs	32T
AxoNN	Inter-Layer	384 GPUs	100B

Accounts

- Tutorial link: https://github.com/axonn-ai/distrib-dl-tutorial

Abhinav Bhatele and Siddharth Singh

Department of Computer Science

bhatele@umd.edu, ssingh37@umd.edu