ICML
International Conference
On Machine Learning

TEXAS
The University of Texas at Austin

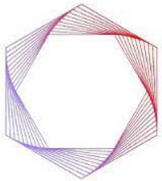
LOWERING THE PRETRAINING TAX FOR GRADIENT-BASED SUBSET TRAINING

Yeonju Ro, Zhangyang Wang, Vijay Chidambaram, and Aditya Akella

University of Texas at Austin

Tue 25 Jul 2PM-3:30PM
Exhibit Hall 1 #439

Emergence of Large-scale Models and Datasets



Dataset	# Eng Img-Txt Pairs
MS-COCO	300K
CC3M	3M
Visual Genome	5.4M
WIT	5.5M
CC12M	12M
CLIP-WIT (private)	400M
LAION-5B	2.3B
BASIC (private)	6.6B

* Images from StableAI, Meta, HuggingFace

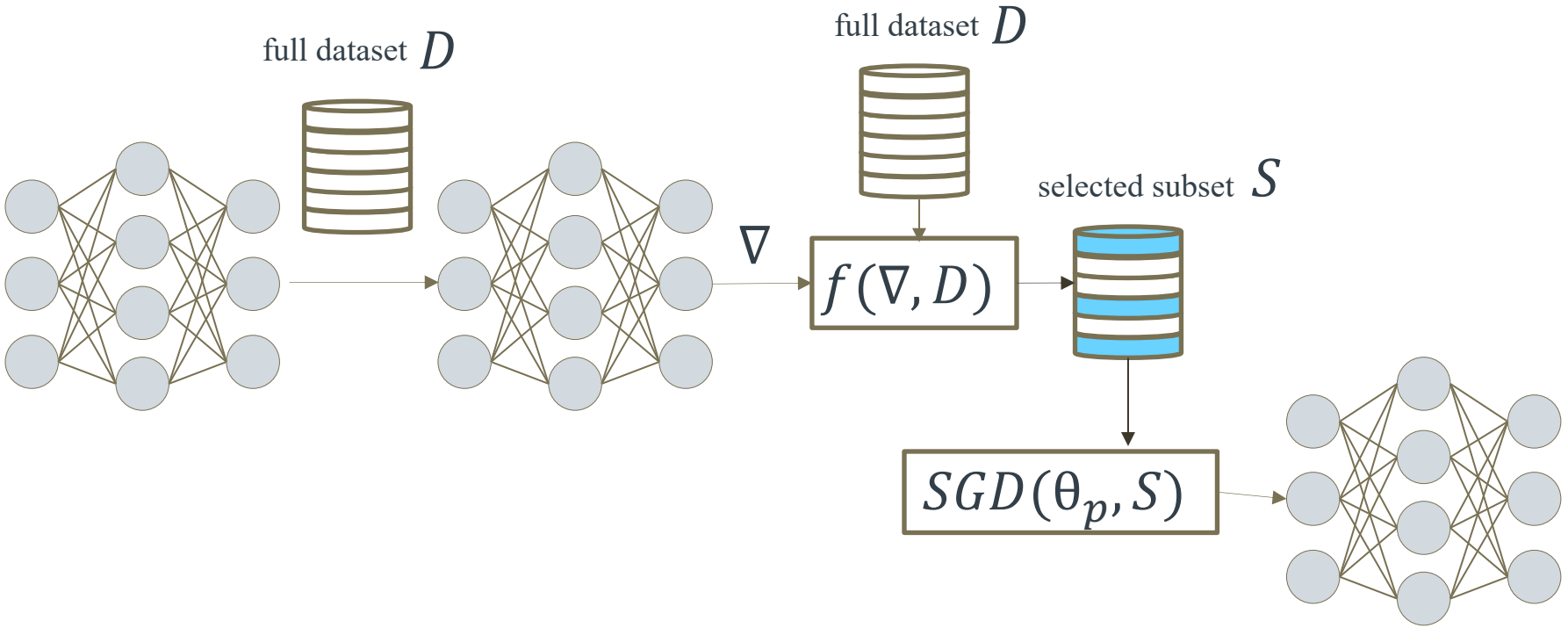
* Data from LAION-5B

Subset Training

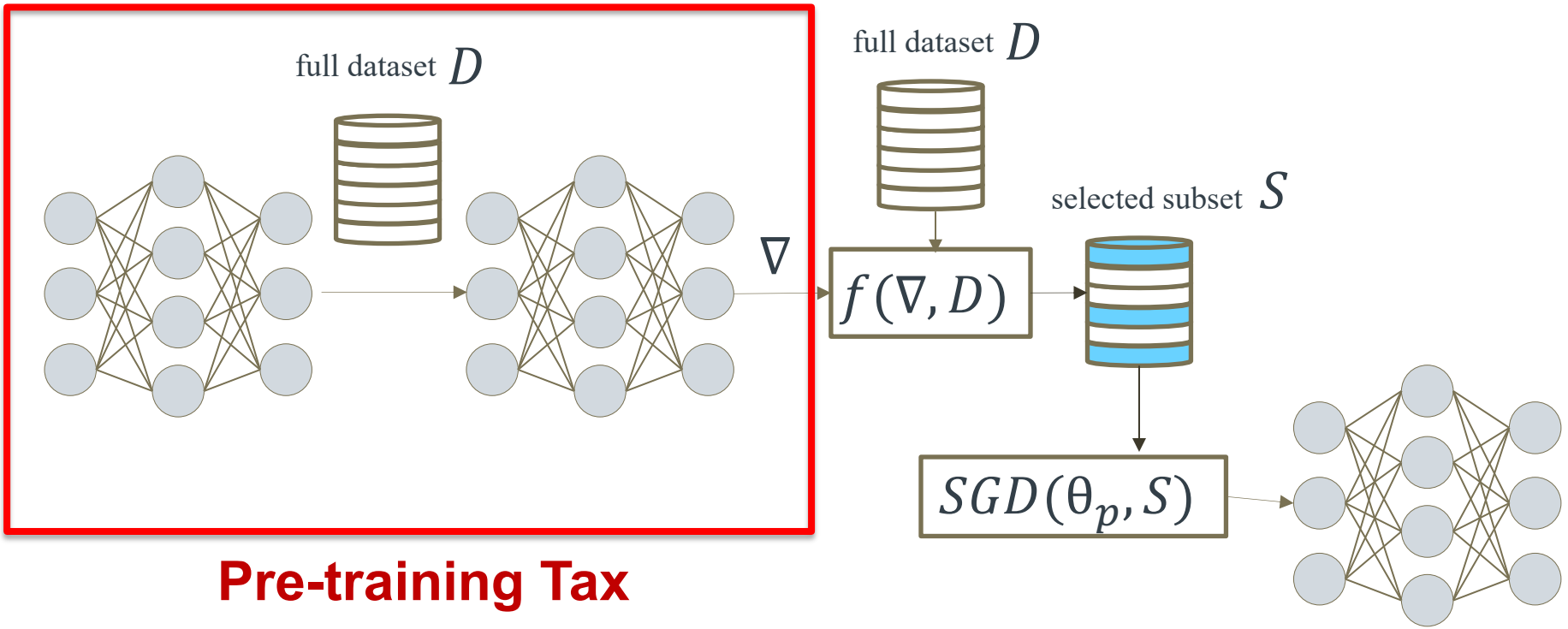
“Not all data samples are equally important”

- There are more important data and less important data in the dataset. Based on the importance value of each sample, we select a subset of the dataset and train a model with the selected subset with minimal loss of accuracy.
- With a subset training, we can reduce 1) training time, 2) memory requirement, 3) storage space.

Subset Training Pipeline and Pre-training Tax



Subset Training Pipeline and Pre-training Tax

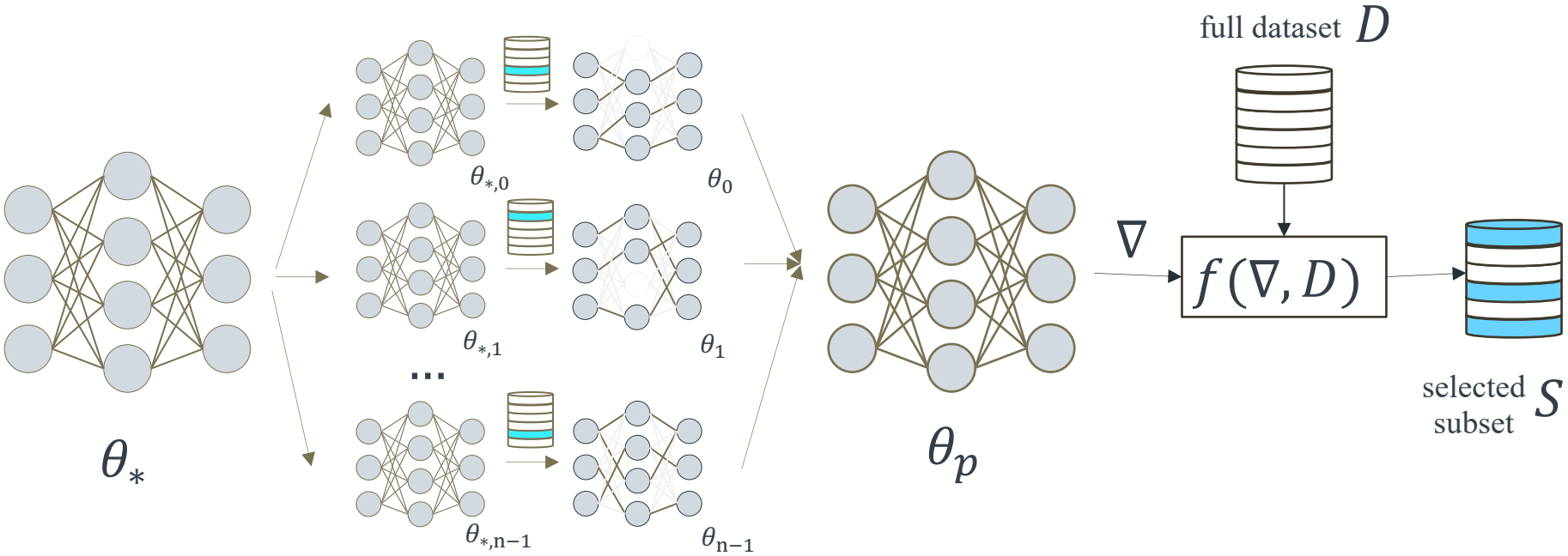


Pre-training Tax

Requirements for Distributed Pre-training

- Make it scalable so it can be run in a distributed environment with minimal communication costs.
  - Workers do not synchronize nor communicate during the pre-training.
  - Do not ship the full data to each worker to reduce communication costs and local training costs at each worker.
- Provide robust and reliable initial gradients for subset selection algorithms.

Distributed Lightweight Pre-training Toolkit



More details..

- We introduce an efficient pre-training framework with ensemble strategy by model averaging *at initialization.*
- To strengthen local worker training, we leverage **data-driven sparsity** as well as **aggressive data augmentation** to mitigate overfitting and boost ensemble diversity.
- Our result shows the proposed method reduces the variance and improves the final accuracy.

Tue 25 Jul 2PM-3:30PM
Exhibit Hall 1 #439