Egeria: Efficient DNN Training with Knowledge-Guided Layer Freezing

Yiding Wang, Decang Sun, Kai Chen (HKUST),
Fan Lai, Mosharaf Chowdhury (University of Michigan)

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

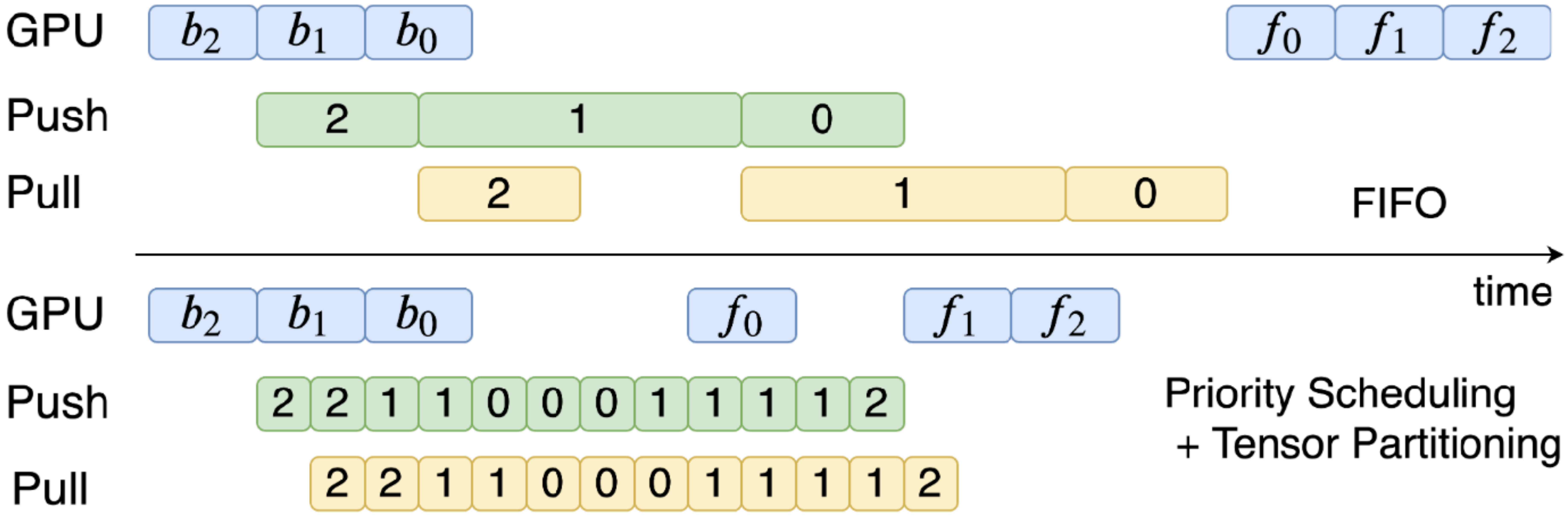
Motivation

- Large models (#layers and #parameters) and datasets make DNN training **time-consuming**.
- An iteration of data-parallel training over a mini-batch of dataset includes:
    1. Forward pass (light computation)
    2. **Backward pass** (heavy computation)
    3. Parameter synchronization (communication)

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Motivation

- **Communication scheduling** (ByteScheduler, SOSP '19). Theoretically optimal.



THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

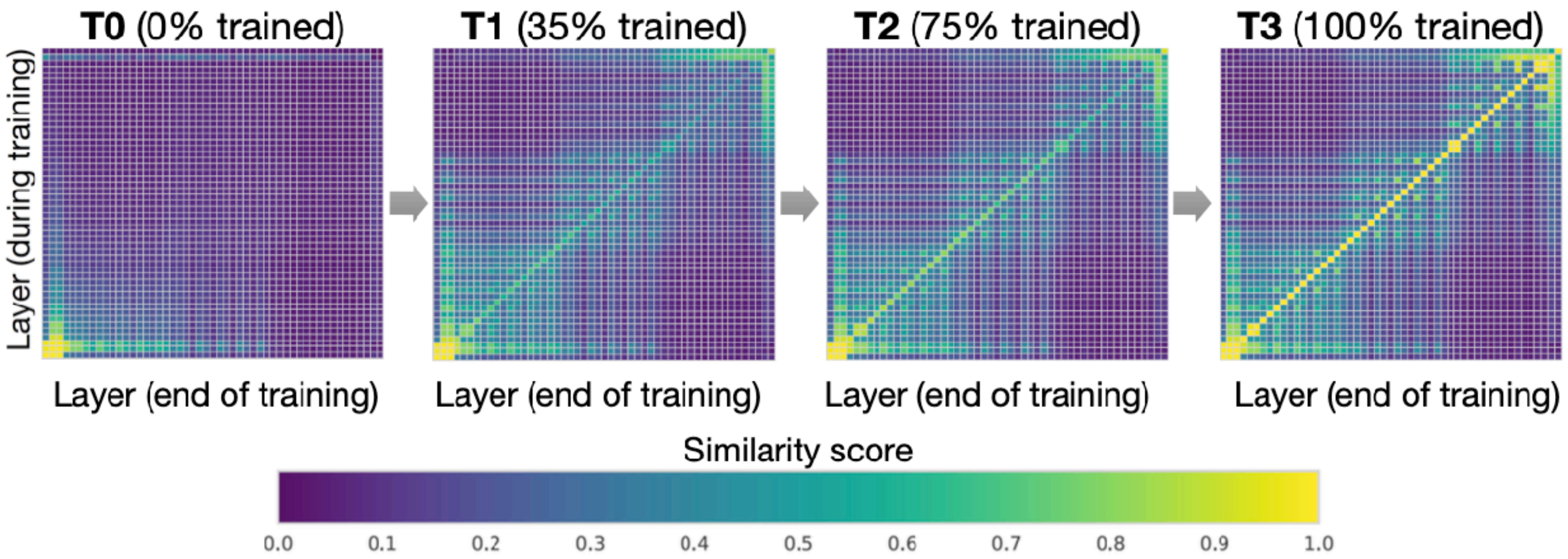
Motivation

- System optimizations (e.g., communication scheduling and pipeline parallelism) accelerate ML workload by making operations efficient (e.g., less idle GPU time).
- One more step: Can we further **reduce the ML workload** (with the same model quality) to accelerate training from the source?
- Reducing workload (lossy) should work together with existing (lossless) optimizations.

Motivation

- The training progress of DNN layers differs significantly: The front layers process general features and deep layers handle task-specific features.

- **The front layers converge earlier than deep layers:**

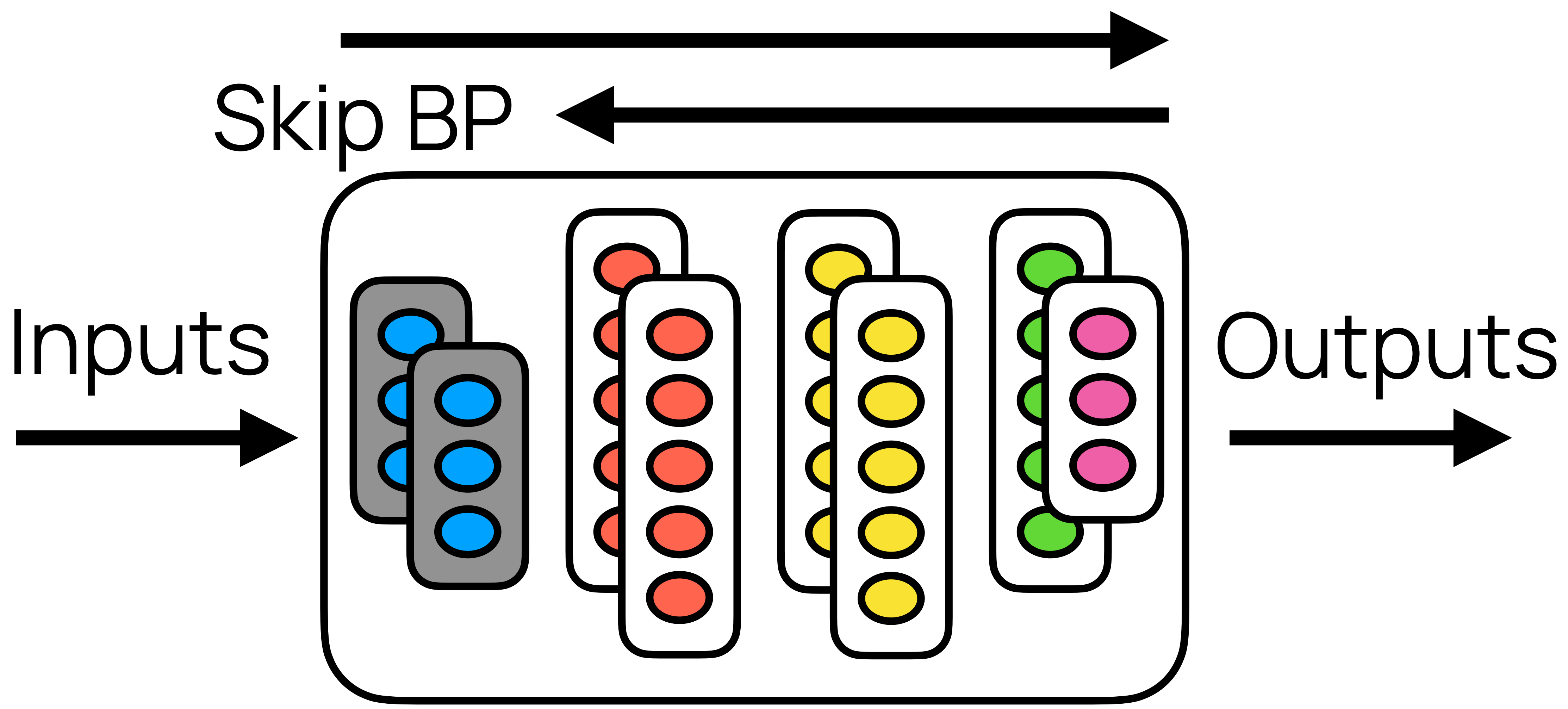


How transferable are features in deep neural networks? Ghahramani etc., NeurIPS '14. Image: PWCCA, Morcos etc.

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

A Potential Solution

- Intuition: We can **freeze the front layers** when they are converged, so that the **backward pass and parameter sync can be skipped**!



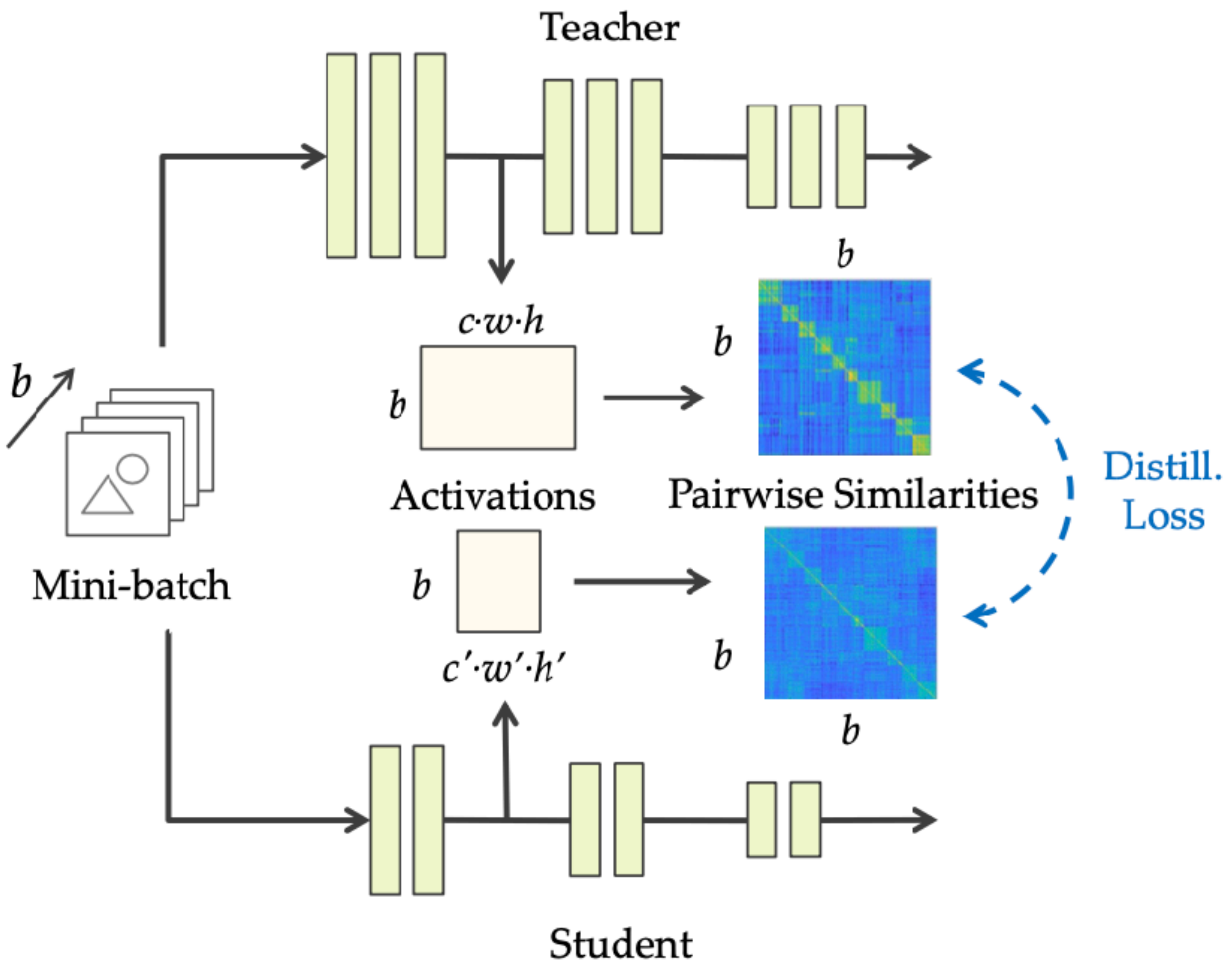
We tried a gradient-based method, but it causes accuracy loss.

- Challenge: How to accurately identify the freezable layers to **accelerate training while maintain accuracy**?

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Learning From ML Research

- We turn to knowledge distillation (KD).
- KD: Using the difference of **internal activation** of layers compared to a **trained teacher** to train a student model by minimizing the distill. loss.
- Hard label (gradients) is not enough.
- **Our goal is the same!** (understanding a layer's training progress)



Similarity-Preserving Knowledge Distillation Tung etc., ICCV '19

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

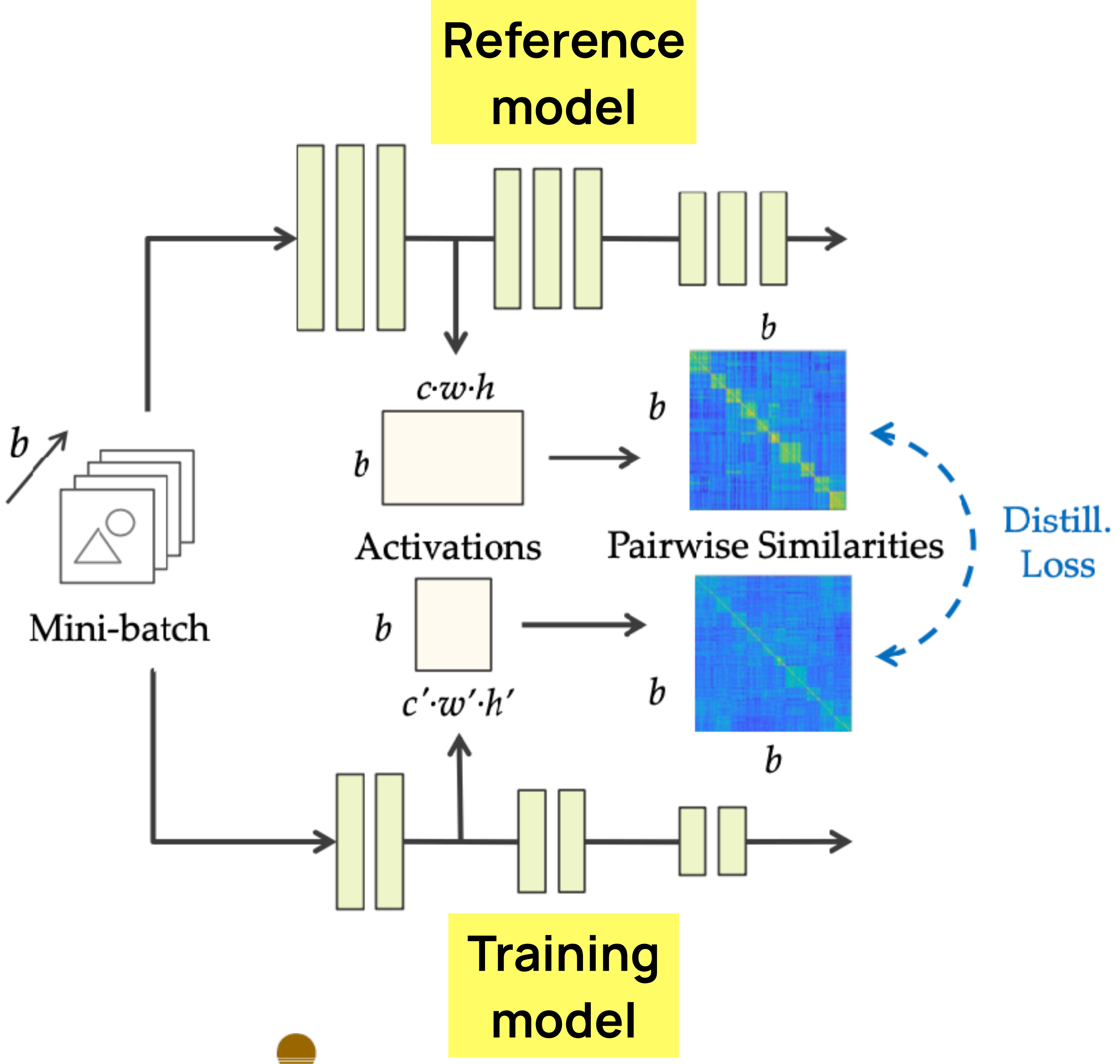
Using a Reference Model in Training

- Egeria compares the training model's layer activation to a **reference model** (a **snapshot** of the training model) to understand the progress!

- One KD loss used in ML work is Similarity Preserving (SP) loss.

- We define a system metric **plasticity** as the negative and normalized SP loss.

- Low plasticity over time → slow change.



THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Freezing Criteria & Hyperparameters

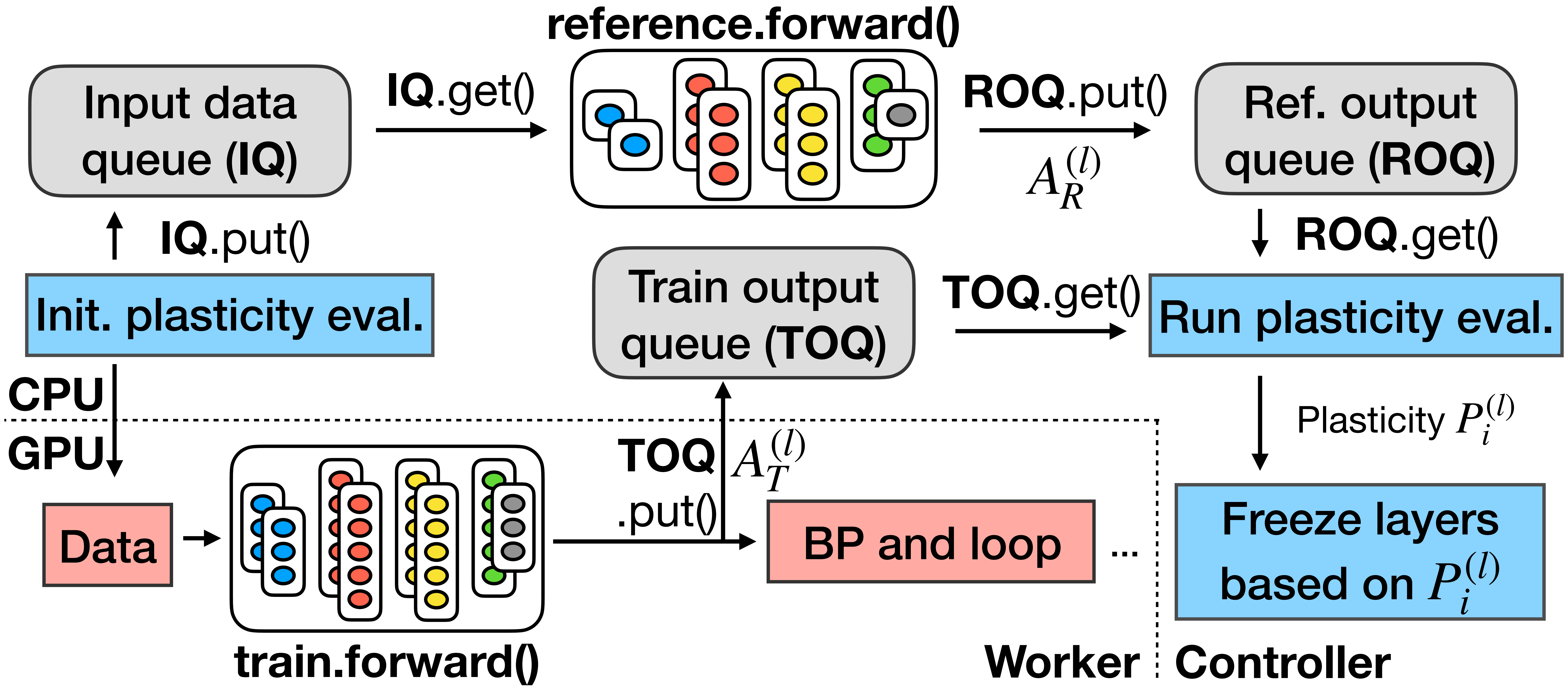
- Freezing layers is a lossy training acceleration technique: Misfreezing would hurt the DNN accuracy.
- We design an algorithm to analyze plasticity values and make freezing decisions.
- The intuition is that if some layers' plasticity is no longer changing, then we can freeze these layers and move on.
- Hyperparameters: Evaluation frequency, thresholds of how small and how long. We use human expertise to set them for now. Not strict.

Egeria Overview

- System efficiency: Reference model and control plan are on CPUs.
- Accuracy: Freezing a layer when confident. Unfreezing & re-freezing.

System Efficiency

- Avoid blocking GPU computation with **asynchronous execution**.



THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

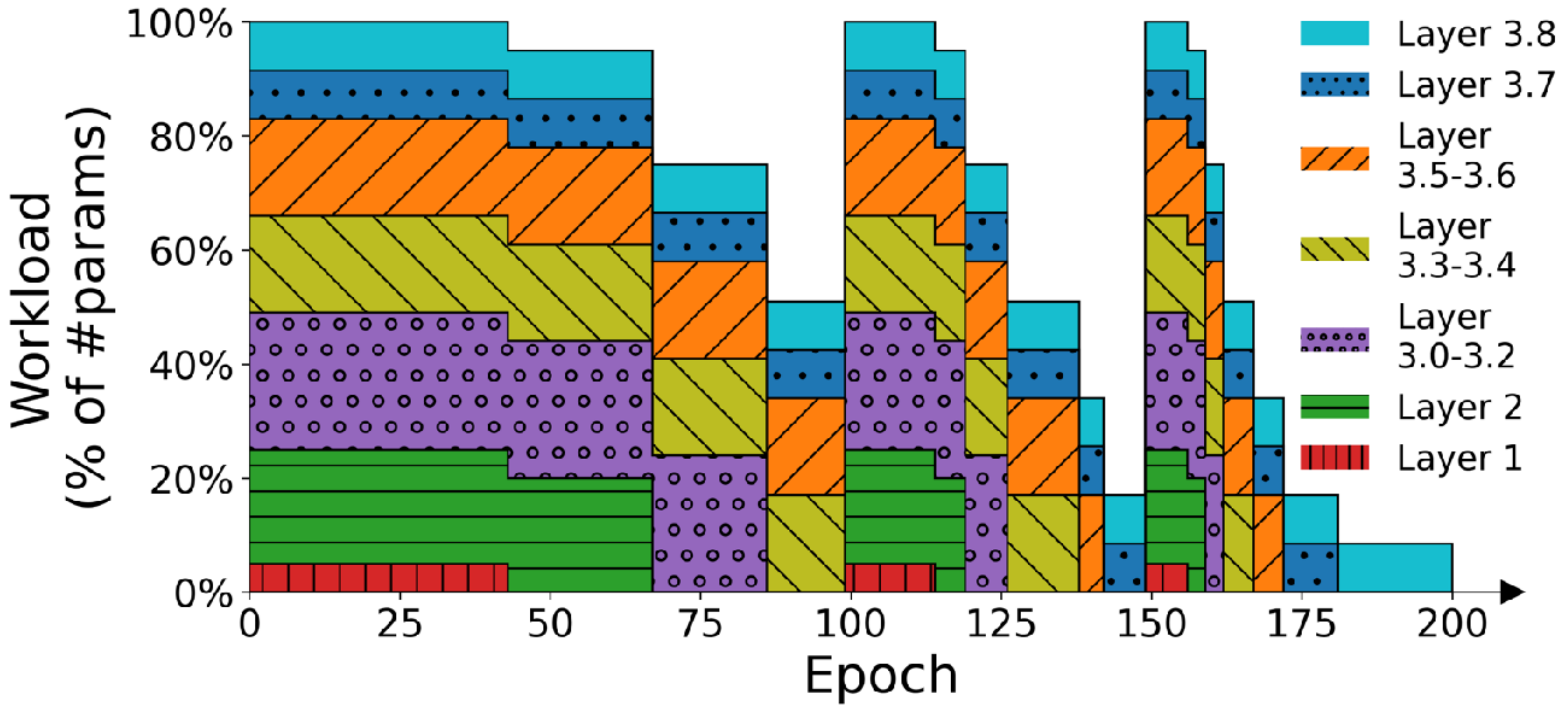
Evaluation: TTA Speedup

- Baseline: Naive PyTorch. Single GPU and data parallelism.

Task	Model	Dataset	Accuracy target	# Servers × # GPUs/server	# Building layer modules	TTA speedup
Image classification	ResNet-50 [127]	ImageNet [166]	Top 1 75.9%	1×2 2×2 - 5×2	48 (residual blocks)	28% 27%-33%
	MobileNet V2 [167]		71.2%	1×2	17 (inverted residual blocks)	22%
	ResNet-56 [127]	CIFAR-10 [168]	92.1%	1×2	54 (residual blocks)	23%
Semantic segmentation	DeepLabv3 [169]	VOC [170]	mIoU 63.3%	1×2	49 (residual blocks and DeepLab head)	21%
Machine translation	Transformer-Base [72]	WMT16 EN-DE [171]	Perplexity 4.7	4×2 2×2 - 5×2	12 (6 encoders & 6 decoders)	43% 33%-43%
	Transformer-Tiny		53.3	1×8	4 (2 & 2)	19%
Question answering	BERT-Base [12] (fine-tuning)	SQuAD 1.0 [172]	F1 score 87.6	1×2	12 (Transformer blocks)	41%

Evaluation: Freezing Breakdown

- How Egeria freezes and unfreezes layers during ResNet-56 training. Blanks are skipped layers.



THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Evaluation: Reference Model

- A quantized reference model is accurate enough and faster on the CPU.
- The time overhead of reference model is only 1.5% of the overall time.

Performance	int8	float16	float32
Final accuracy	92.1%	92.0%	92.2%
CPU inference speed	3.59×	1.69×	1×
Reference acc. gap	-0.6%	-0.2%	0

Table 2. Using difference precisions for the reference model. EGERIA hits the sweet spot between efficiency and accuracy.

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Thank you!

- Egeria **accurately freezes the converged layers** and saves their computation and communication costs.
- It uses knowledge distillation and transfer learning techniques.
- We propose the training plasticity metric to quantify layers' training progress since different layers converge differently during training.
- It accelerates DL training by 19%-43% without sacrificing accuracy.

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY