Pipe-BD: Pipelined Parallel Blockwise Distillation

Hongsun Jang, Jaewon Jung, Jaeyong Song, Joonsang Yu,
Youngsok Kim, and Jinho Lee

Accelerated Intelligent Systems Lab.

Introduction

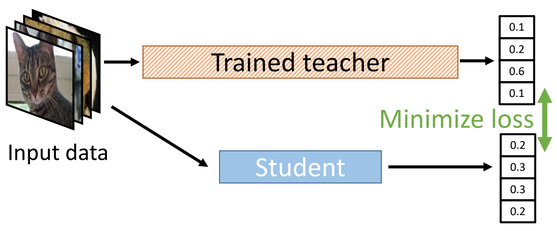
- The existing SOTA methods for parallelizing blockwise distillation
  -> Relying on traditional data-parallelism, exhibiting low resource utilization and redundancies caused by not fully exploiting independent nature of the blocks.

- We suggest novel parallel training method for blockwise distillation, **Pipe-BD**.
- **Pipe-BD** can automatically make all scheduling decisions for high throughput.

Background

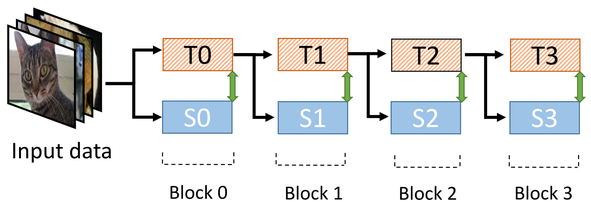
1. Knowledge Distillation (KD)

- KD trains student model with **pretrained** teacher's output.



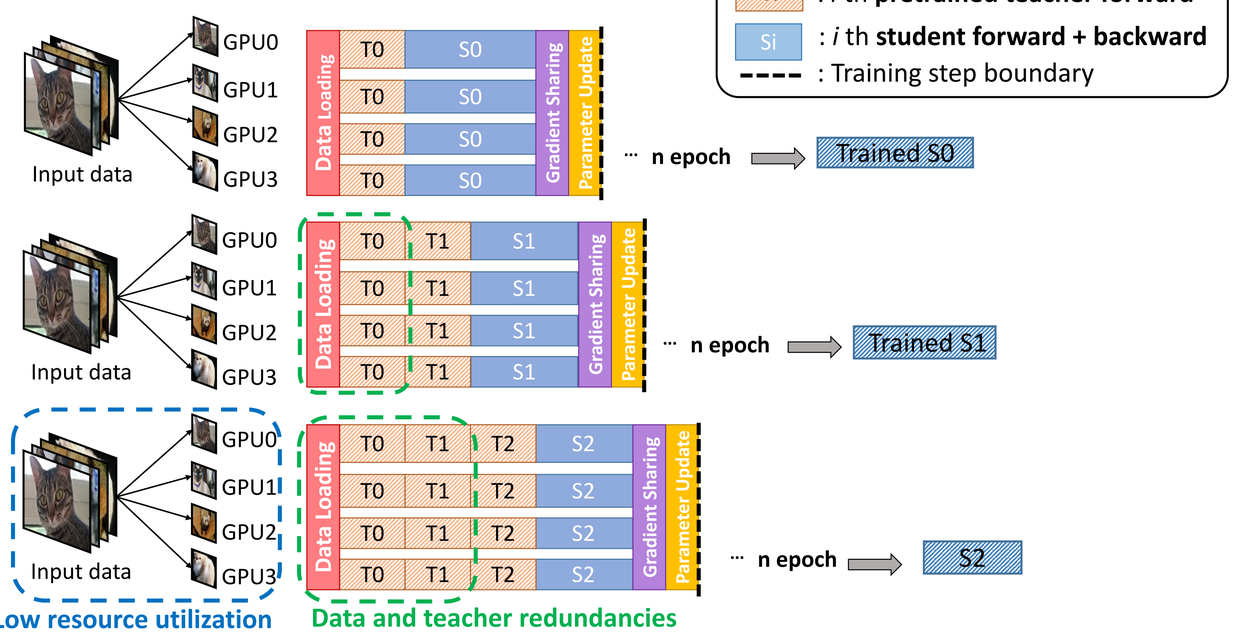
2. Blockwise Distillation (BD)

- BD splits models into smaller blocks and **uses KD in blockwise manner**.

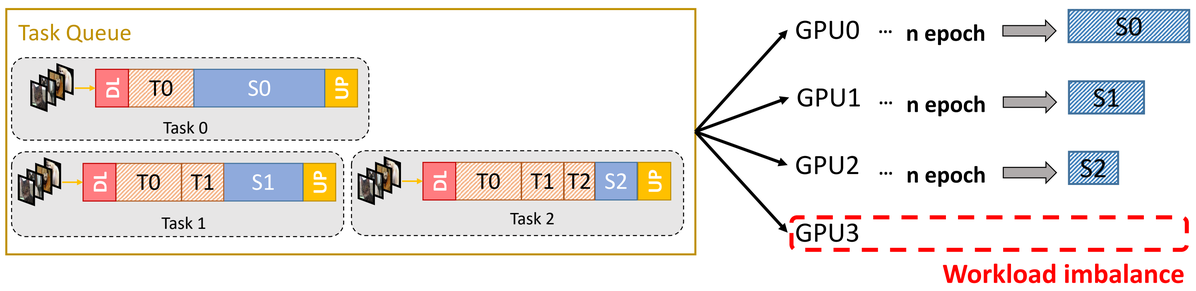


3. Baseline Parallelization of Blockwise Distillation

- Data parallelism (**DP**)



- Layer-wise parallelism (**LS**): Similar to task parallelism

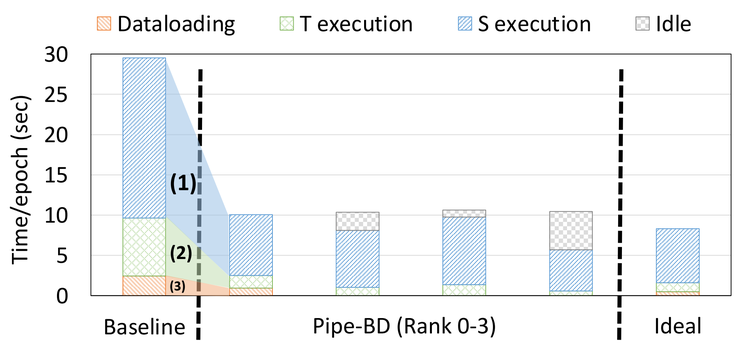


- Comparison table among baselines and Pipe-BD

	Teacher redundancy	Data redundancy	Low utilization	Workload imbalance	Scheduling decision
Baseline (DP)	O	O	O	X	X
Baseline (LS)	O	O	X	O	Bin packing, work stealing
Pipe-BD (TR)	O	△	X	O	Manual
Pipe-BD (DPU)	X	X	X	O	Manual
Pipe-BD (AHD)	**X**	**X**	**X**	△	**Auto**

Motivation

- Baseline approach (DP) suffers from three significant inefficiencies.
  (1) **Low resource utilization** (2) **Teacher redundancy** (3) **Data redundancy**

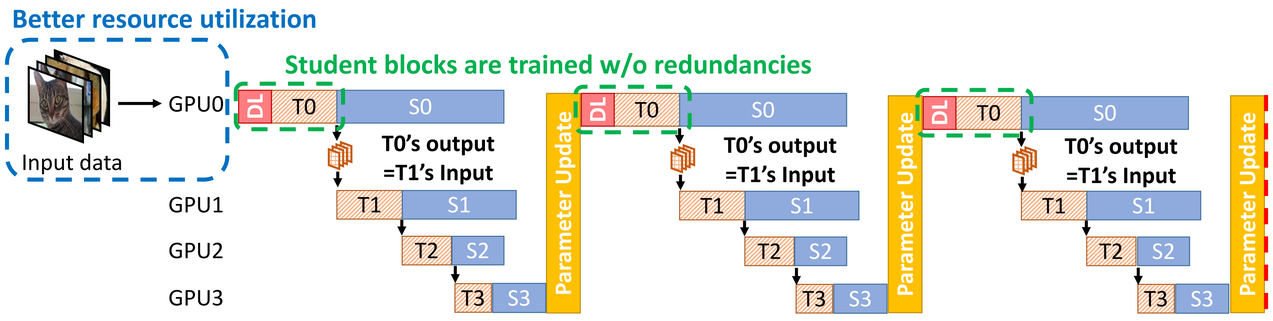


Pipe-BD:
Successfully handles (1), (2), (3) almost close to ideal case.

Pipe-BD

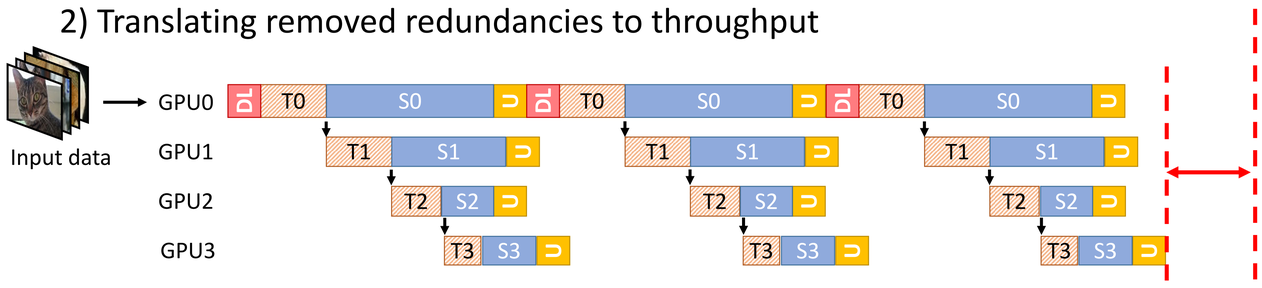
1. Teacher Relaying (TR)

- Teacher's intermediate activations are relayed between GPUs.
  1) Increasing GPU resource utilization
  2) Eliminating data redundancies and teacher redundancies



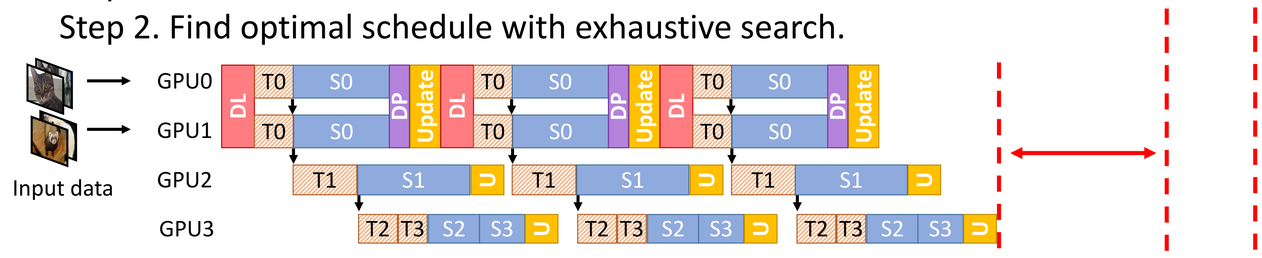
2. Decoupled Parameter Update (DPU)

- The parameters of student blocks are updated w/o waiting for other devices.
  1) Using **special characteristic of BD** (No weight dependency between blocks!)
  2) Translating removed redundancies to throughput



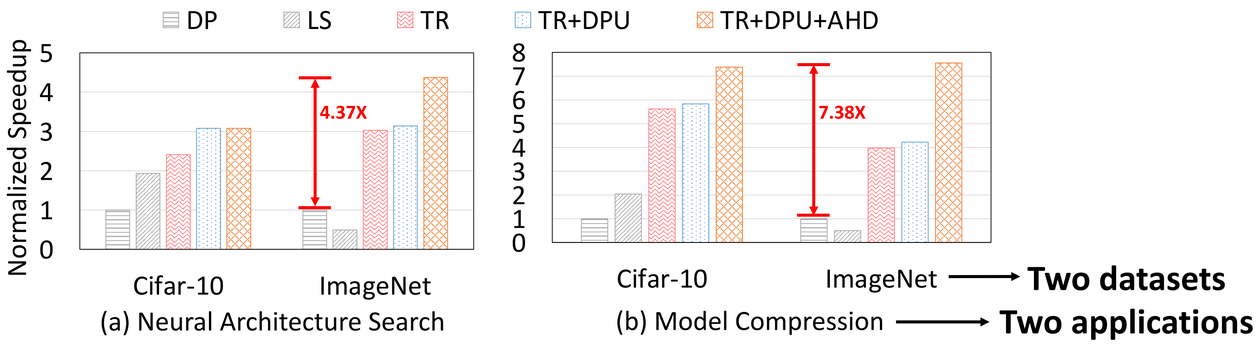
3. Automatic Hybrid Distribution (AHD)

- Batch-level degree of freedom is provided for workload scheduling.
- A schedule is automatically decided based on online profiling.
  Step 1. Measure each block execution time on feasible batch sizes.
  Step 2. Find optimal schedule with exhaustive search.

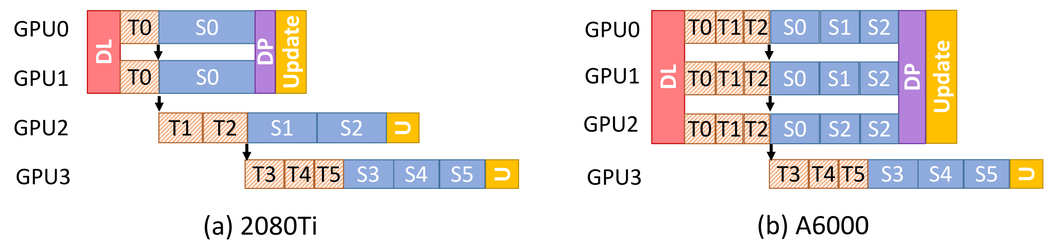


Evaluation

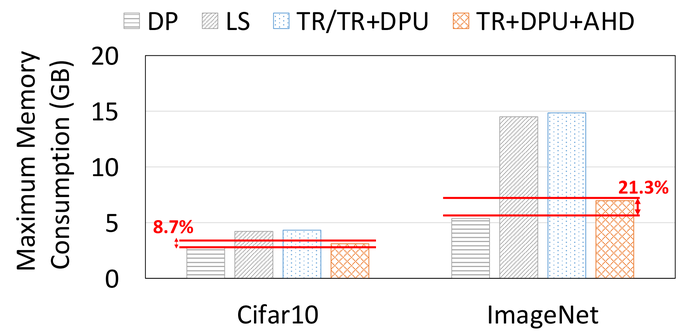
- Pipe-BD has overall speedup **2.37X to 7.38X** compared to baselines.



- Automatically **decided schedules** under two environments (2080Ti, A6000)



- **Memory overhead**
  Pipe-BD requires minor 8.7% and 21.3% additional memory overheads over baseline.



Conclusion

- **Pipe-BD** provides **multi-fold speedup without any modifying mathematical formulation of blockwise distillation**.
- We open-source Pipe-BD at https://github.com/hongsunjang/Pipe-BD.