Tensor-Parallelism with Partially Synchronized Activations

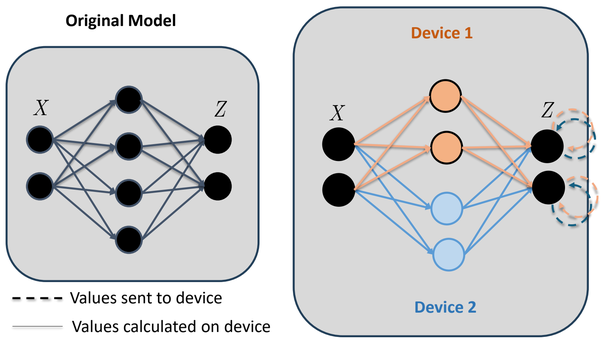
Itay Lamprecht†‡, Asaf Karnieli†, Yair Hanani†, Niv Giladi°* and Daniel Soudry‡

† Intel ‡ Technion ° AWS

Our Code

Background

Tensor-parallelism is a technique used in inference and training which splits model weights between multiple devices.



Tensor-parallel trade-off:

Inference latency ↓ Per device memory ↓ Communication overhead ↑

There have been many approaches to reduce tensor-parallel communication overhead:

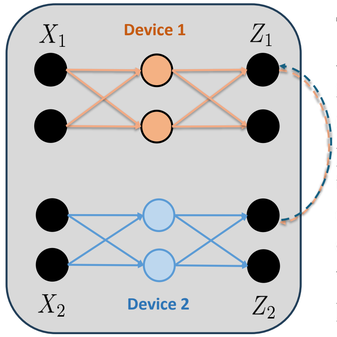
Communication computation overlapping	✗Not scalable ✗Not compatible with other parallelization methods
Compression	✗Accuracy degradation ✗Compression overhead
Inference time methods	✗Does not address training communication

Partial Synchronization

Our Insight

Activations don't need to be fully synchronized after communication.

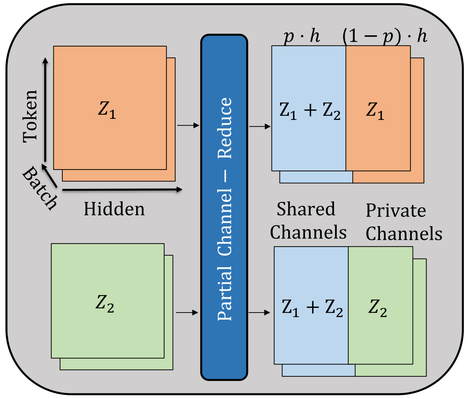
CAAT-Net



The Communication Aware Architecture for Tensor-parallelism (CAAT-Net) uses partial channel-reduce to overcome the communication overhead associated with tensor-parallelism.

Partial Channel-Reduce

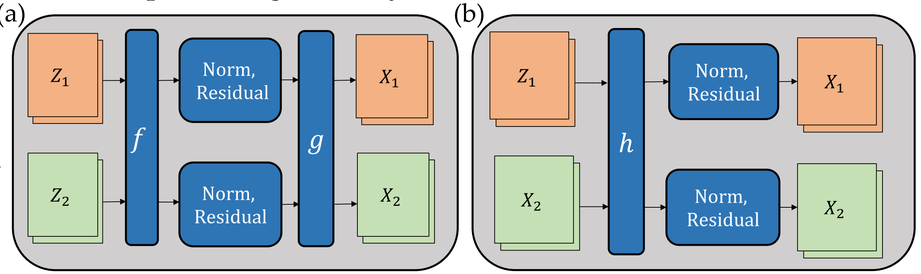
In partial channel-reduce, only a subset of the channels in the hidden dimension are synchronized, with synchronization factor p.



Implementing Partial Synchronization

In common training frameworks, partial synchronization leads to a forward-backward mismatch. To address this, the backward pass is modified.

Numerical stability is also assessed.



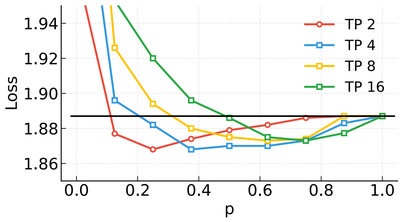
(a) Vanilla transformers in current training frameworks. The operation f is an all-reduce in the forward pass and identity in the backward pass. The operation g is an all-reduce in the backward pass and identity in the forward pass. (b) With partial synchronization, h denotes the reduction operation in both the forward and the backward pass.

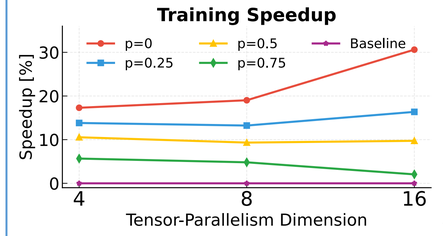
Results

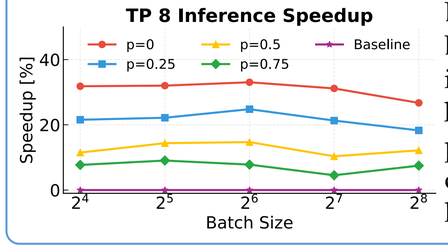
A CAAT-Net variant of Llama2-7b trained with tensor-parallel dimension 8 and $p = 0.5$ achieves results on par with the baseline.

Model	LAMBADA (acc)	Hellaswag (acc)	WinoGrande (acc)	PIQA (acc)
Baseline	**61.34 ± 0.68**	45.85 ± 0.50	61.48 ± 1.37	**72.91 ± 1.06**
CAAT-Net	61.05 ± 0.68	**46.10 ± 0.50**	**62.19 ± 1.36**	72.86 ± 1.04
	OpenBookQA (acc)	BOOL-Q (acc)	WikiText (ppl)	Validation Loss
Baseline	**26.60 ± 1.98**	**64.89 ± 0.83**	12.51	1.01
CAAT-Net	24.00 ± 1.87	62.51 ± 0.85	**12.46**	**1.00**

For a 130M parameter model, validation loss can improve for some values of p.







9% training speedup and 14% inference speedup achieved in the configuration used to train Llama2-7b. Better speedup is achieved for higher tensor-parallel dimension and lower p.

* Work was done while author was at Intel.