Layer-wise Decomposition of ResNets for Distributed Training

Chen Dun, Cameron Wolfe, Anastasios Kyrillidis
Rice University, Department of Computer Science

Motivation

Training deep learning models is expensive, but distributed training can make it faster.

Synchronous, distributed methods are widely used but introduce a major communication bottleneck because of frequent synchronization between machines.

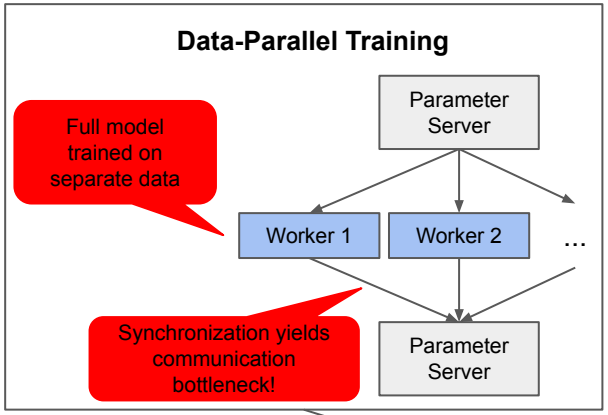
Residual Networks (ResNets) are robust to the removal of entire layers (Huang et. al 2016).

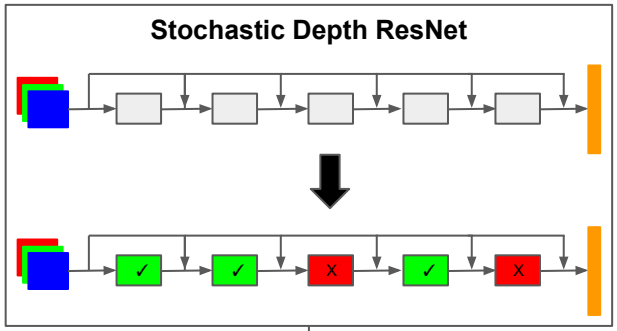
ResNets as Euler discretizations. Dropping layers yields a coarse approximation of the same continuous transformation (Lu et. al. 2020).

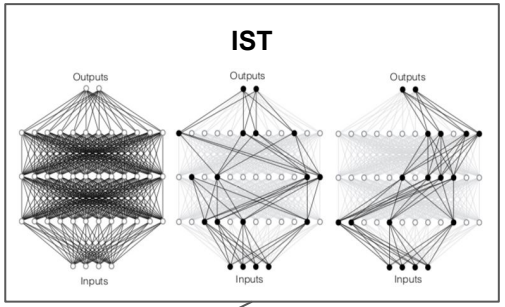
$$\mathbf{h}_{t+1} = \mathbf{h}_t + \Delta t f(\mathbf{h}_t, \theta_t)$$

Independent Subnet Training (IST) (Yuan et. al., 2020) has shown that independently-trained subnetworks can be combined to form a high-performing global model.

If layer-wise subnetworks can be constructed from ResNets, IST can be applied to form a simple and synchronous method of training ResNets with minimal communication.



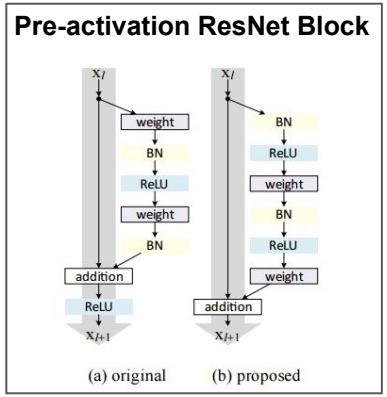


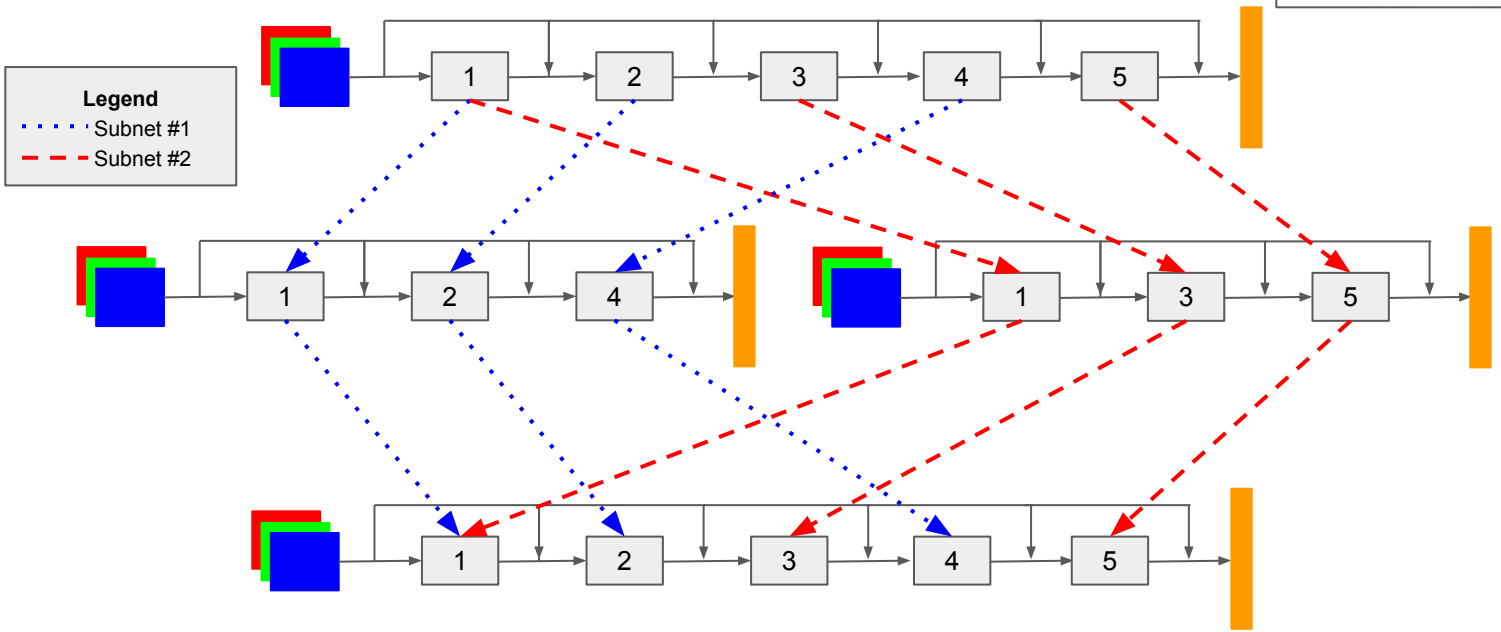


Communication-Efficient Synchronous Distributed Training for ResNets

Methodology

1. Randomly initialize a global ResNet (i.e., Pre-activation ResNet-101).
2. Use layer-wise decomposition of the ResNet to product N subnetworks.
3. Distribute each subnetwork to a separate machine.
   a. Only communicate parameters partitioned to a subnet!
4. Train subnetworks independently for several iterations.
5. Aggregate subnetwork parameters into the global ResNet.
   a. Weights shared between subnetworks are averaged.
6. Repeat the above steps until convergence.



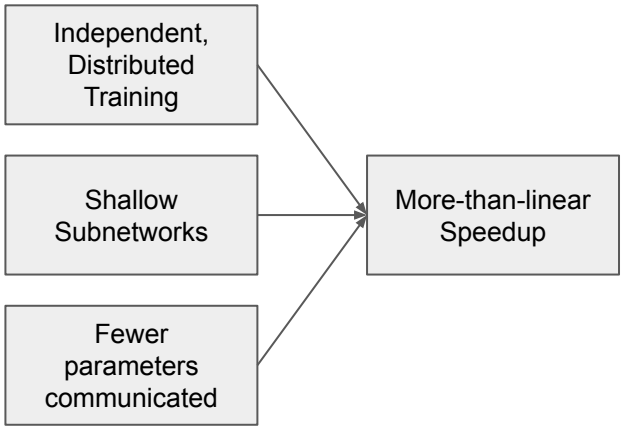


Results

The proposed methodology yields comparable performance for two, four, and eight machines, while significantly accelerating training.

Method	CIFAR10	CIFAR100	SVHN
Baseline	92.39%	71.01%	95.76%
2-Site	92.34%	70.68%	95.41%
4-Site	90.59%	69.06%	95.17%
8-Site	89.19%	68.93%	94.48%

Why does this speed up distributed training?



Analysis:

- Overlapping layers between subnets improves training with many subnetworks
- Pre-activation style ResNets work best with this formulation, as they stabilize input to each residual block.
- Residual blocks that lie within repeated chains of identical convolutions are easiest to partition.
- Strided convolutions or early convolutions within the network should not be partitioned, but shared between all sites.

Future work:

- Explore “smart” methods of partitioning ResNet layers to subnetworks.
- Explore adding stochasticity into the construction of each subnet.
- Explore the application of the proposed methodology within the non-IID federated learning setting.
- Develop theoretical justification for the effectiveness of this approach.