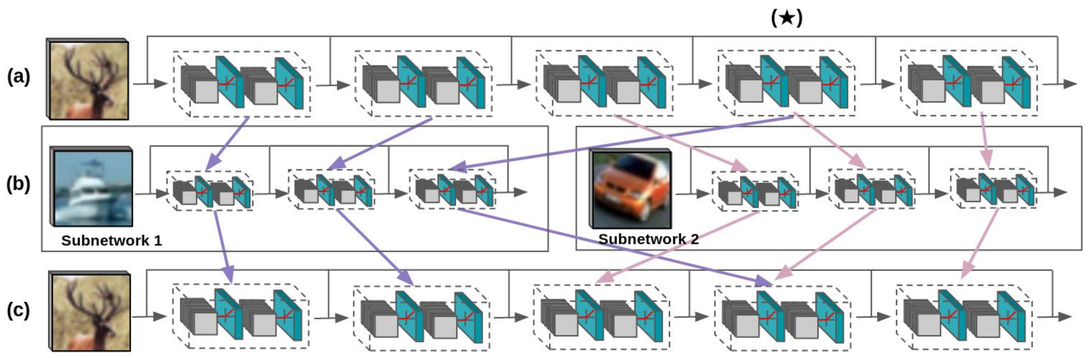
ResIST: Layer-wise Decomposition of ResNets for Distributed Training

Chen Dun, Cameron Wolfe, Christopher M. Jermaine, Anastasios Kyrillidis
Rice University, Department of Computer Science



Algorithm 1 ResIST Meta Algorithm

Parameters: T synchronization iterations, S sub-ResNets, ℓ local iterations, W ResNet weights.
$h(W) \leftarrow$ randomly initialized ResNet.
for $t = 0, \ldots, T-1$ **do**
  $\{h_s(W_s)\}_{s=1}^{S} = \texttt{subResNets}(h(W),\ S)$.
  Distribute each $h_s(W_s)$ to a different worker.
  for $s = 1, \ldots, S$ **do**
    Train $h_s(W_s)$ for ℓ iterations using local SGD.
  end for
  $h(W) = \texttt{aggregate}\left(\{h_s(W_s)\}_{s=1}^{S}\right)$.
end for

Motivation

Computer vision (CV) task is one of the most important real life scenarios for Federated learning, such as face recognition on mobile device and medical image classification on remote device. While deep ResNet[1] is the most successful model for CV, when applied in Federated Learning scenario, its large model size results in high communication cost, high computation cost and high memory cost on each edge device.

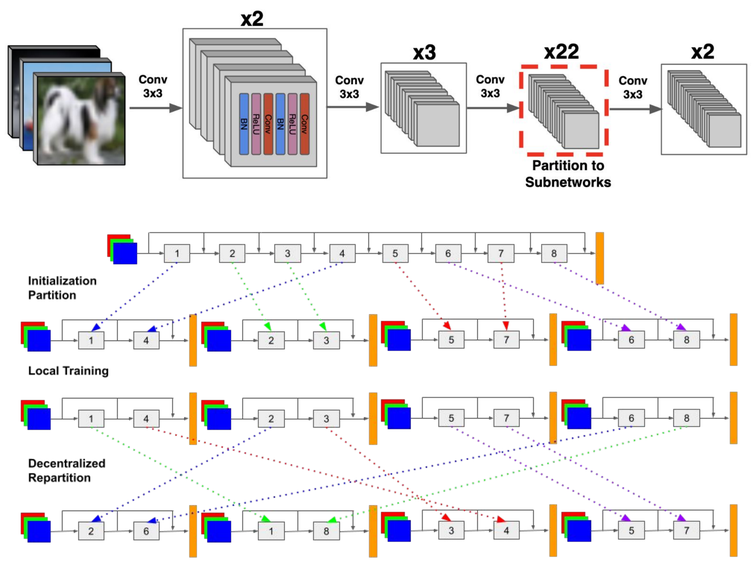
Contribution

We propose ResIST: A novel efficient Federated Learning algorithm for ResNet, which significantly reduces communication, computation and memory cost without losing accuracy.

	Data Parallel	Model Parallel	ResIST
Memory Cost	**High**	**Low**	**Low**
Comm Frequency	**Low**	**High**	**Low**
Comm Volume in each round	**High**	**Low**	**Low**
Computation Cost	**High**	**High**	**Low**

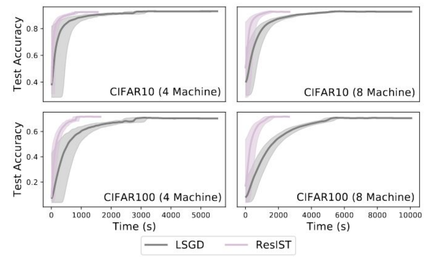
Method

ResIST randomly decomposes a global ResNet into several shallow sub-ResNets that are trained independently on each edge device for several local iterations, before having their updates synchronized and aggregated into the global model. In the next round, new sub-ResNets are randomly generated and the process repeats. By construction, per iteration, ResIST communicates only a small portion of network parameters to each machine and never uses the full model during training.



Result

ResIST is tested on Federated Learning scenario with (1) image classification datasets CIFAR10, CIFAR100 and Imagenet and (2) object detection dataset Pascal VOC. Local SGD (LSGD) is used as baseline.

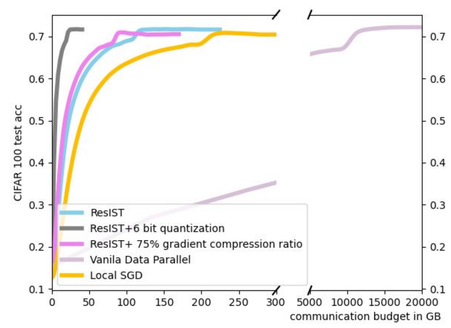


	# Machines	CIFAR10	CIFAR100
Local SGD	2	92.36% ± 0.01	70.67% ± 0.03
	4	92.90% ± 0.06	71.51% ± 0.04
	8	92.00% ± 0.07	69.64% ± 0.05
ResIST	2	91.95% ± 0.32	70.06% ± 0.51
	4	92.35% ± 0.22	71.30% ± 0.20
	8	91.45% ± 0.30	70.26% ± 0.21

	# Machines	Dataset	Total Time	Speedup
Local SGD	4	C10	5486 ± 7.05	-
		C100	5528 ± 65.90	-
	8	C10	10072 ± 5.12	-
		C100	10058 ± 8.71	-
ResIST	4	C10	1532 ± 0.83	**3.60**×
		C100	1545 ± 1.27	**3.58**×
	8	C10	2671 ± 3.25	**3.77**×
		C100	2639 ± 3.89	**3.81**×

	# Machines	Imagenet	Imagenet V2 Test Set MF	T-0.7	TI	Training Time	Speedup	Communication	Cost Ratio
Local SGD	2	73.32%	60.72%	69.47%	75.48%	48.61 hours	-	7546.80 GB	-
	4	72.66%	59.88%	68.34%	74.27%	29.29 hours	-	7546.80 GB	-
ResIST	2	71.60%	58.92%	67.51%	73.56%	36.79 hours	**1.32**×	5831.2 GB	**1.29**×
	4	70.74%	57.56%	66.46%	72.65%	22.37 hours	**1.31**×	6007.6 GB	**1.26**×

ResIST is also fully compatible with Quantization[2] and Sparsification[3] method in Federated Learning. When combined, ResIST can further reduces the communication cost.



[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
[2] Alistarh, D., Grubic, D., Li, J., Tomioka, R., & Vojnovic, M. (2017). QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, *30*, 1709-1720.
[3] Aji, A. F., & Heafield, K. (2017). Sparse communication for distributed gradient descent. *arXiv preprint arXiv:1704.05021*.