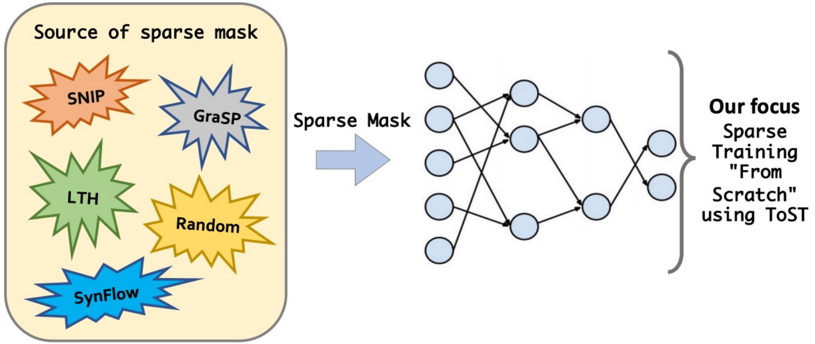
JUNE 2022

TRAINING YOUR SPARSE NEURAL NETWORK BETTER WITH ANY MASK

Ajay Jaiswal, Haoyu Ma, Tianlong Chen, Ying Ding, and Zhangyang Wang

AJAY JAISWAL
PhD Student, The University of Texas at Austin

Introduction



- DNNs are overparameterized, and recent research effort is focused on designing sophisticated pruning methods to yield high quality independently trainable sparse subnetworks.
- **Under-explored theme**: *improving training techniques for existing pruned sub-networks, i.e. sparse training.*
- **Big question:** Can we carefully customize the sparse training techniques to deviate from the default dense network training protocols?

Our contribution

A curated and easily adaptable training toolkit (ToST) for training ANY sparse mask from scratch:

- *"ghost" skip-connection (injecting additional non-existent skip-connections in the sparse masks),*
- *"ghost" soft neurons (changing the ReLU neurons into smoother activation functions such as Swish or Mish),*
- *as well as modifying initialization and labels.*

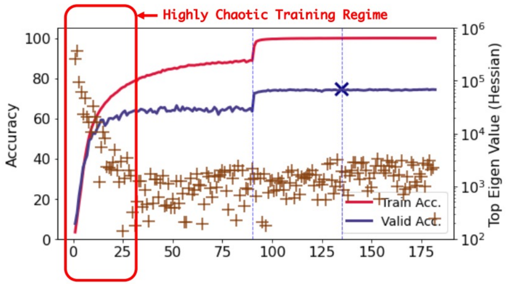


Figure 2. Top eigenvalues (Hessian) analysis of the training trajectory of a ResNet-18 sparse mask (90% sparsity) identified by LTH (Frankle & Carbin, 2018) using CIFAR-100.

Activation	Layer 1	Layer 2	Layer 3	Layer 4
ReLU	27.14%	39.33%	39.48%	57.93%
Swish	0.31%	0.26%	0.24%	0.20%
Mish	1.09%	1.14%	1.03%	0.95%

Table 1. Layer-wise Activation sparsity of ResNet-18 sparse mask (90% sparsity) identified by LTH (Frankle & Carbin, 2018) and trained with CIFAR-100.

Our Toolkit (ToST)

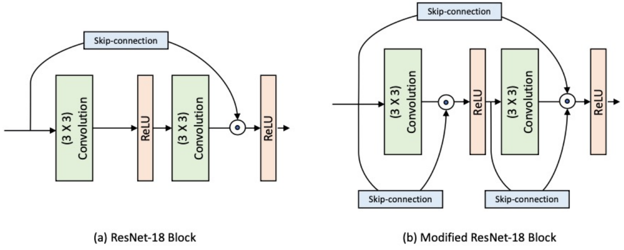


Figure 4. Our modified ResNet-18 block to introduce additional "ghost" skip-connections for the initial stage of sparse training.

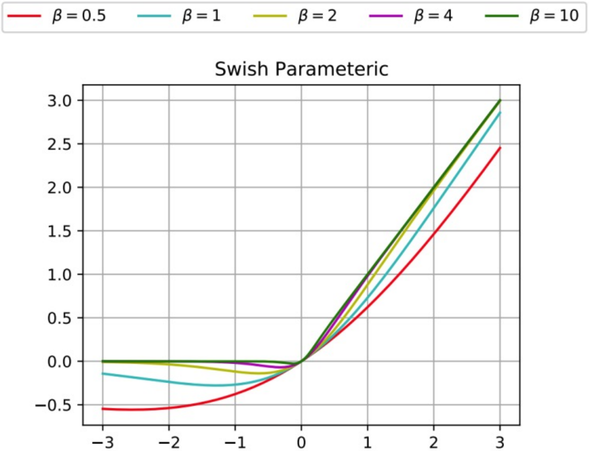


Figure 5. PSwish Visualization with different β values.

Ghost Swish (GSw), we gradually increase the β value of GSw, leading to be alike ReLU.

Label Smoothening

$$L_{\mathrm{LS}} = -\sum_{k=1}^{K} y_k \log(p_k)$$

$$y_k^{LS} = y_k(1-\alpha) + \alpha/K$$

Ghost Skips (GSk), we introduced gate functions regulated by a hyperparameter α, which controls the contribution of GSk during the training.

Layer-wise Re-scaled initialization (LRsI): Balance between random re-initialization of sparse subnetworks and directly copying the default dense initialization. LRsI keep original initialization of sparse masks intact for each parameter block and just re-scaled it by a learned scalar coefficient.

Experimental Results

Sparse Mask	CIFAR-10			CIFAR-100		
	90%	95%	98%	90%	95%	98%
ResNet-32 [No Pruning]	94.80	-	-	74.64	-	-
Random Pruning	89.95±0.23	89.68±0.15	86.13±0.25	63.13±2.94	64.55±0.32	19.83±3.21
Random Pruning + ToST	**91.53±0.11**	**91.44±1.01**	**88.20±0.89**	**65.19±1.36**	**64.61±1.21**	**33.98±6.64**
SNIP (Lee et al., 2018)	92.26±0.32	91.18±0.17	87.78±0.16	69.31±0.52	65.63±0.15	55.70±1.13
SNIP + ToST	**92.83±0.15**	**92.01±0.21**	**88.12±0.13**	**70.00±0.09**	**68.46±0.62**	**60.21±1.96**
GraSP (Wang et al., 2020)	92.20±0.31	91.39±0.25	88.70±0.42	69.24±0.24	66.50±0.11	58.43±0.43
GraSP + ToST	**92.98±0.07**	**92.77±0.14**	**89.92±0.56**	**70.18±0.22**	**67.20±0.74**	**62.30±1.06**
SynFlow (Tanaka et al., 2020)	92.01±0.22	91.67±0.17	88.10±0.25	69.03±0.20	65.23±0.31	58.73±0.30
SynFlow + ToST	**93.39±0.59**	**92.06±0.32**	**91.82±0.73**	**70.25±0.06**	**67.90±1.22**	**61.72±0.84**
LTH (Frankle & Carbin, 2018)	93.14±0.30	92.98±0.12	92.22±0.61	71.11±0.57	70.37±0.19	69.02±0.22
LTH + ToST	**94.01±0.23**	**93.60±0.70**	**93.34±1.06**	**72.30±0.61**	**71.99±0.95**	**70.22±0.61**
ResNet-50 [No Pruning]	94.90	-	-	74.91	-	-
Random Pruning	85.11±4.51	88.76±0.21	85.32±0.47	65.67±0.57	60.23±2.21	28.32±10.35
Random Pruning + ToST	**92.73±0.22**	**90.95±1.22**	**87.11±2.21**	**67.75±1.32**	**63.60±0.11**	**41.99±4.51**
SNIP (Lee et al., 2018)	91.95±0.13	92.12±0.34	89.26±0.23	70.43±0.43	67.85±1.02	60.38±0.78
SNIP + ToST	**92.89±0.53**	**92.56±0.12**	**90.56±0.19**	**70.79±0.22**	**68.06±0.09**	**61.51±1.41**
GraSP (Wang et al., 2020)	92.10±0.21	91.74±0.35	89.97±0.25	70.53±0.32	67.84±0.25	63.88±0.45
GraSP + ToST	**92.64±0.17**	**92.33±0.09**	**90.94±0.35**	**70.89±0.21**	**68.09±0.12**	**65.01±0.33**
SynFlow (Tanaka et al., 2020)	92.05±0.20	91.83±0.23	89.61±0.17	70.43±0.30	67.95±0.22	63.95±0.11
SynFlow +ToST	**92.55±0.10**	**92.57±0.18**	**90.27±0.29**	**70.86±0.21**	**68.83±0.15**	**65.40±0.13**
LTH (Frankle & Carbin, 2018)	93.69±0.31	93.18±0.17	92.79±0.14	71.89±0.11	71.05±0.13	70.41±0.28
LTH + ToST	**94.37±0.06**	**94.01±0.32**	**92.94±0.21**	**73.69±0.13**	**72.20±0.15**	**71.93±0.34**

Table 2. Classification accuracies of various pruning algorithm for varying sparsities $s \in \{90\%, 95\%, 98\%\}$ and network architectures (ResNet-18 and 32) with and without our sparse training toolkit (ToST).

Algorithm	**85%**	**90%**	**95%**
SNIP (Lee et al., 2018)	58.91±0.23	56.15±0.31	51.19±0.47
SNIP + ToST	59.44±0.09	57.19±0.21	53.21±0.08
LTH (Frankle & Carbin, 2018)	60.11±0.13	58.46±0.17	53.19±0.31
LTH + ToST	61.52±0.32	58.96±0.08	54.76±0.22

Table 3. Classification accuracies on TinyImageNet for varying sparsities $s \in \{90\%, 95\%, 98\%\}$ using ResNet-50.

Experimental Results

Method	**75%**	**80%**	**85%**	**90%**	**95%**
LTH (Frankle & Carbin, 2018)	73.21±0.17	72.94±0.12	71.91±0.22	71.12±0.30	69.57±0.19
LTH + GSk	73.77±0.11	**73.69±0.25**	72.86±0.30	**72.17±0.23**	**71.72±0.22**
LTH + GSw	73.45±0.13	73.22±0.43	**73.27±0.31**	72.03±0.12	70.85±0.52
LTH + LRsI	**73.93±0.15**	73.12±0.13	72.30±0.19	71.83±0.32	69.98±0.29
LTH + LS	73.58±0.28	73.70±0.32	72.65±0.25	71.93±0.20	70.19±0.14
LTH + ToST	74.29±0.31	74.03±0.14	73.90±0.49	73.23±0.27	72.08±0.10

Table 4. Breakdown of the performance of individual tweaks in ToST tweaks when applied on training ResNet-18 sparse masks (LTH) with varying sparsities $s \in \{75\%, 80\%, 85\%, 90\%, 95\%\}$ and trained on CIFAR-100.

	Dense NN (0%)	**20%**	**75%**	**95%**
“GSk”	-0.77%	+0.03%	+0.56%	+2.15%
“GSw”	+0.11%	+0.29%	+0.24%	+1.28%

Table 5. Performance benefit of “GSk” and “GSW” when applied to dense networks (0%) sparsity, low sparsity (20%), mid-level sparsity (75%), and high sparsity (95%). We have used LTH sparse mask of ResNet-18 trained on CIFAR-100.

Experimental Results

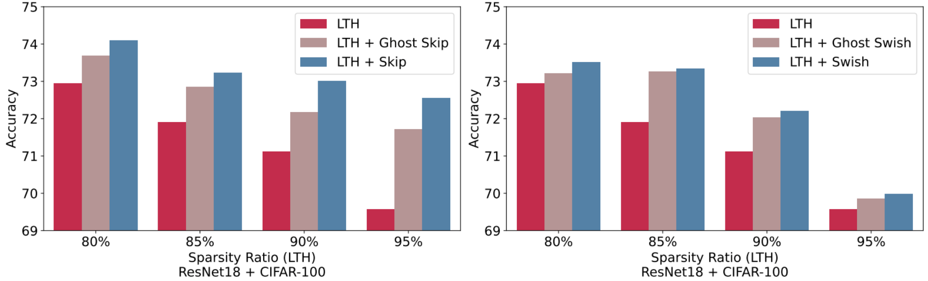


Figure 6. Performance comparison of the "Ghostiliness" behaviour of GSk and GSw with the default prolonged injection of swish and skip connections for LTH sparse masks with varying sparsities $s \in \{80\%, 85\%, 90\%, 95\%\}$.

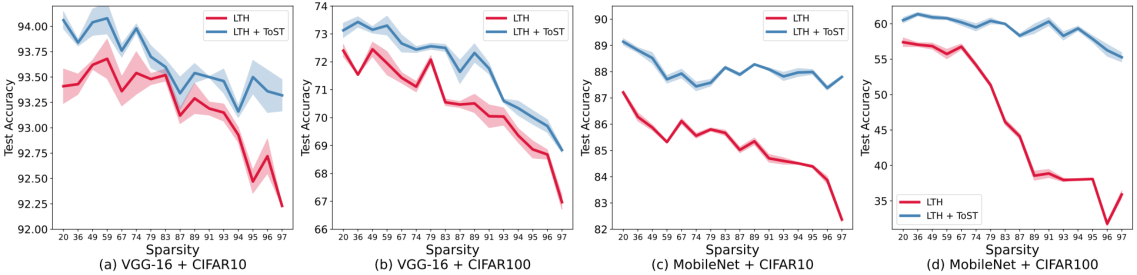


Figure 7. Performance comparison of sparse masks by LTH at varying sparsities $s \in [20\% - 97\%]$ on CIFAR-10 and CIFAR-100.

Thank you!