Sparse Winning Tickets are Data-Efficient Image Recognizers

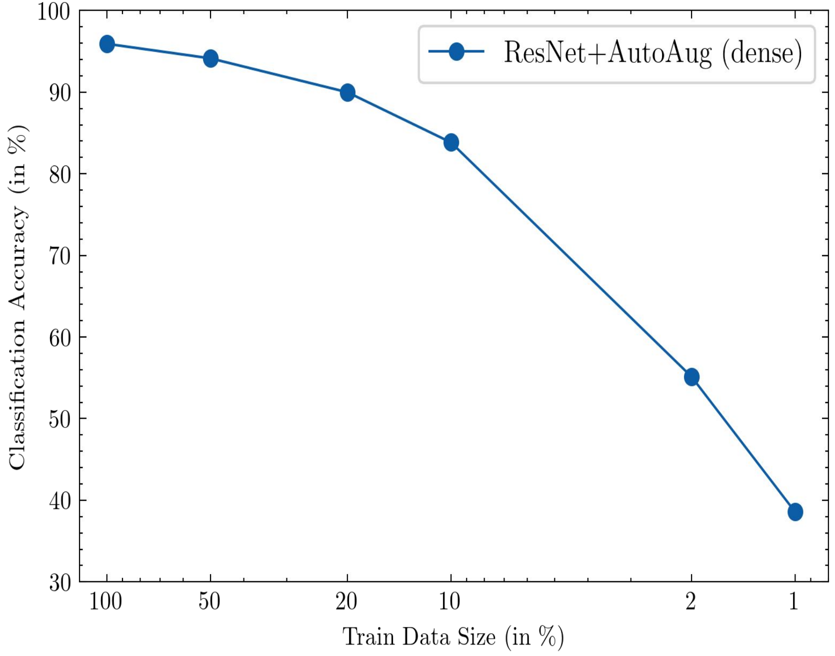
Mukund Varma T[1], Xuxi Chen[2], Zhenyu Zhang[2], Tianlong Chen[2], Subhashini Venugopalan[3], Zhangyang Wang[2]

[1]Indian Institute of Technology Madras, [2]University of Texas at Austin, [3]Google Research

Deep Neural Networks are ‘Data Hungry’

DNNs require **large amount** of training data to perform well.

However, there exists several domains and tasks where training data - labelled or unlabeled - is **limited**.



10,000 imgs/class → 50 imgs/class

Can Sparsity Help?

Sparsity as a regularization can reduce overfitting.

Lottery Tickets identified by Iterative Magnitude Pruning (IMP) induces an inductive bias specific to the task to be learned.

IMP reduces sample complexity.

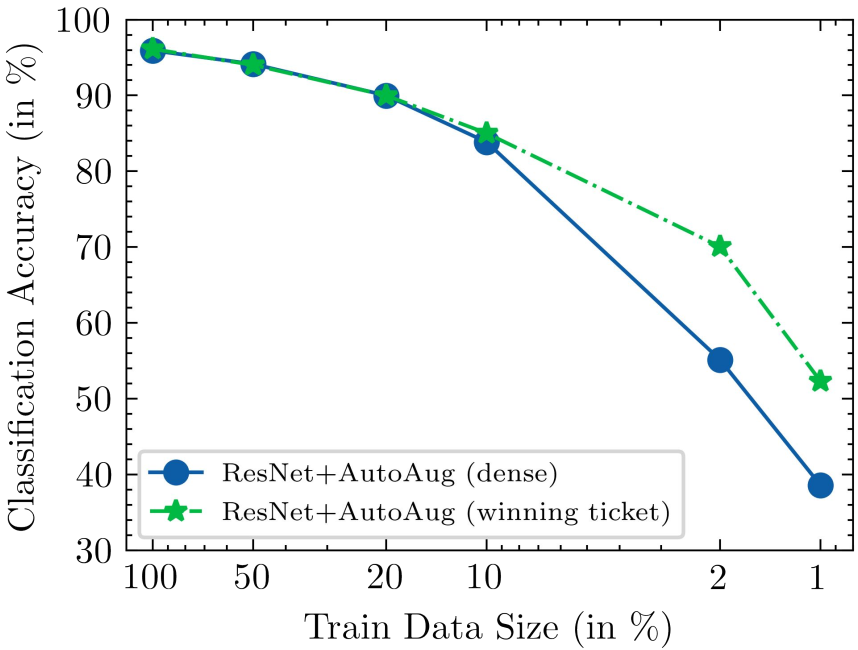
Can Sparsity Help?

Sparsity as a regularization can reduce overfitting. *(Shalev et al. '14)*

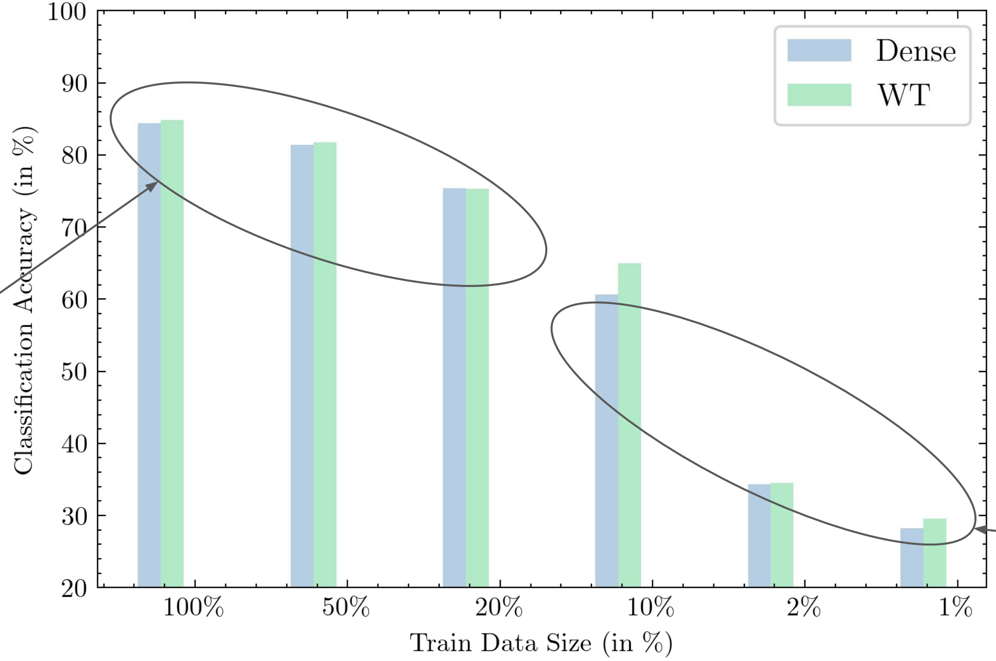
Lottery Tickets identified by Iterative Magnitude Pruning (IMP) induces an inductive bias specific to the task to be learned. *(Pellegrini et al. '21)*

IMP reduces sample complexity. *(Zhang et al. '21)*

Yes!



Sparse Winning Tickets show 'improved' performance in low-data regimes

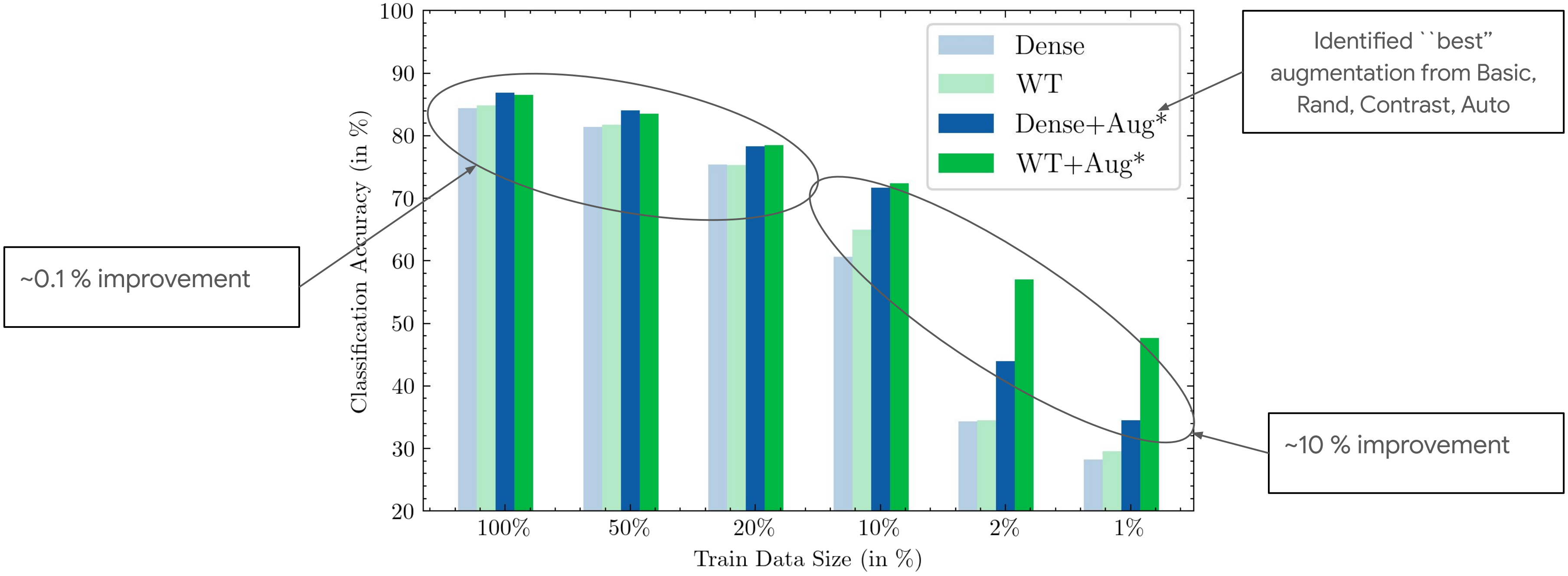


~0.2 % improvement

~0.7 % improvement

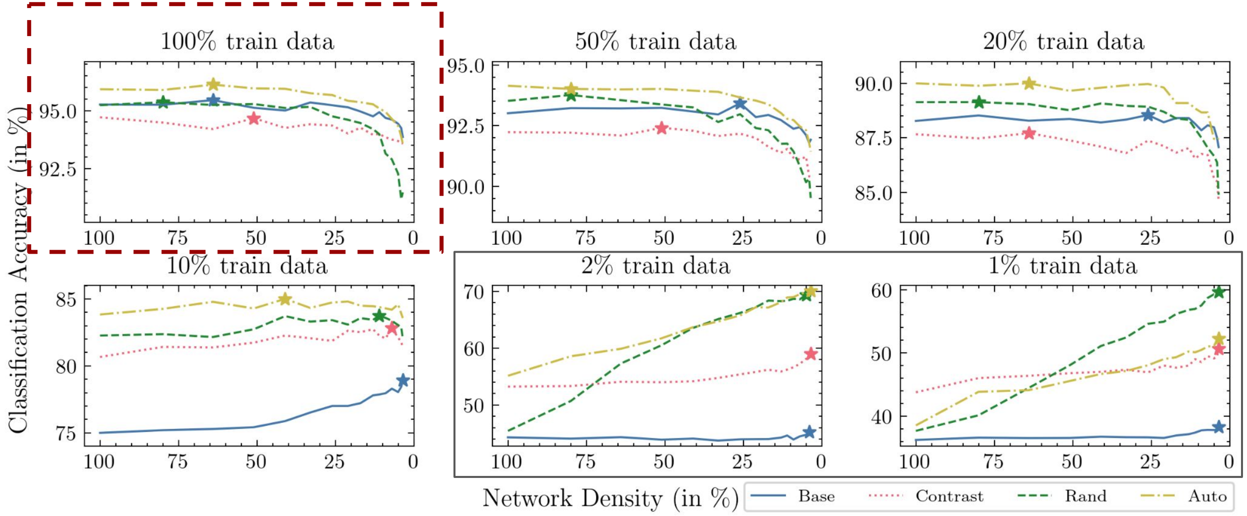
with Augmentations

Sparse Winning Tickets ^ show 'superior' performance in low-data regimes

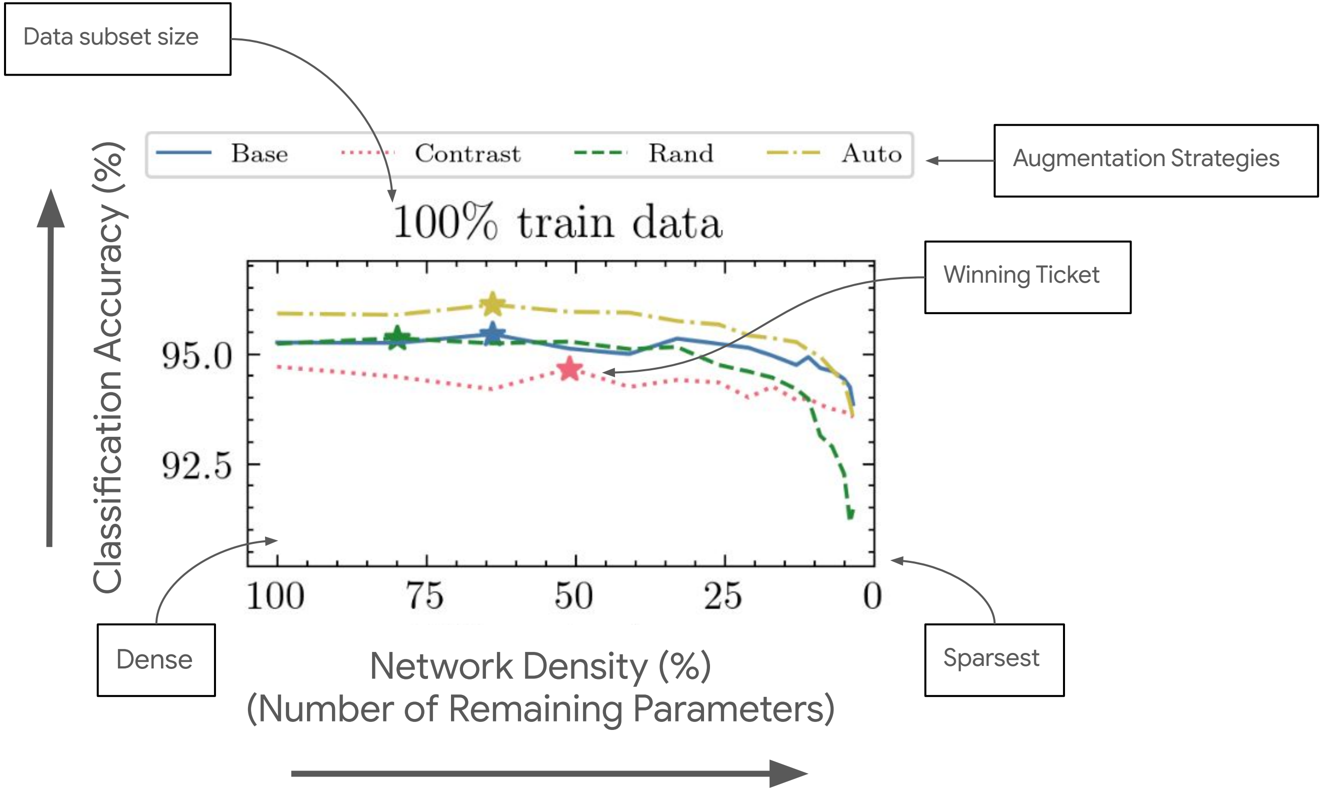


Note: Augmentation substantially improves performance of pruned networks. However, just augmenting does not help and it's the combination that yields significantly better results.

Sparse Winning Tickets show ‘superior’ performance in low-data regimes



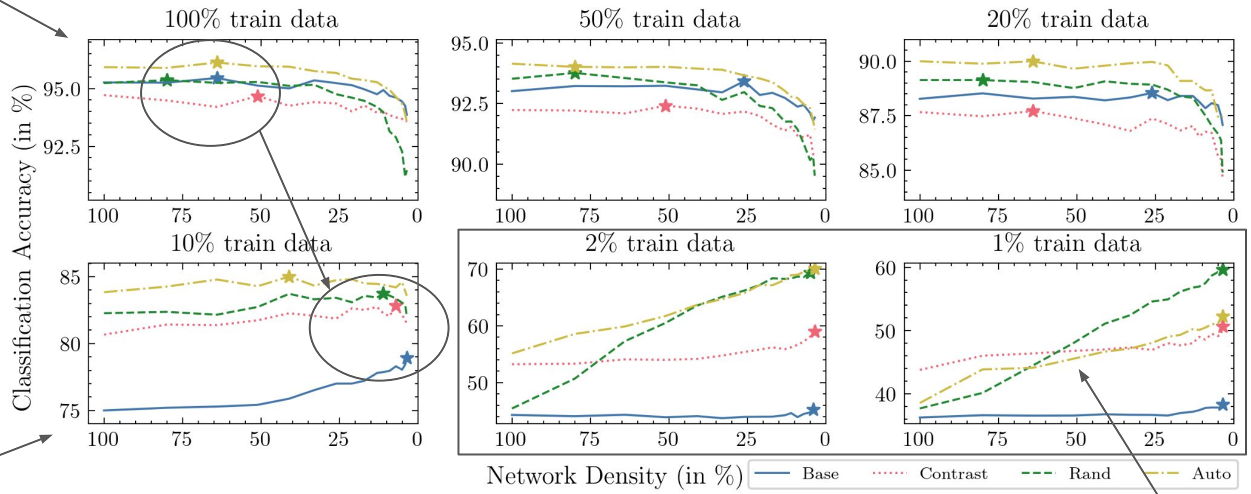
Sparse Winning Tickets show ‘superior’ performance in low-data regimes



Sparse Winning Tickets show 'superior' performance in low-data regimes

At higher data sizes, WT exist at ~(75-50%) density

At lower data sizes, WT exist at ~(10-3%) density

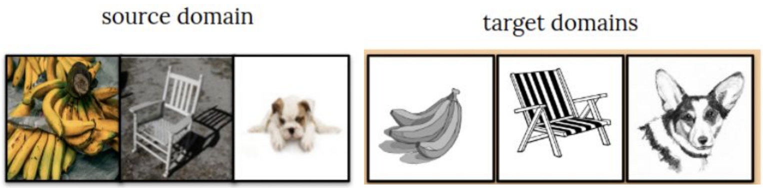


At least data sizes, strictly increasing trend between network performance and density

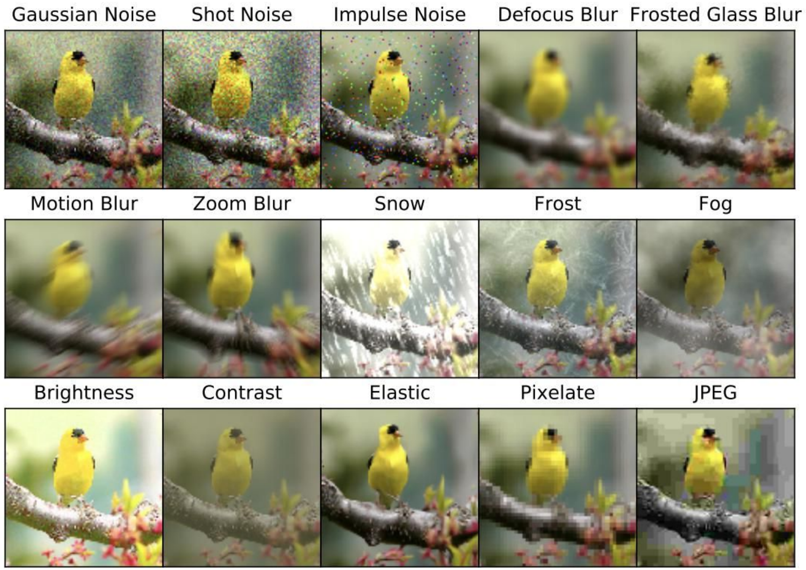
Sparse Winning Tickets are robust to distributional shifts

Does sparsity reduce overfitting by avoiding memorization of training samples?

Evaluate robustness of networks on unseen distributional shifts.



Domain Shifted*



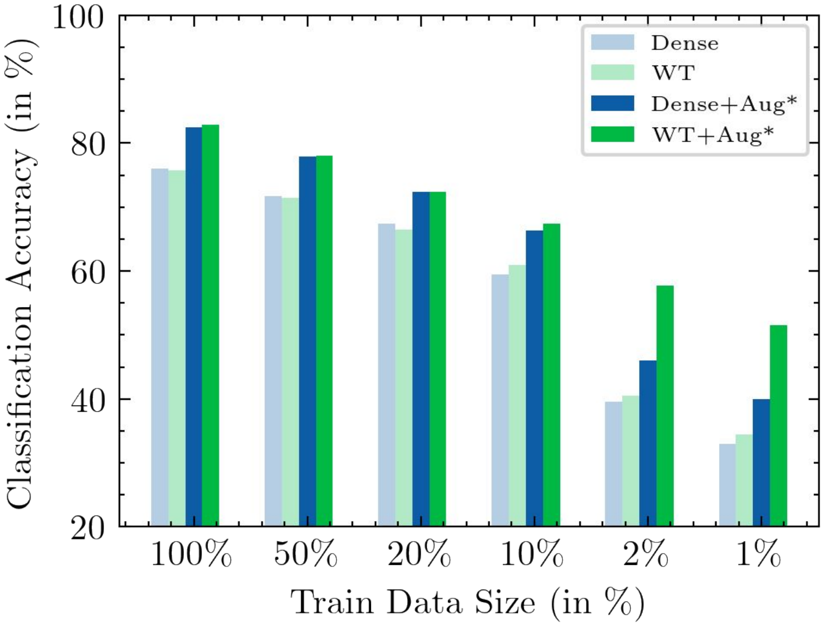
Synthetic Corruptions

(Hendrycks et al. '19)

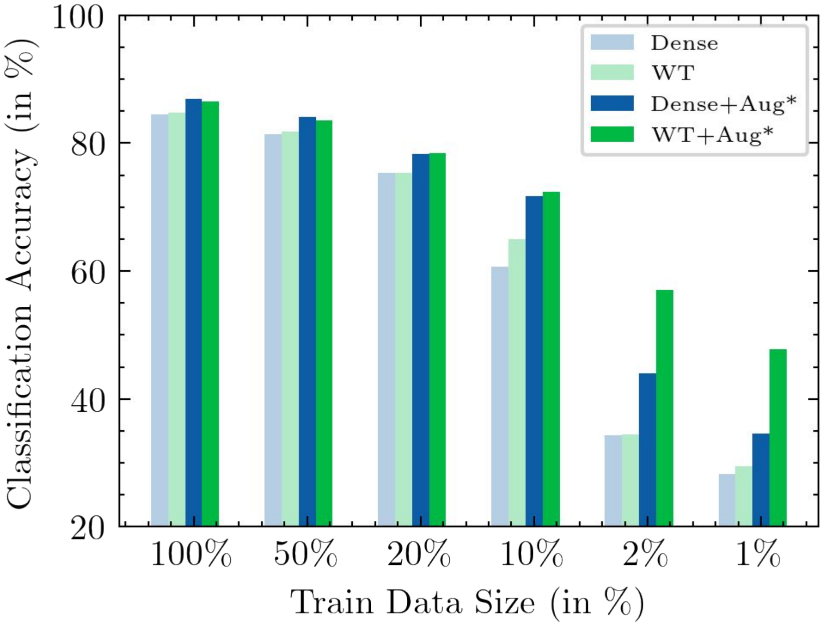
* Representative figure from http://ai.bu.edu/visda-2021/

Sparse Winning Tickets are robust to distributional shifts

Synthetic Corruptions



Domain Shifted



As data size **decreases**, winning tickets exhibit **superior robustness** to several corruptions.

IMP compliments existing data-efficient training

How well does IMP fare against more specific methods for data-efficiency?

Can their data-efficiency be further improved with pruning?

	Method	CIFAR10 (2%) D+Aug*	CIFAR10 (2%) WT+Aug*	CIFAR10 (1%) D+Aug*	CIFAR10 (1%) WT+Aug*
Fine Tuning	Random Init. (R18)	55.14%	70.05%	43.8%	59.66%
	ImageNet Init. (R18)	75.50%	77.92%	66.00%	69.46%
	SimCLR Init. (R18)	52.58%	64.09%	37.56%	44.39%
Reg. Loss	Cosine Loss	64.82%	72.63%	45.87%	64.67%
	t-vMF Loss	62.23%	72.81%	41.70%	64.54%
Reg. Arch	Full Conv.	62.16%	73.24%	49.30%	64.20%
	Harmonic Nets	61.36%	66.48%	22.97%	49.85%
Mobile. Arch	MobileNetv2	64.64%	71.01%	51.8%	61.63%

IMP compliments existing data-efficient training

IMP always further improves performance

How well does IMP fare against more specific methods for data-efficiency?

Can their data-efficiency be further improved with pruning?

	Method	CIFAR10 (2%) D+Aug*	CIFAR10 (2%) WT+Aug*	CIFAR10 (1%) D+Aug*	CIFAR10 (1%) WT+Aug*
Fine Tuning	Random Init. (R18)	55.14%	70.05%	43.8%	59.66%
	ImageNet Init. (R18)	75.50%	77.92%	66.00%	69.46%
	SimCLR Init. (R18)	52.58%	64.09%	37.56%	44.39%
Reg. Loss	Cosine Loss	64.82%	72.63%	45.87%	64.67%
	t-vMF Loss	62.23%	72.81%	41.70%	64.54%
Reg. Arch	Full Conv.	62.16%	73.24%	49.30%	64.20%
	Harmonic Nets	61.36%	66.48%	22.97%	49.85%
Mobile. Arch	MobileNetv2	64.64%	71.01%	51.8%	61.63%

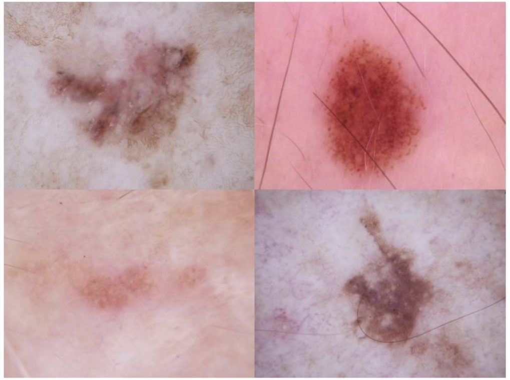
Random Init + IMP performs better than several specialized data efficient methods

IMP compliments and be combined with existing data-efficient techniques, and further improves performance on an average by **8% and 15% at 2% and 1%** data sizes respectively.

Generalization to other low-sample datasets

Do these results hold in cases of highly specialized datasets of images from medical, scientific domain, or just images from different distribution - differing greatly in size, color, or channels than seen in typical image datasets?

CLaMM



ISIC

EuroSAT

Generalization to other low-sample datasets

Do these results hold in cases of highly specialized datasets of images from medical, scientific domain, or just images from different distribution - differing greatly in size, color, or channels than seen in typical image datasets?

Method	CLaMM		ISIC		EuroSAT	
	D	WT	D	WT	D	WT
Full Conv.	46.77%	57.91%	56.42%	58.20%	76.05%	77.06%
Cosine Loss	49.60%	60.15%	59.03%	61.00%	82.28%	88.68%
t-vMF Loss	24.50%	59.67%	56.22%	59.30%	71.14%	88.26%
Harmonic Nets	42.58%	44.14%	50.56%	52.51%	79.40%	83.28%
ImageNet Init. (R18)	47.46%	55.86%	59.72%	62.80%	90.75%	91.32%
ImageNet Init. (R50)	51.66%	57.03%	61.73%	64.88%	92.89%	93.39%
Random Init.	50.29%	55.76%	57.34%	59.73%	83.44%	87.85%

The winning tickets outperform the dense model on an average by **12%, 2.3%, and 5.5%** on the CLaMM, ISIC and EuroSAT datasets respectively.

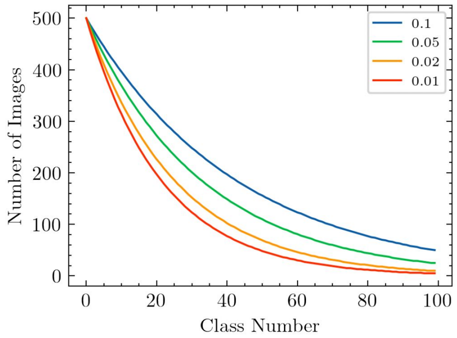
Generalization to complex low-sample datasets

We verify if our results also hold true in the case of small data-subsets from:

- Complex datasets containing much larger number of classes with only 5-50 samples per class.

Dataset	D+Aug	WT+Aug
ImageNet (5%)	28.82%	31.04%
CIFAR100 (2%)	17.21%	25.06%
CIFAR100 (1%)	11.21%	16.44%

- Simulated datasets with imbalanced (long-tail of) classes.

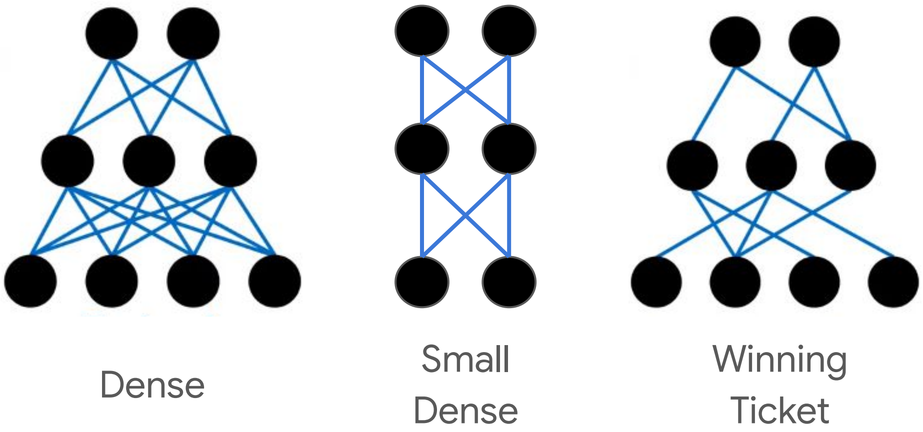


Imbalance Factor	D+Aug.	WT+Aug.
0.10	64.77	64.92
0.05	58.43	59.43
0.02	50.91	51.72
0.01	45.13	46.73

What properties of winning tickets make them `data-efficient’

Can data-efficiency just be contributed to fewer parameters?

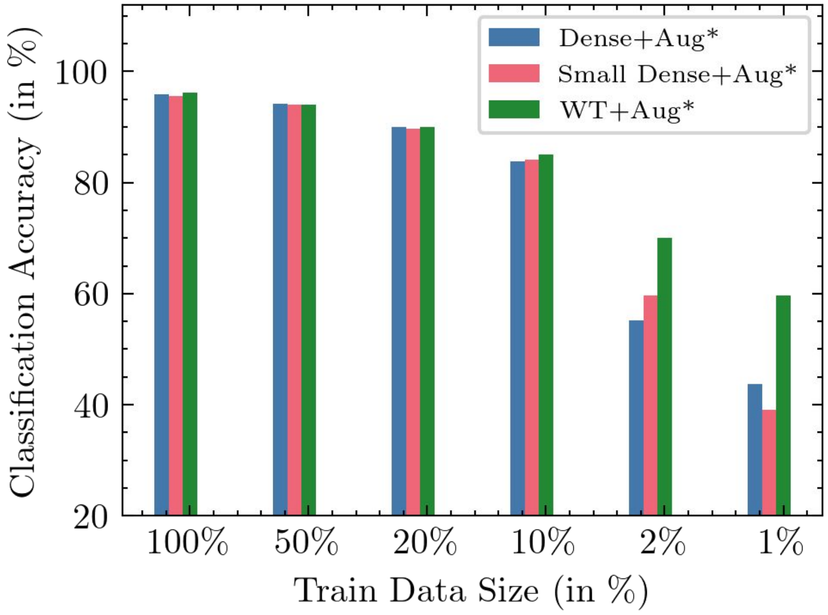
Do the learned connections play any role?



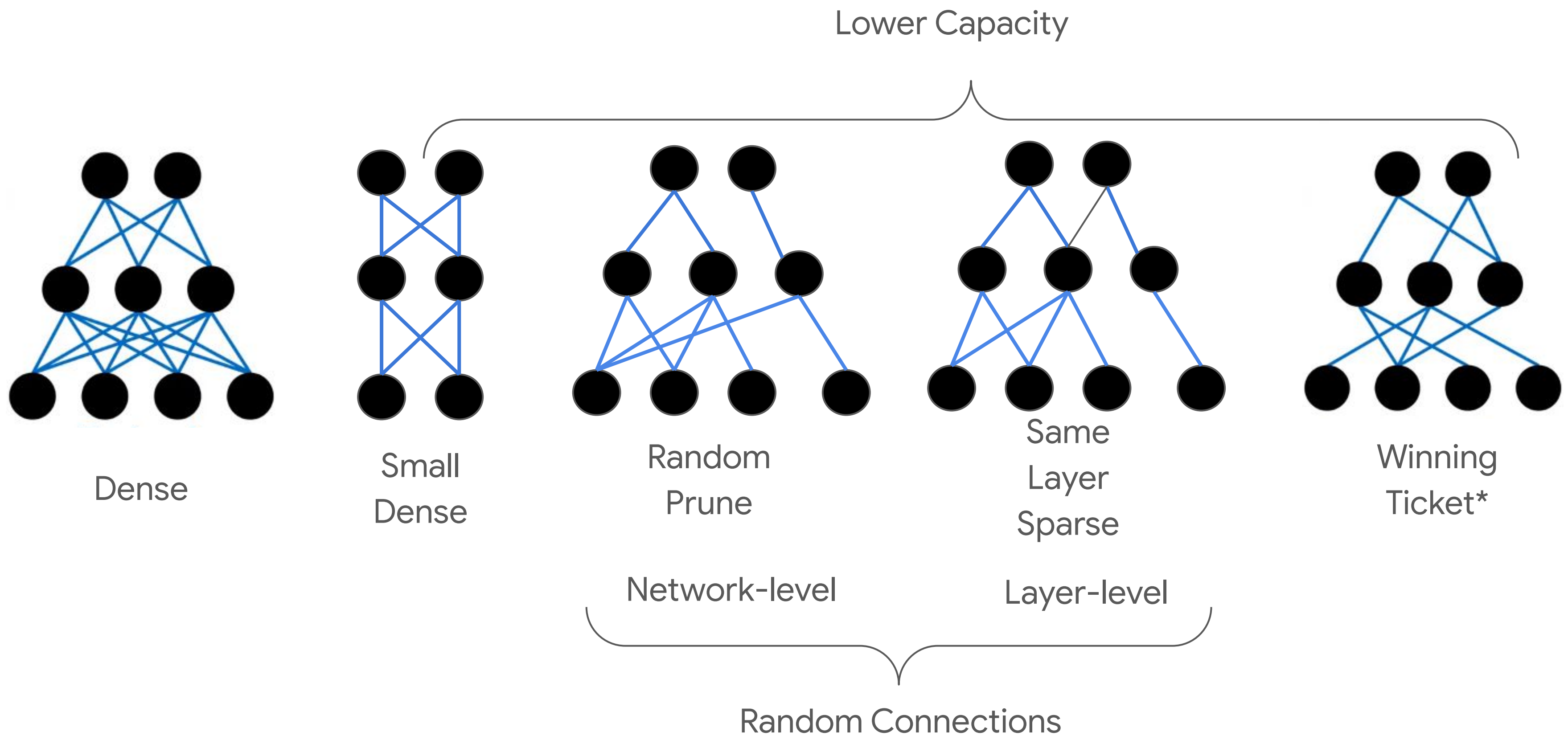
Lower Network Capacity improves data-efficiency

Smaller capacity networks showcase performance improvements - more drastically at least data sizes.

However, the winning ticket still outperforms a dense network of similar capacity quite significantly, indicating that beyond capacity perhaps the network connections also play an important role.



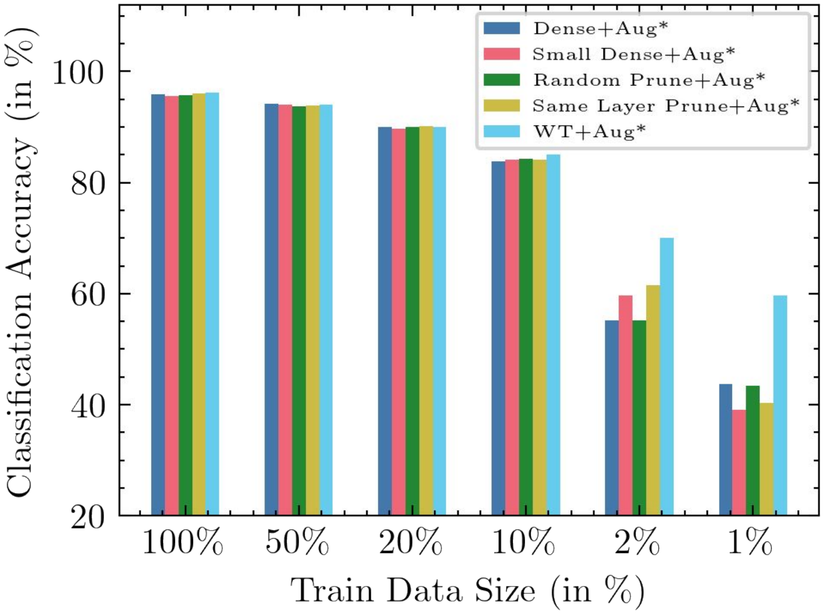
Both Network Capacity and Connectivity are important!



Both Network Capacity and Connectivity are important!

Learned connections significantly outperforms random connections both at a network, layer level.

Both network capacity and connectivity play a vital role in improving data-efficiency of sparse networks.

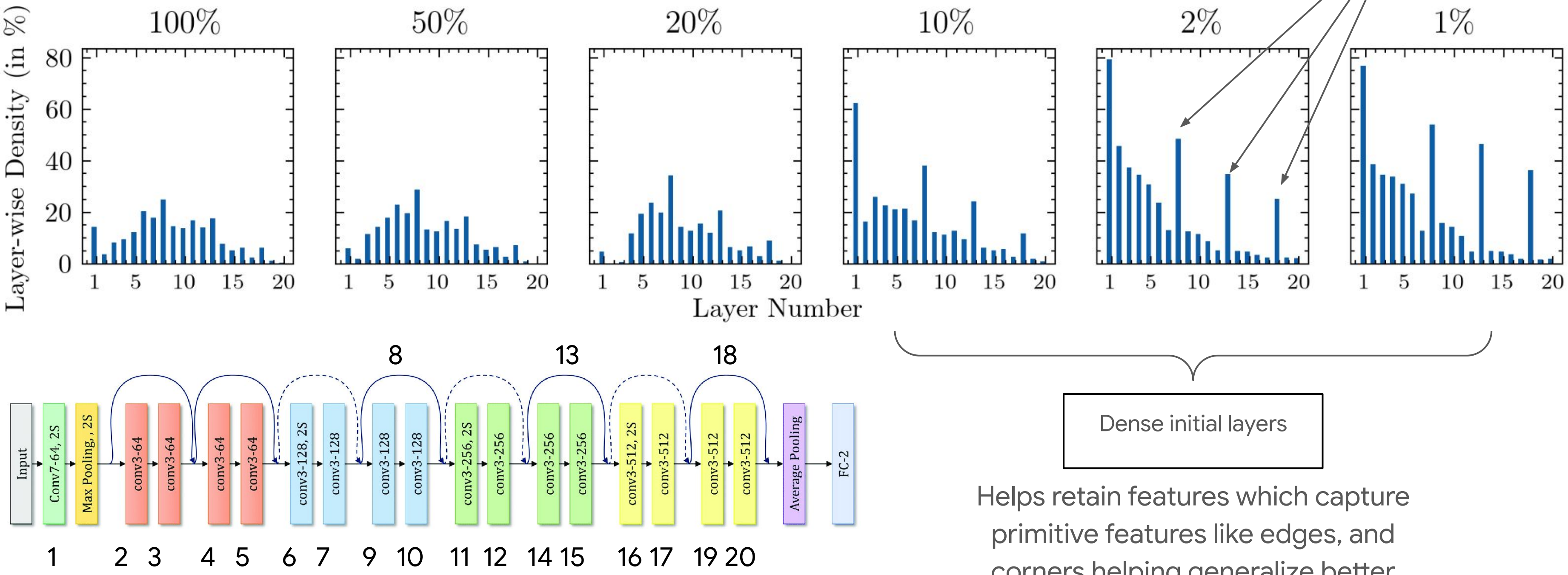


Which Layers are Getting Pruned?

Helps enable minimal changes to output even upon input change (identity connections), making them robust

Dense residual layers

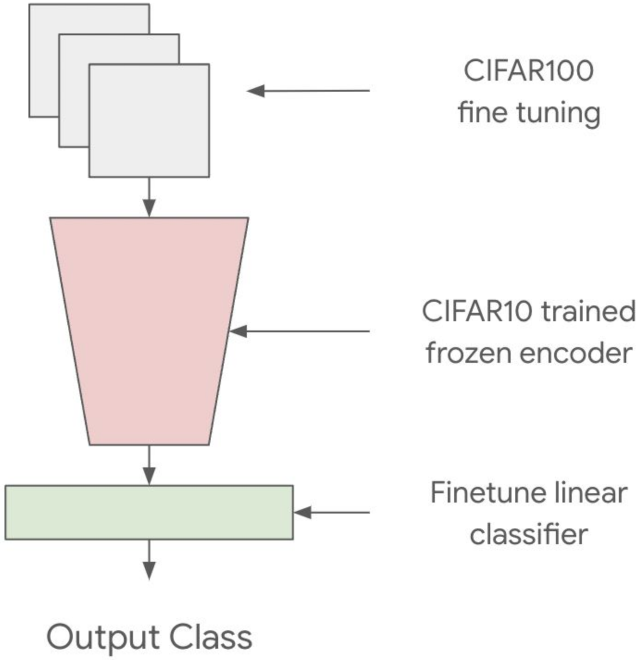
Dense middle layers



Dense initial layers

Helps retain features which capture primitive features like edges, and corners helping generalize better

Empirical Evidence to Generalizability of the learned representations

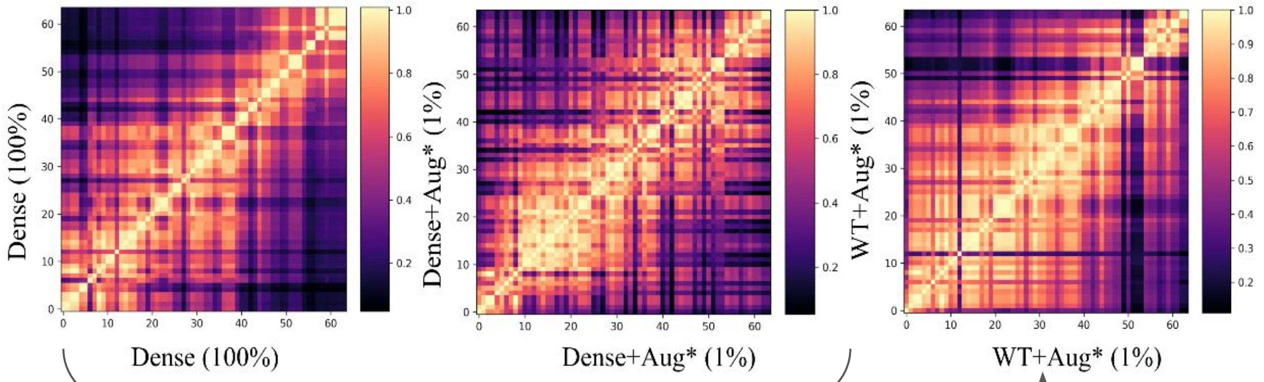


Sub-Network	Accuracy
Reference (dense, 100%)	20.79%
Dense (1%)	6.42%
Dense+Aug. (1%)	4.94%
WT+aug. (1%)	23.67%

Winning ticket trained only on 1% data outperforms corresponding dense counterparts.

More surprisingly, it outperforms a dense network trained on full 100% CIFAR10 data.

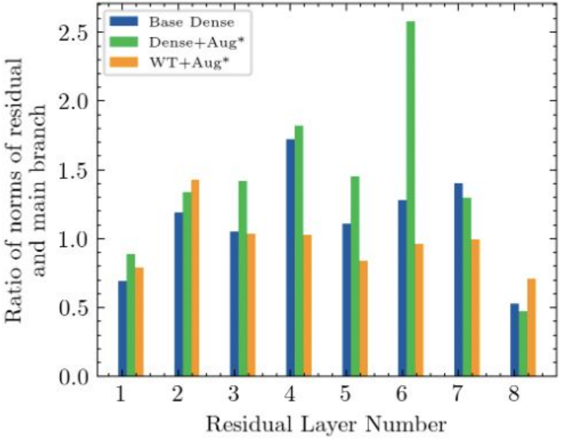
Layer-Wise Representation Similarity



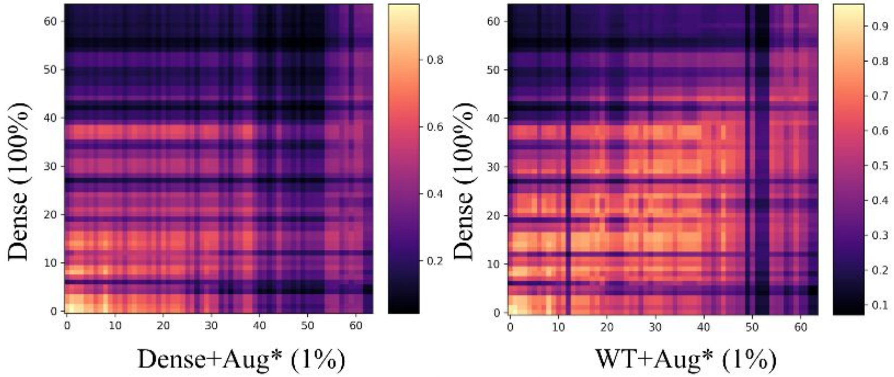
Intra Block similarity

Uniform Representation across layers

Layer-Wise Representation Similarity



Winning tickets exhibit lower propagation of information via the residual streams indicating the overall uniform representation is directly related to extraction of globally generalizable features.



Winning ticket trained only on 1% data exhibits greater similarity to a network trained on 100% data.

Key Takeaways

- With decreasing training data, the winning ticket gets sparser and when combined with augmentations considerably outperforms the dense network.
- Winning tickets avoid memorisation to prevent overfitting, showcase improved robustness to several distribution shifts.
- IMP compliments several data-efficient strategies to further improve performance.
- These results also hold in the case of diverse datasets, simulated imbalanced datasets with 50-100 images per class only.
- Lower capacity, and learned connectivity help the winning ticket learn more generalizable representations.