When Does Re-initialization Work?

Sheheryar Zaidi[1*], Tudor Berariu[2*], Hyunjik Kim[3], Jörg Bornschein[3], Claudia Clopath[2,3], Yee Whye Teh[1,3], Razvan Pascanu[3]

*Equal contribution, [1]University of Oxford, [2]Imperial College London, [3]DeepMind

Spotlight @ ICBINB Workshop, NeurIPS 2022

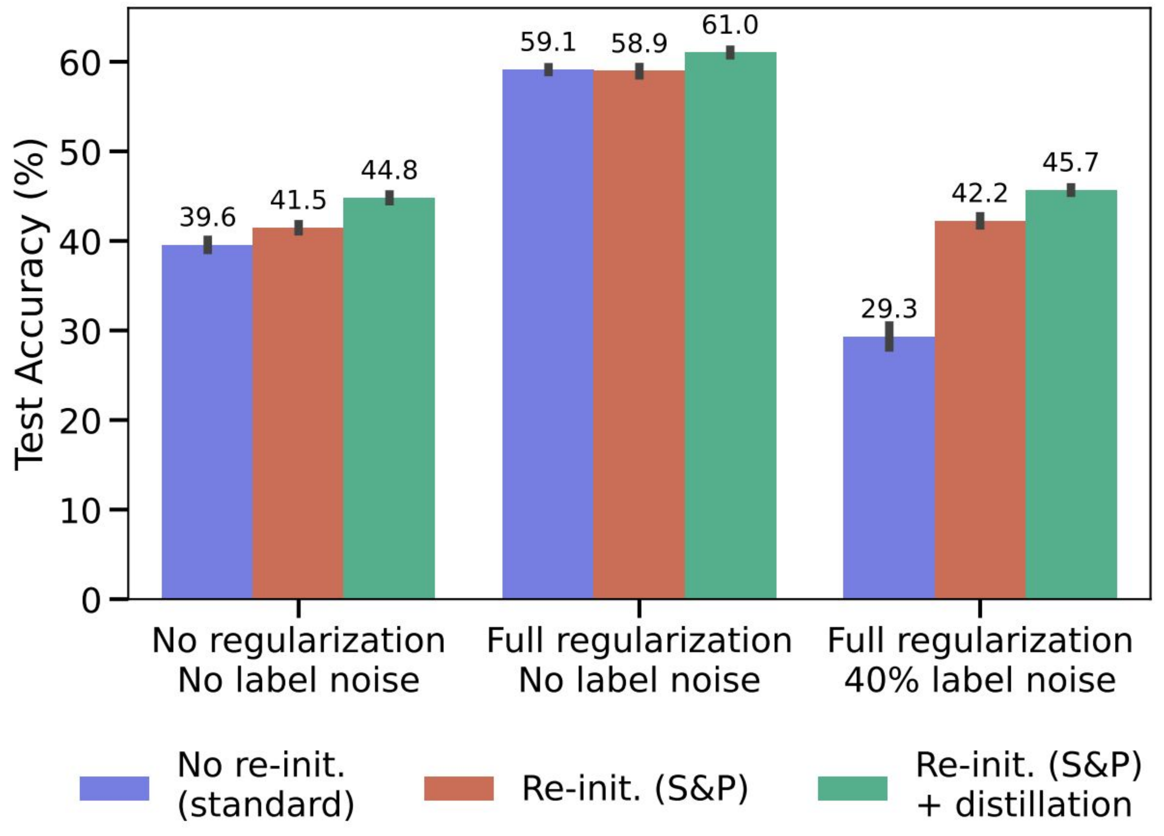
Summary

A study of the effect of repeatedly **re-initializing** a neural network during training on generalization performance.

Re-initialization regularizes learning and improves generalization compared to standard training (*i.e.* without re-initialization) in the absence of other regularization techniques.

In SOTA training protocols, re-initialization offers little benefit, apart from **robustness to optimization hyperparameters**.

Under label noise, re-initialization significantly improves performance, even alongside other regularization techniques.



Comparison of standard training with re-initialization using *Shrink & Perturb* in three scenarios on Tiny ImageNet with PreAct-ResNet-18.

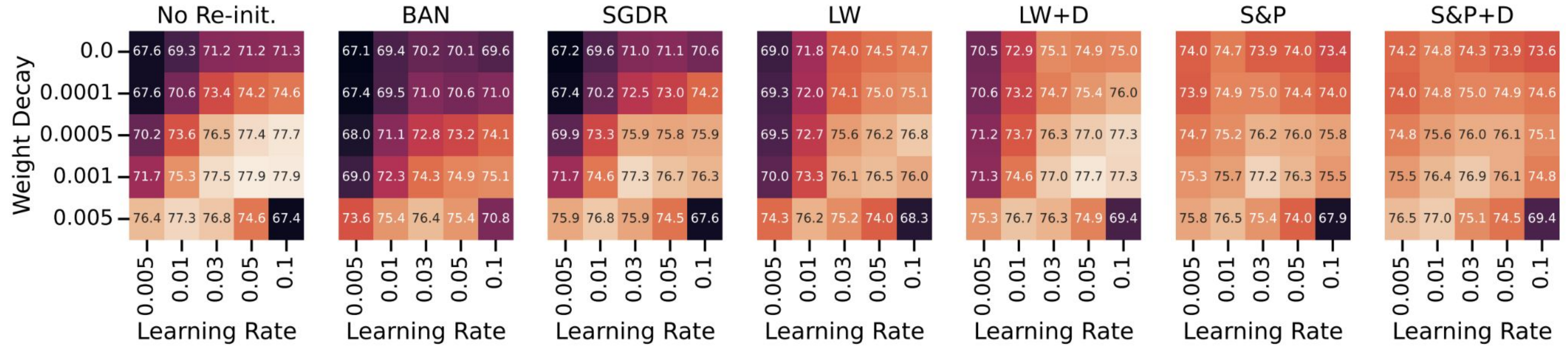
Re-initialization Alongside Other Regularization

With other regularization (data augmentation, learning rate schedule and weight decay), re-initialization methods offer no further performance benefit.

But performance becomes less sensitive to LR/WD hyperparameter choices.

Data Aug.	Cosine Anneal.	Weight Decay	No Re-initialization (standard training)	Self-distillation (fixed-budget BAN)	SGDR	Layer-wise Re-initialization w/o dist.	Layer-wise Re-initialization w/ dist.	Shrink & Perturb w/o dist.	Shrink & Perturb w/ dist.
✗	✗	✗	55.5 ±0.6	56.4 ±0.5	N/A	61.0 ±0.6	62.5 ±0.2	63.1 ±0.6	**63.5** ±0.3
✓	✗	✗	70.8 ±0.1	70.5 ±0.5	N/A	72.1 ±0.3	**74.7** ±0.2	71.9 ±0.1	74.0 ±0.6
✓	✓	✗	71.2 ±0.2	70.9 ±0.4	71.0 ±0.6	74.6 ±0.5	**75.4** ±0.2	**75.4** ±0.3	**75.4** ±0.4
✓	✓	✓	**77.9** ±0.2	77.2 ±0.1	77.5 ±0.2	77.5 ±0.1	77.3 ±0.3	77.5 ±0.2	77.0 ±0.3

ResNet-18 trained on CIFAR-100.

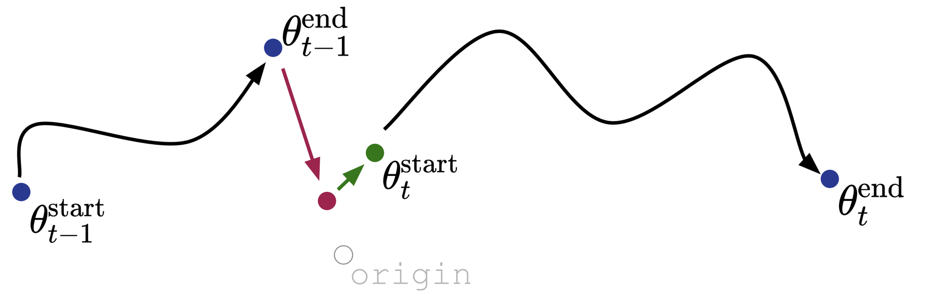


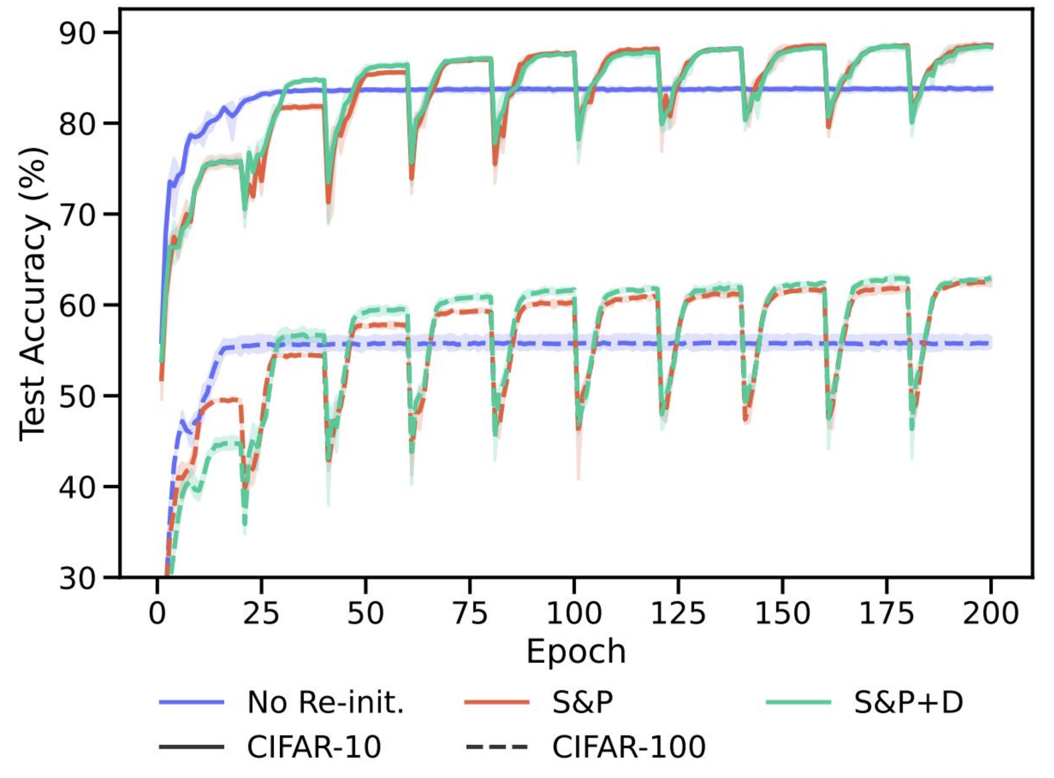
Test accuracy vs. LR/WD choices on CIFAR-100 with ResNet-18.

The Regularizing Effect of Re-initialization

In a simple setting without any other regularization, re-initialization—even upto 25 times during training—considerably improves generalization.

Shrink & Perturb (S&P) is a re-initialization method that multiplicatively *shrinks* and additively *perturbs* (with Gaussian noise) the weights: $\theta_{\mathrm{RI}} = \lambda\theta + \gamma\theta_{\mathrm{init}}$



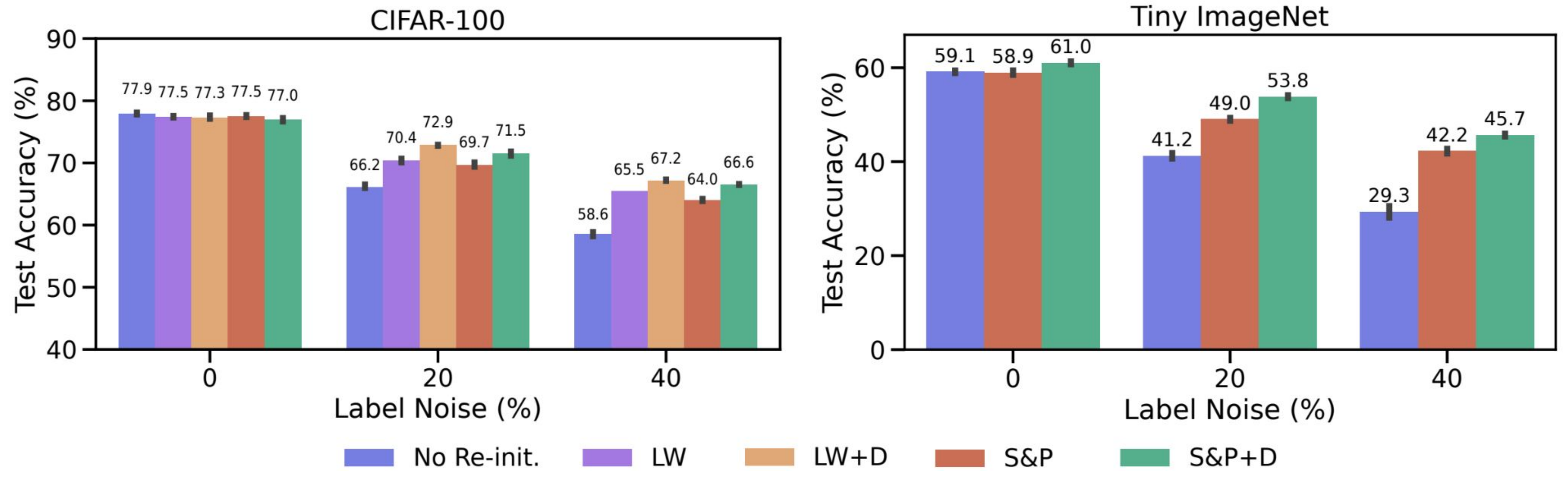


ResNet18 trained with 10 stages of re-initialization.

Re-initialization Under Label Noise

Even if all other regularization is used and carefully tuned, re-initialization offers significant (>10 points) benefit over standard training!

See paper for more findings.



Training under label noise with all other regularization: the benefit of re-initialization improves as noise increases.