Are wider nets better given the same number of parameters?

Anna Golubeva, Guy Gur-Ari, Behnam Neyshabur @ Blueshift, Alphabet
PI QUANTUM INTELLIGENCE LAB V VECTOR INSTITUTE
ICLR 2021 arXiv:2010.14495

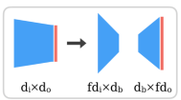
wider layers ⇔ more parameters ⇒ better performance

▶ Is the performance gain due to more params or larger width?

How to increase width independently of the number of params?

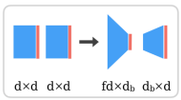
Bottleneck Methods

linear: split each layer in two



changes depth
strongly affects trainability

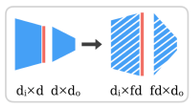
non-linear: modify layers in pairs



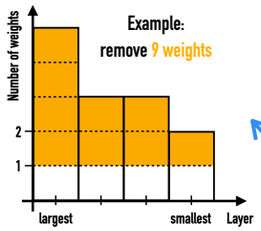
leads to worse performance

Static Sparsity

random, applied at init



does not alter the NN structure



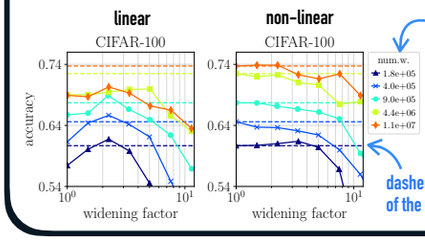
- per-layer distribution according to layer size
- in-layer distribution uniform across all layer dimensions

Our approach in summary:

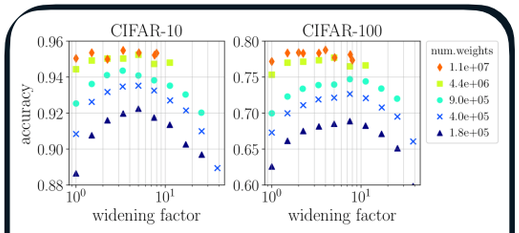
- select model type and architecture
  baseline: dense model (full connectivity) — e.g. ResNet18 with 32 output channels in the first conv layer
- fix the number of weights
- build a family of models having different widths and sparsity, but same number of weights
  wide & sparse: increase the width and remove excess weights
- train and compare performance (task: image classification)

Bottleneck results

Best test accuracy obtained by ResNet-18 models widened using the bottleneck methods:

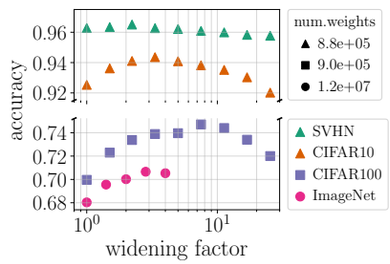


ResNet-18 on CIFAR



improvement is strongest for smaller models

ResNet-18: results in overview



Test accuracy of ResNet-18 as a function of width: performance improves as width is increased, even though the number of weights is fixed!

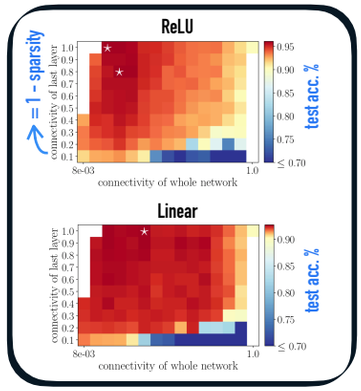
ResNet-18 on ImageNet

The improvement obtained by the sparse models with increasing width is on par with the dense models:

width	64	90	128	181	256
dense	68.03 (11.7)	69.11 (22.8)	70.22 (45.7)	70.91 (90.7)	71.89 (180.6)
sparse	–	69.56 (11.7)	70.02 (11.7)	70.66 (11.7)	70.53 (11.7)

top-1 test acc. % num. weights in 10^6

MLP-1 on MNIST



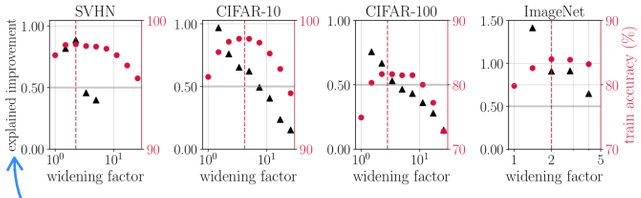
∞width limit and sparse GP kernel

hypotheses:
- perf improvement is correlated with having a GP kernel that is closer to the ∞width kernel
- the distance to the ∞width kernel can be reduced by increasing network width

▶ compute GP kernel of a sparse ReLU net with 1 hidden layer in theory and experiments

How much improvement is due to width only?

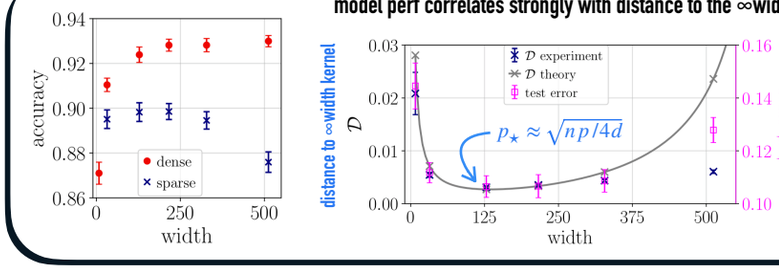
compare perf increase for wide & sparse to wide & dense models



$$= \frac{\Delta \text{ test acc. (sparse)}}{\Delta \text{ test acc. (dense)}}$$

MLP-1 on MNIST: test accuracy and GP kernel distance

model perf correlates strongly with distance to the ∞width kernel



theory predicts optimal connectivity

$$p_\star \approx \sqrt{\frac{np}{4d}}$$

with np = const.
input dimension d
layer width n