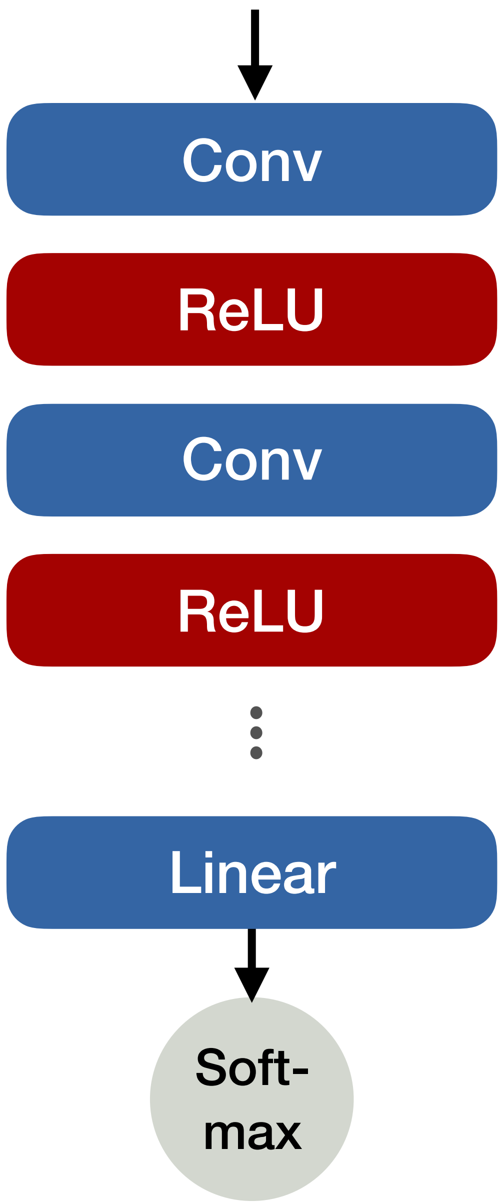
Open Problem: Understanding generalization

Generalization in deep learning

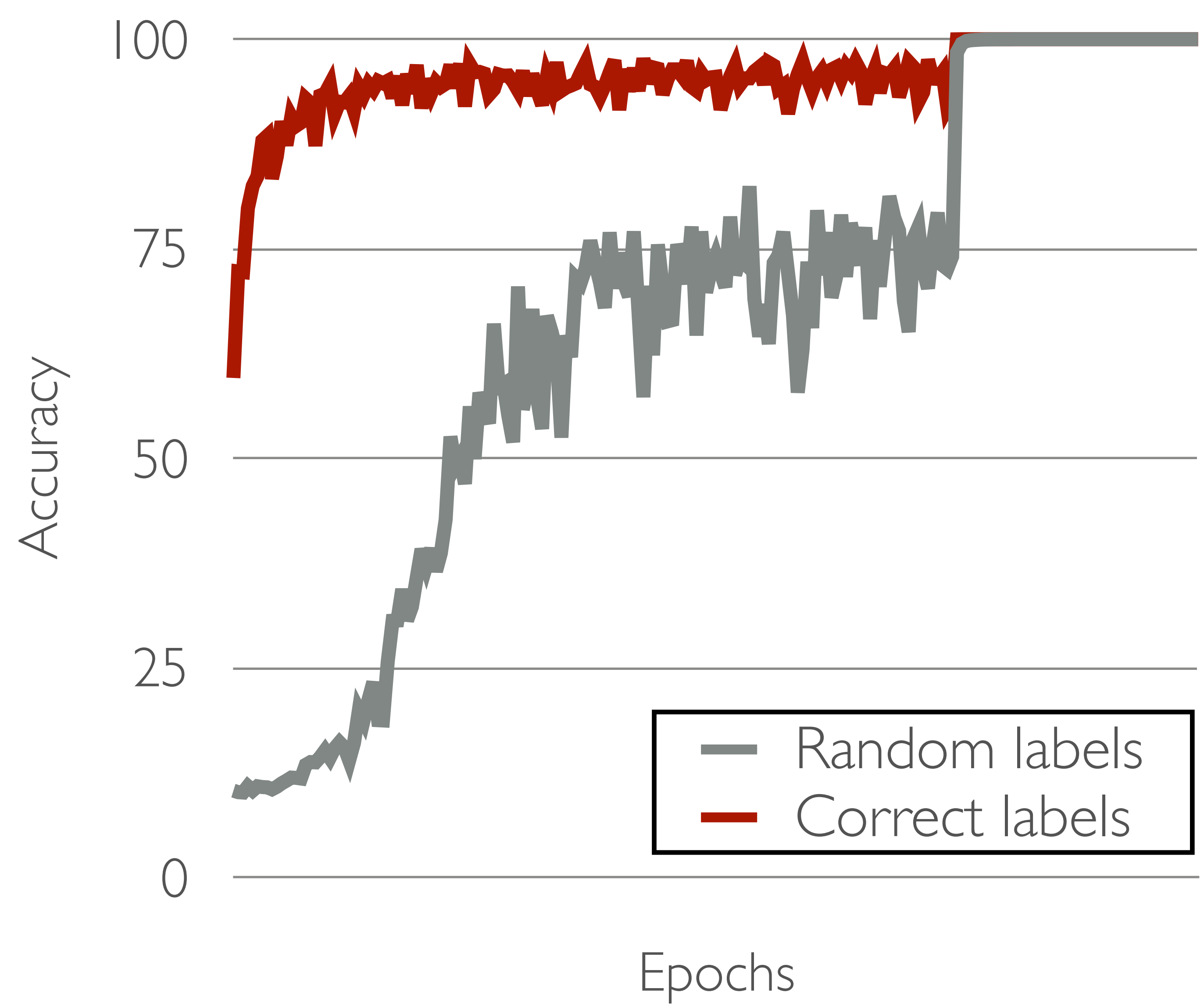
- Standard wisdom
  - Bigger/wider models overfit more



Deep networks are big enough to remember all training data

- Deep networks easily fit random labels
  - Memorize all data
  - Works even for random noise inputs

wide resnet on cifar-10



Understanding deep learning requires rethinking generalization, Zhang etal. 2017

Why does SGD still work?

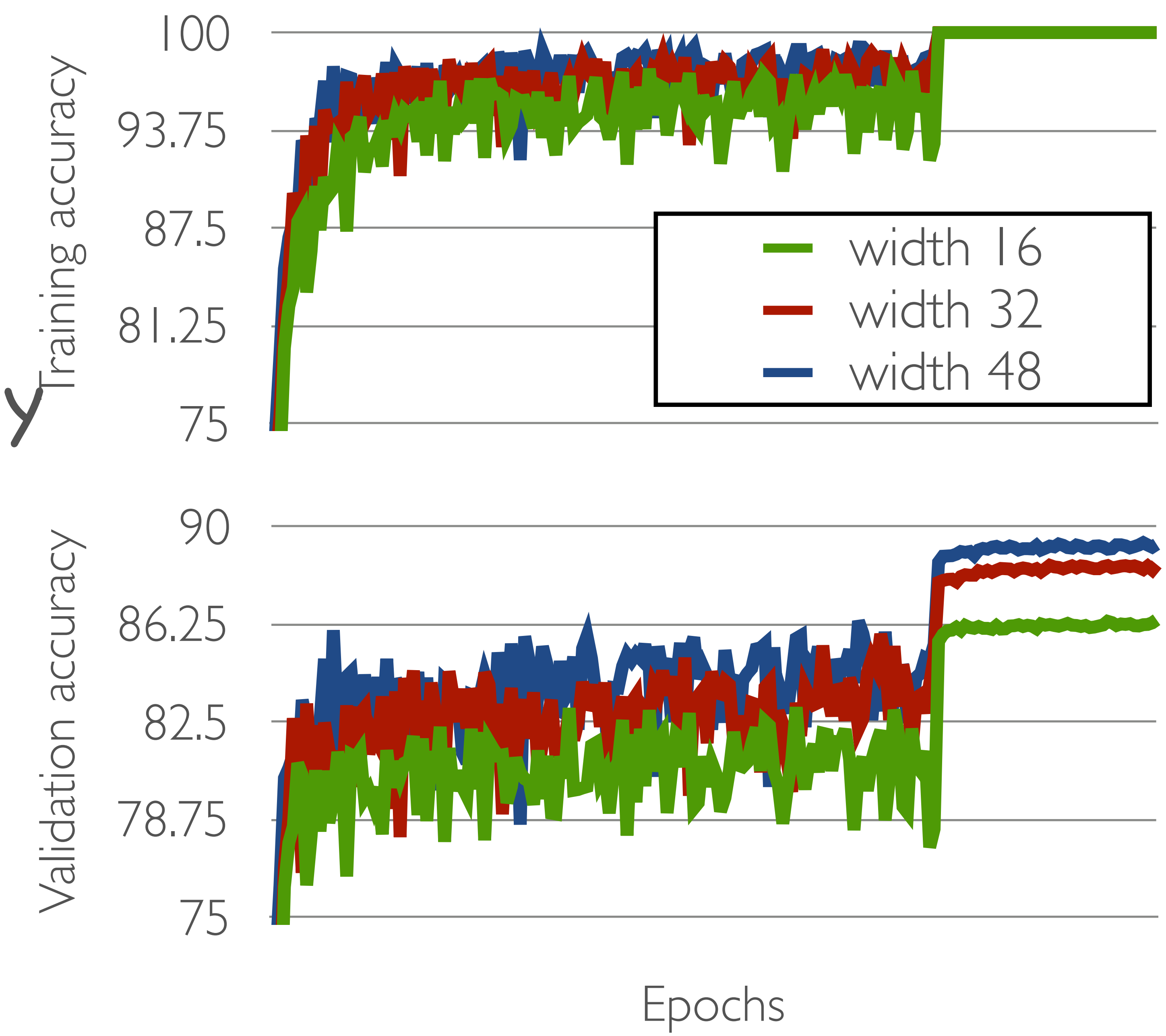
- SGD gradually minimizes objective
- Prefers solutions close to initialization
- Implicitly regularizes
- Random labels take SGD on a longer path

Exploring generalization in Deep Learning, Neyshabur etal. 2017

Larger networks overfit less

- Without data augmentation
  - 100% training accuracy
  - Larger models generalize better
  - Hence overfit less

wide resnet on cifar-10

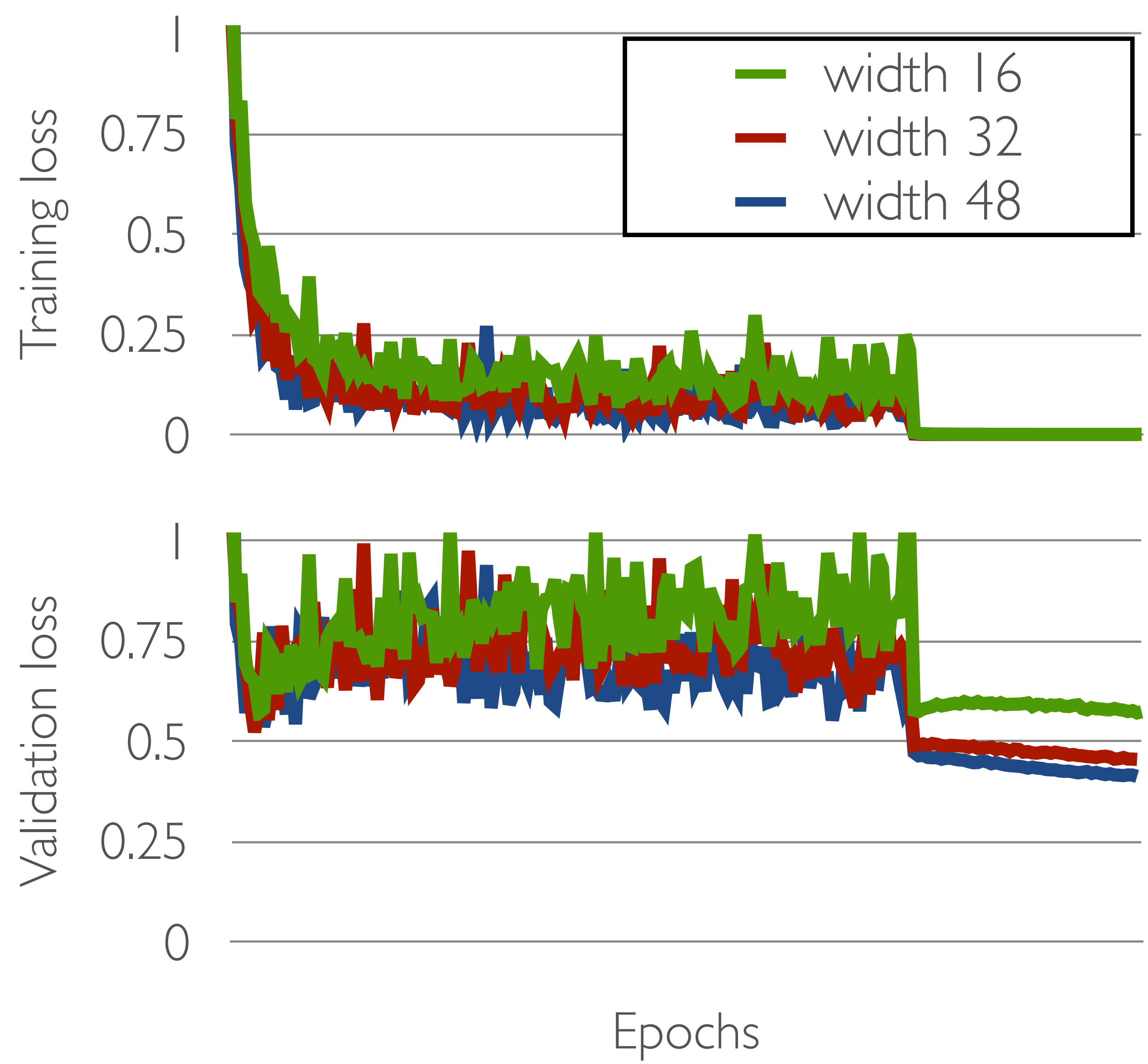


Understanding deep learning requires rethinking generalization, Zhang etal. 2017

Larger networks overfit less

- All models overfit massively on loss (log likelihood)

wide resnet on cifar-10



On Calibration of Modern Neural Networks, Guo etal. 2017

Larger networks overfit less

- Do we need a new learning theory?
- Do we need new intuitions?

In summary

- Models can overfit, but do not with SGD and data augmentation
  - Implicit regularization
  - How to do make it explicit?
  - Overfitting is dependent on learning algorithms (e.g. Adam overfits more)
- How can we measure overfitting?