Open Problem: Pruning and compression

How do we train a small network?

- Idea 1:
  - Randomly initialize and train network
- Idea 2:
  - Train a larger network and make it small

Network distillation

- Train an ensemble of large networks
  - Train a small network to mimic it's output (with cross entropy)
  - Important: Reduce confidence of ensemble prediction (soft targets)

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv 2015.

Why does distillation work?

- Dark knowledge
- Networks learn about (visual) relationships of classes
- Boost training signal

Network pruning / factorization

- Train a wide network (many channels)
  - Remove channels/weights that are used the least
    - 90% of parameters can be removed after training
    - Training the small network is challenging

H Li, A Kadav, I Durdanovic, H Samet, HP Graf, "Pruning Filters for Efficient ConvNets", ICLR 2017
S Han, J Pool, J Tran, W Dally, "Learning both Weights and Connections for Efficient Neural Network" NIPS 2015

Possible explanation: Lottery ticket hypothesis

- Not all initializations are created even
- Train network

A randomly-initialized, dense neural network contains a subnetwork that is initialized such that—when trained in isolation—it can match the test accuracy of the original network after training for at most the same number of iterations.

Jonathan Frankle, Michael Carbin, "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks", ICLR 2019

Lottery ticket hypothesis

- Very nice idea
- Likely not the full story

Zhuang Liu et al., "Rethinking the Value of Network Pruning", ICLR 2019