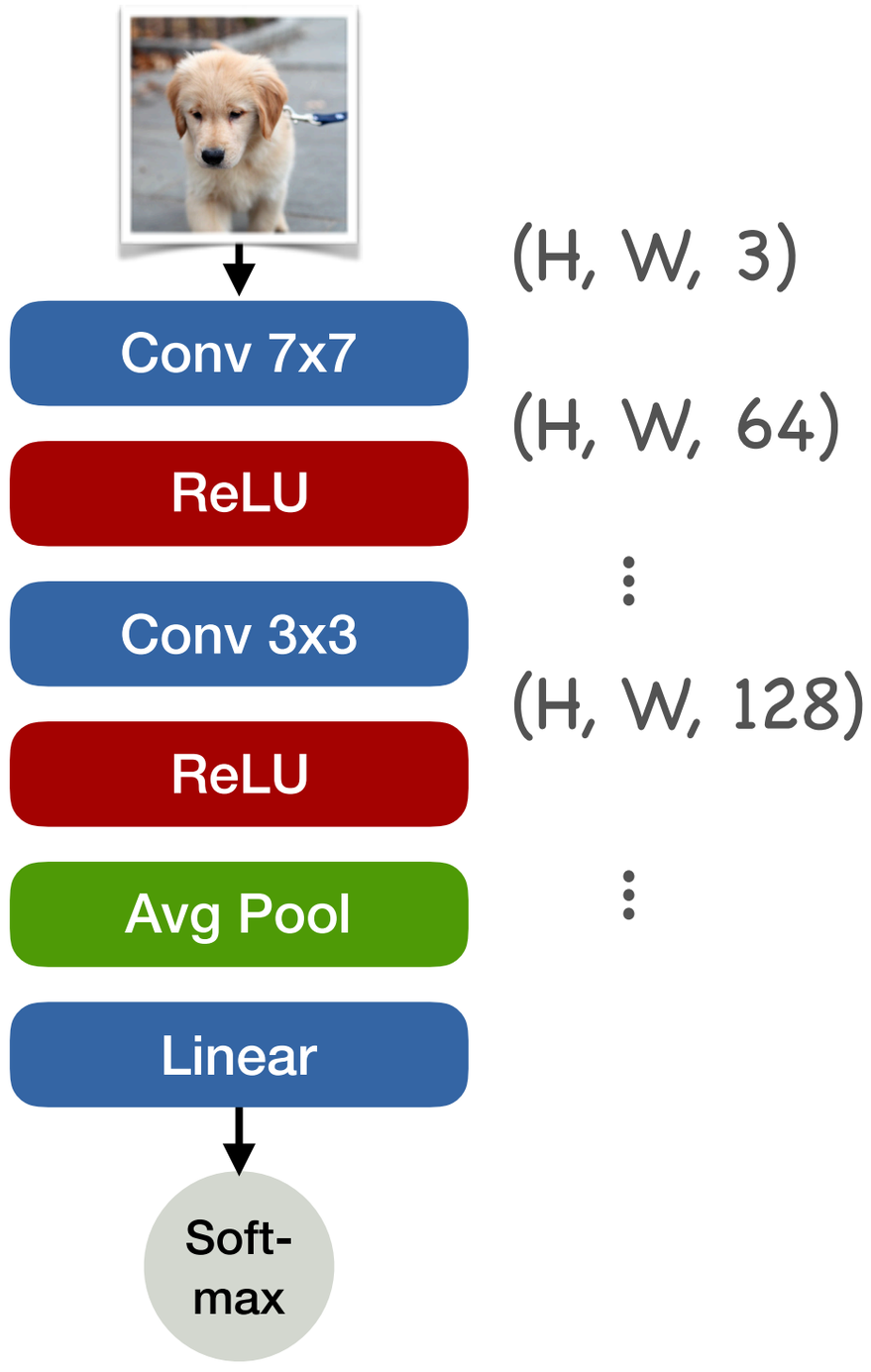
Weight decay

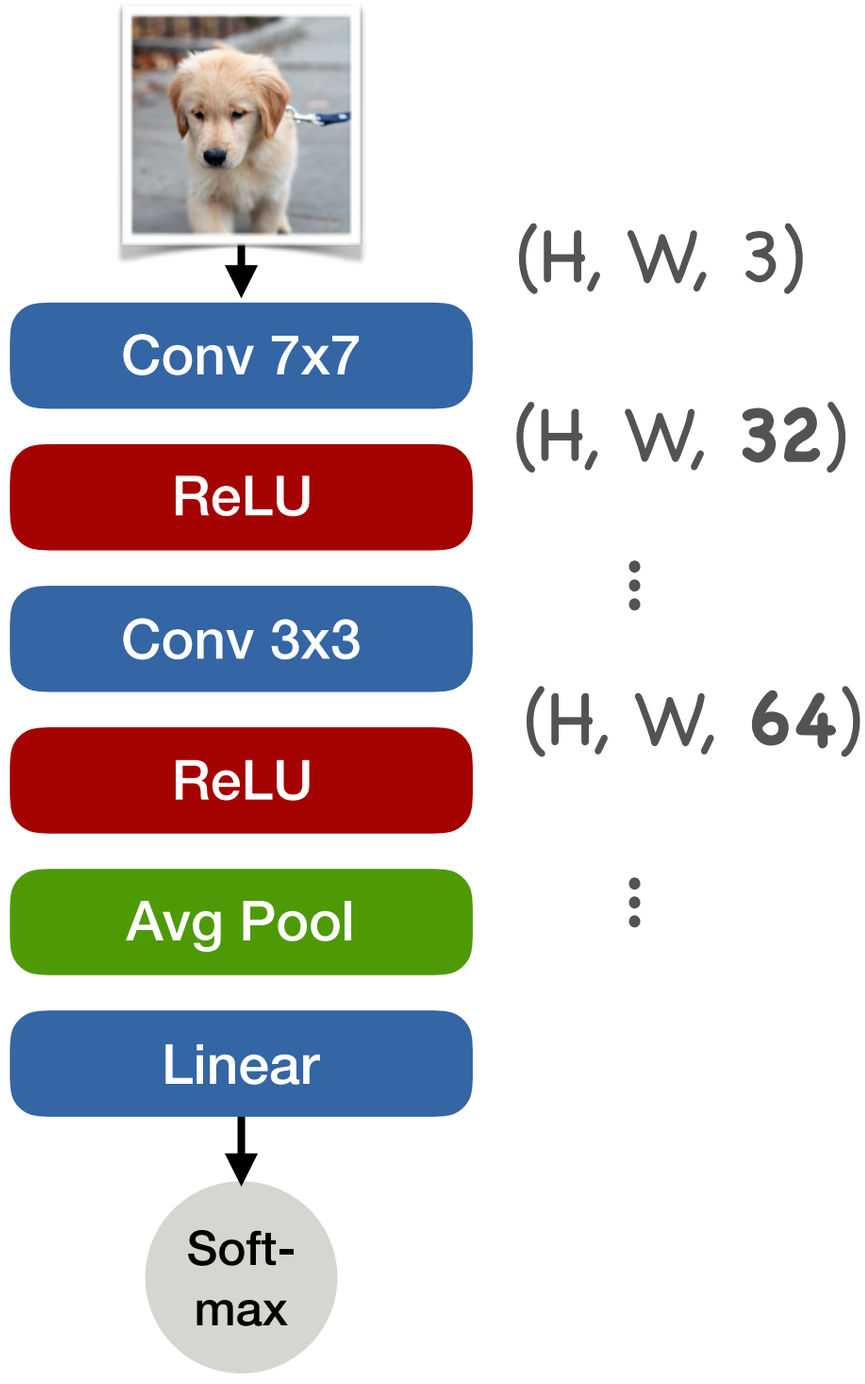
Simpler models

- Traditional wisdom
  - Simpler model = less overfitting



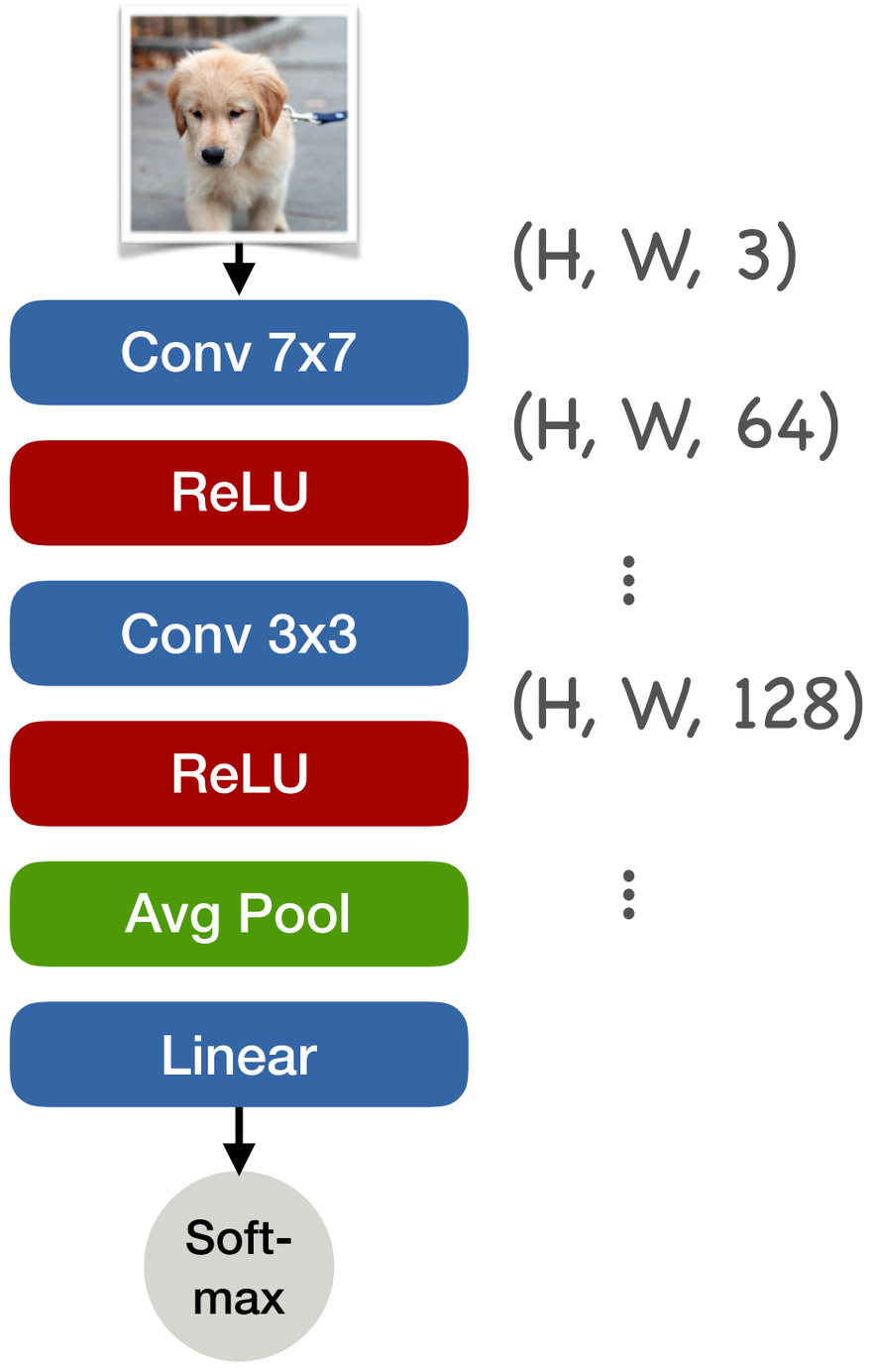
Idea 1: Smaller model

- Overfits less
- Fits less
  - Worse generalization



Idea 2: Big model with regularization

- Weight decay
  - Keep weights small (L2 norm)
- Works sometimes
  - Keep weight at same magnitude



How to use weight decay?

- Parameter in optimizer,
  e.g. torch.optim.SGD or
  torch.optim.Adam
  - weight_decay
  - Use 1e-4 as default

Other reasons to use weight decay

- Network weights cannot grow infinitely large
  - Helps handle exploding gradients