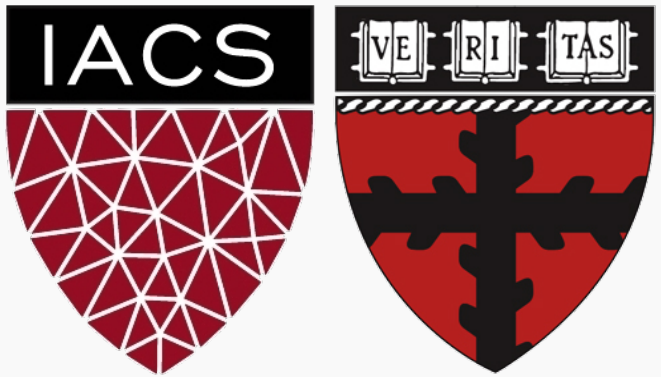
Neural Network Regularization
Dropouts & Batch Normalization

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Tanner



Outline

Regularization of NN

- Norm Penalties
- Early Stopping
- Data Augmentation
- **Dropout**

Co-adaptation

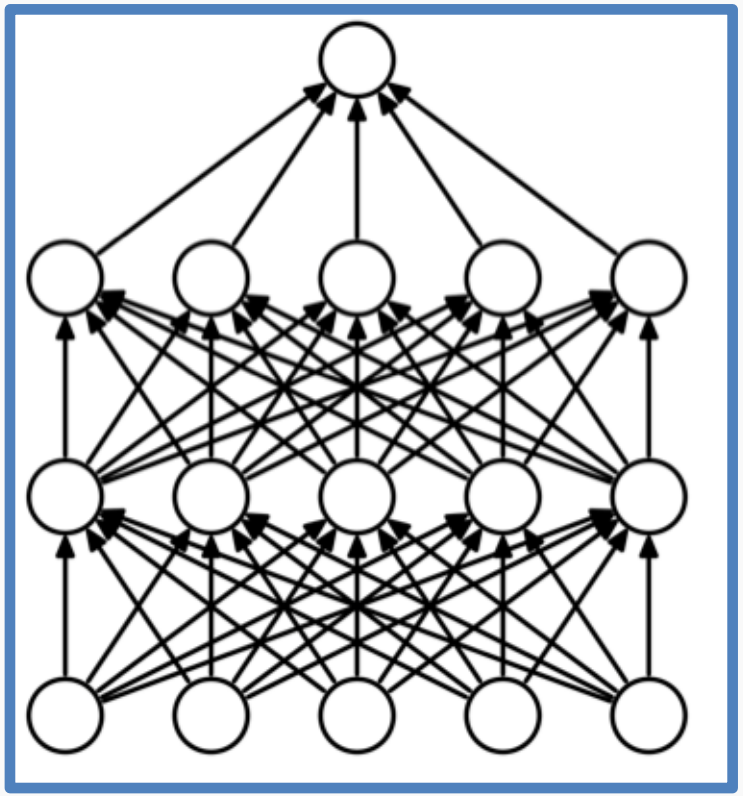
Overfitting is when we are very sensitive to the input and therefore the model fits the noise.

L1 and L2 regularizations 'shrink' the coefficients to avoid this problem.

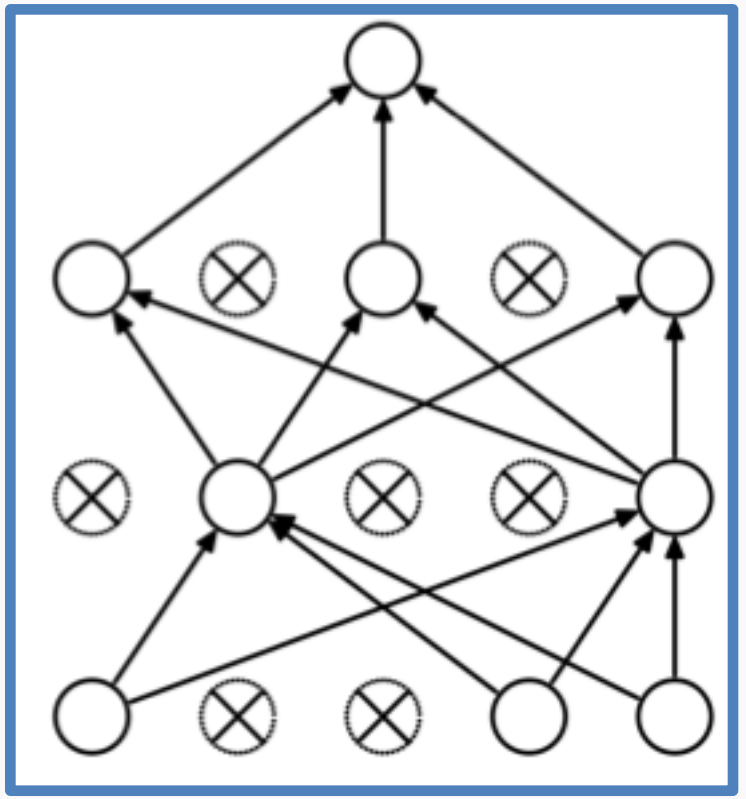
However in a large network many units can collaborate to respond to the input while the weights can remain relatively small. This is called co-adaptation.

Dropout

- Randomly set some neurons and their connections to zero (i.e. “dropped”)
- Prevent overfitting by reducing co-adaptation of neurons
- Like training many random sub-networks



Standard Neural Network



After applying dropout

https://arxiv.org/pdf/1207.0580.pdf

Dropout: Training

For each new example in a mini-batch (could be for one mini-batch depending on the implementation):

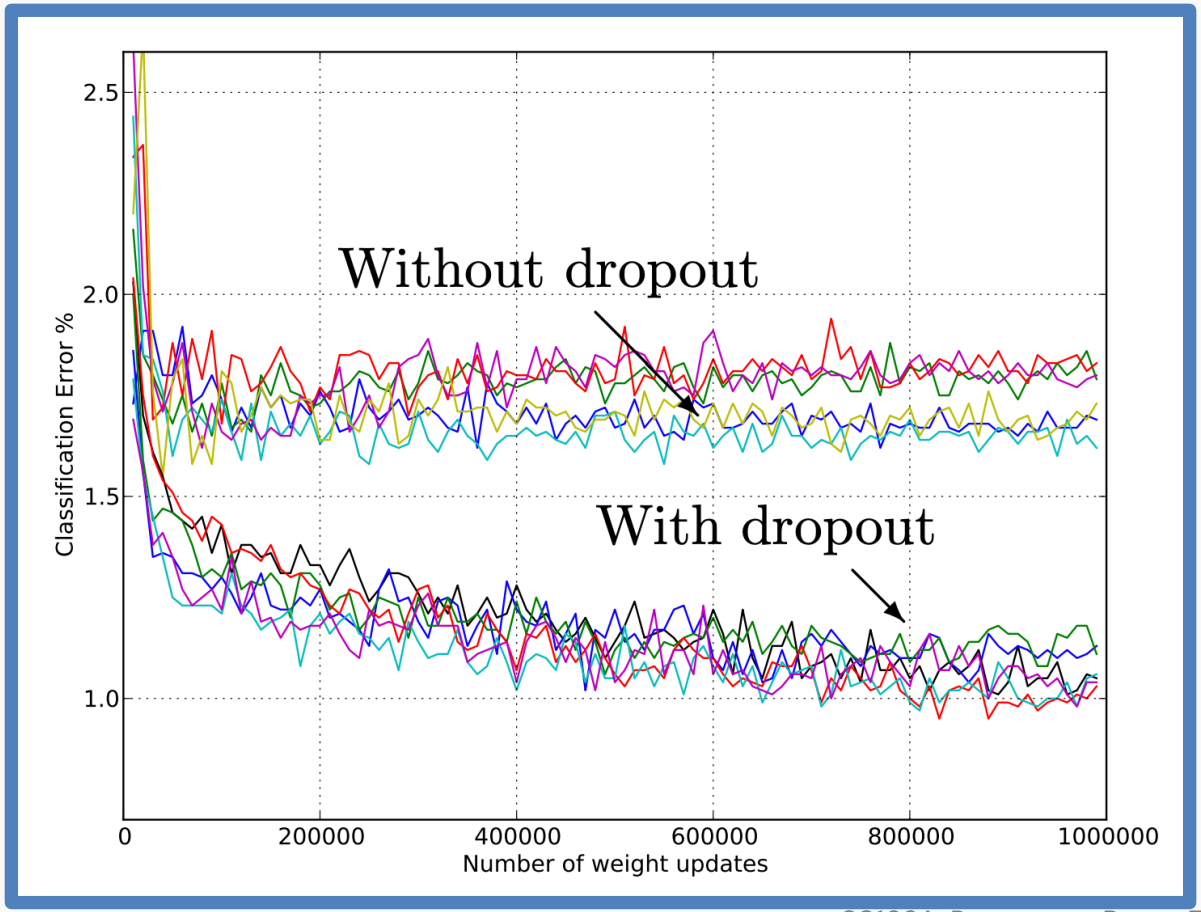
- Randomly sample a binary mask μ independently, where μ_i indicates if input/hidden node i is included
- Multiply output of node i with μ_i, and perform gradient update

Typically:

- Input nodes are included with prob=0.8 (as per original paper, but rarely used)
- Hidden nodes are included with prob=0.5

Dropout

- Widely used and highly effective
- Proposed as an alternative to ensemble methods, which is too expensive for neural nets



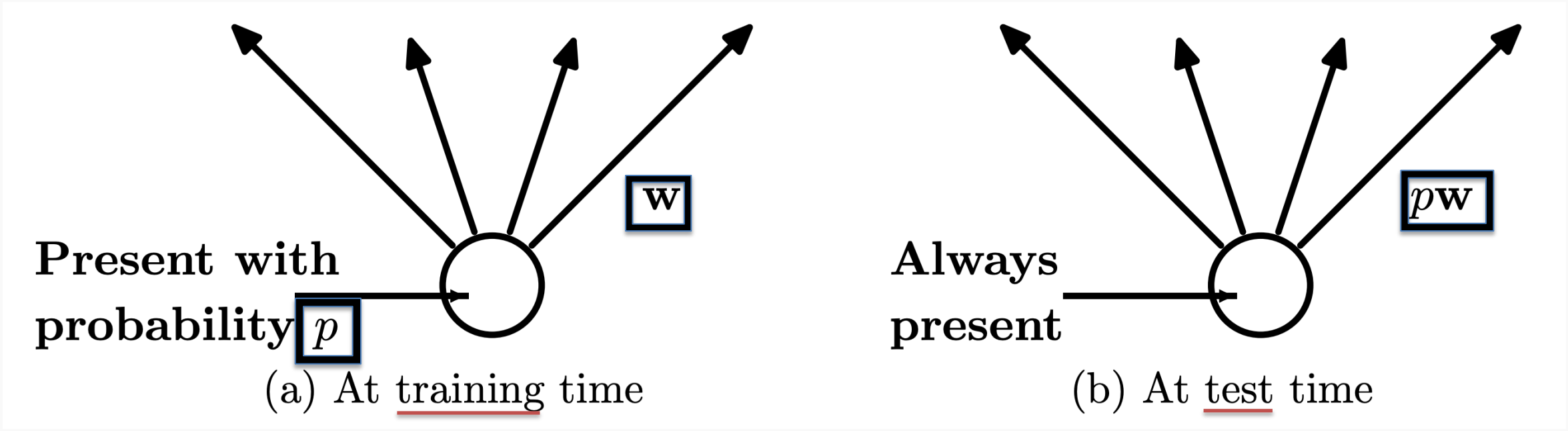
Test error for different architectures with and without dropout.

The networks have 2 to 4 hidden layers each with 1024 to 2048 units.

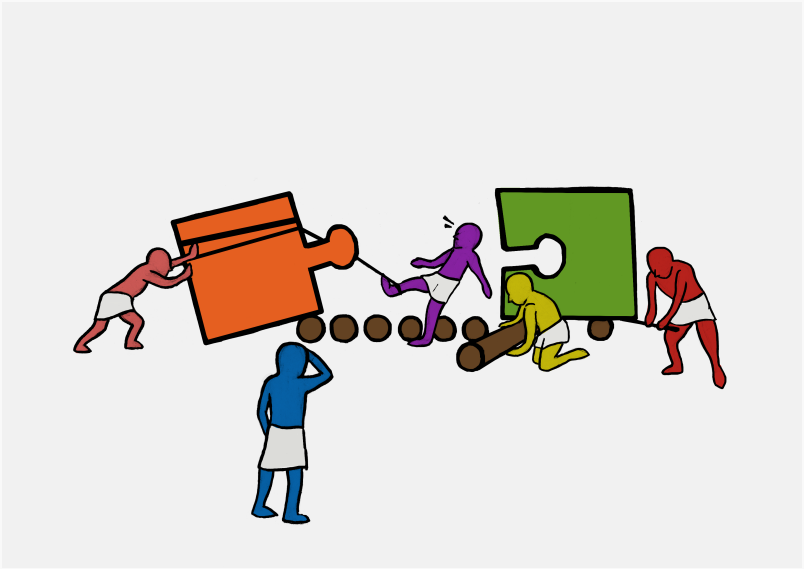
http://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf

Dropout: Prediction

- We can think of dropout as training many of sub-networks
- At **test time**, we can "aggregate" over these sub-networks by **reducing connection weights in proportion to dropout probability, *p***



NOTE: Dropouts can be used for **neural network inference** by dropping during predictions and predicting multiple times to get a distribution



Exercise: Dropout