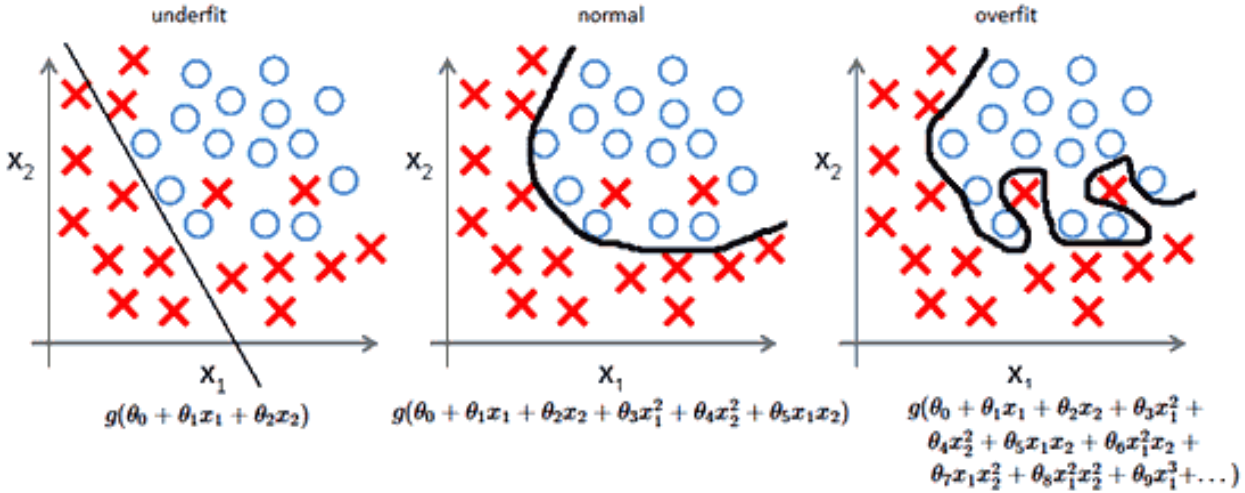
DNN Discussion 2

講者：Isaac

Outline

- Regularization
- Dropout
- Batch Normalization

Regularization

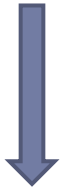


High variance

Regularization

- Add new term to loss/cost function
  - Optimize loss/cost function and regularization term at the same time

$$L_{new}(\theta) = L_{old}(\theta) + regularization\ term$$



Mean square, cross-entropy with softmax, ……

Regularization

- Recall linear algebra

$$\left\| \begin{bmatrix} 0.1 & 0.5 & -0.3 \\ -0.2 & 0.4 & 0.2 \\ -0.1 & 0.3 & 0.3 \end{bmatrix} \right\|_1 = |0.1| + |0.5| + |-0.3| + |-0.2| + |0.4| + |0.2| + |-0.1| + |0.3| + |0.3| = 1.2$$

L1 norm

$$\left\| \begin{bmatrix} 0.1 & 0.5 & -0.3 \\ -0.2 & 0.4 & 0.2 \\ -0.1 & 0.3 & 0.3 \end{bmatrix} \right\|_2 = (0.1)^2+(0.5)^2+(-0.3)^2+(-0.2)^2+(0.4)^2+(0.2)^2 +(-0.1)^2+(0.3)^2+(0.3)^2$$

L2 norm

Regularization

$$\sigma(W^L \dots \sigma(W^2\sigma(W^1x + b^1) + b^2) \dots + b^L)$$

$$L_{new}(\theta) = L_{old}(\theta) + \lambda\|\theta\|_1$$

$$\{W^1, W^2, \dots\dots, W^L\}$$

L1 regularization

$$L_{new}(\theta) = L_{old}(\theta) + \frac{\lambda}{2}\|\theta\|_2$$

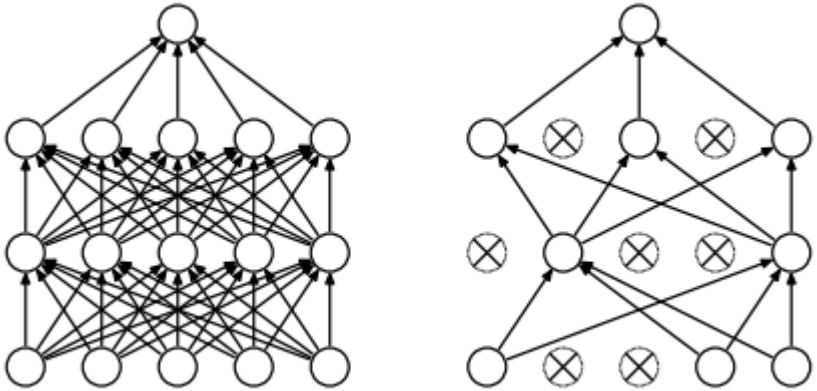
$$\{W^1, W^2, \dots\dots, W^L\}$$

L2 regularization

Make loss function small and weights close to zero

Dropout

- When training, each neuron has P% probability dropout
  - Each mini-batch would resample the dropout neurons

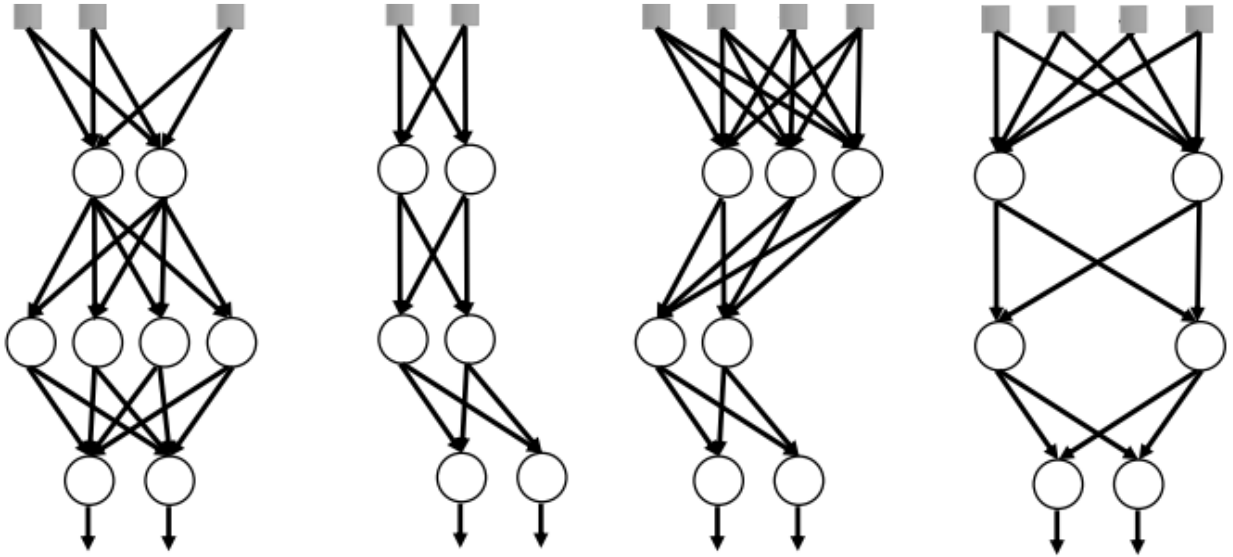


(a) Standard Neural Net

(b) After applying dropout.

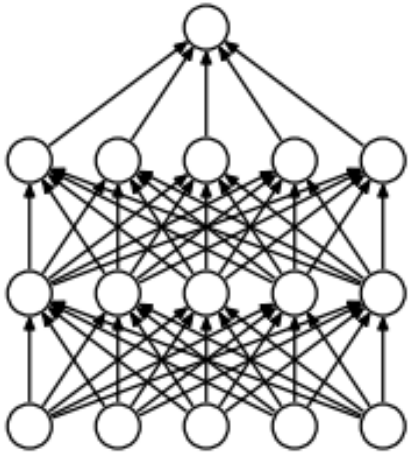
Dropout

Batch 1 Batch 2 Batch 3 Batch 4

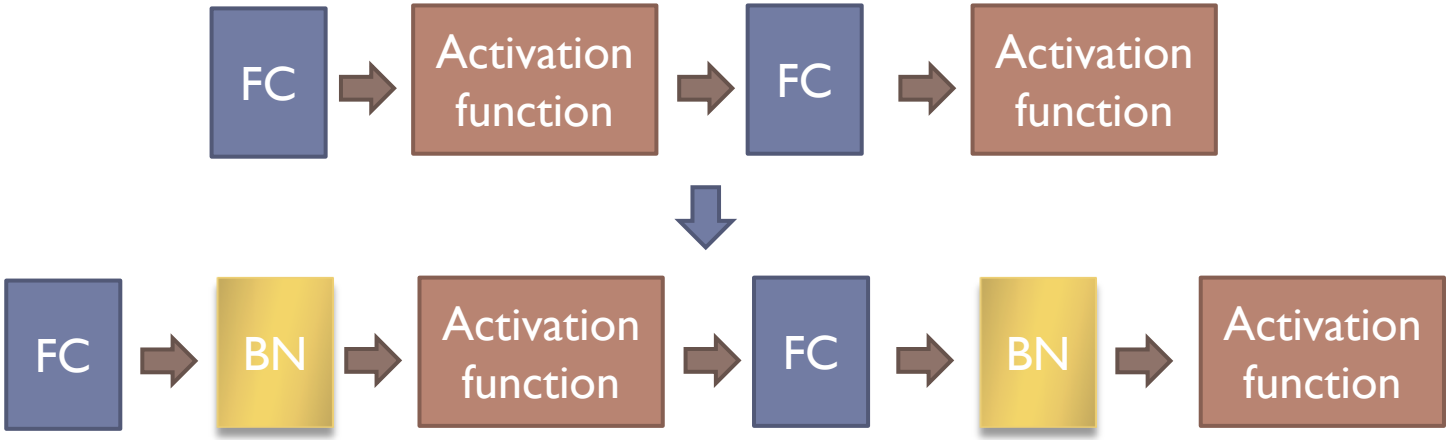


Dropout

- When testing
  - connect all of neurons



Batch Normalization



$$\sigma(W^L \dots \sigma(W^2\sigma(W^1x + b^1) + b^2) \dots + b^L)$$

$$\sigma(BN(W^L \dots \sigma(BN(W^2\sigma(BN(W^1x + b^1))) + b^2) \dots + b^L))$$

Batch Normalization

When training

$$\sigma(BN(W^L \dots \sigma(BN(W^2\sigma(BN(W^1x + b^1))) + b^2) \dots + b^L))$$

$$BN(x_i) = \gamma\left(\frac{x_i - \mu_B}{\sqrt{{\sigma_B}^2 + \varepsilon}}\right) + \beta$$

“1” (usually)

“0” (usually)

“0.001” (usually)

Linear transform to have zero mean and unit variance

- Accelerate training
- Less sensitive to initialization
- Improve regularization

Batch Normalization