بسم الله الرحمن الرحیم

Deep Learning

دانشگاه صنعتی مالک اشتر

Mohammad Ali Keyvanrad

Lecture 5:A Review of Artificial Neural Networks (4)

OUTLINE

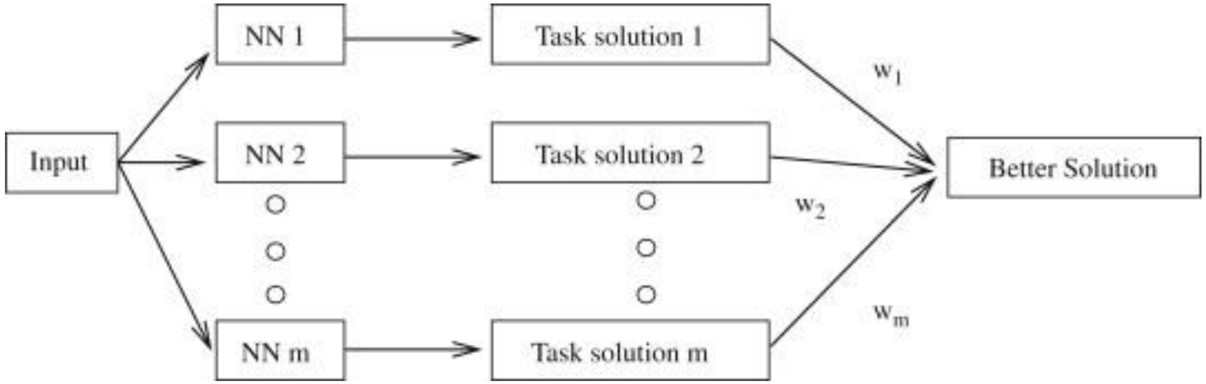
- Model Ensembles
- Regularization
- Dropout
- Regularization: A common pattern

OUTLINE

- **Model Ensembles**
- Regularization
- Dropout
- Regularization: A common pattern

Model Ensembles

- One reliable approach to improving the performance of Neural Networks
  - Train multiple independent models
  - At test time average their predictions
- **Disadvantage**
  - Take longer to evaluate on test example

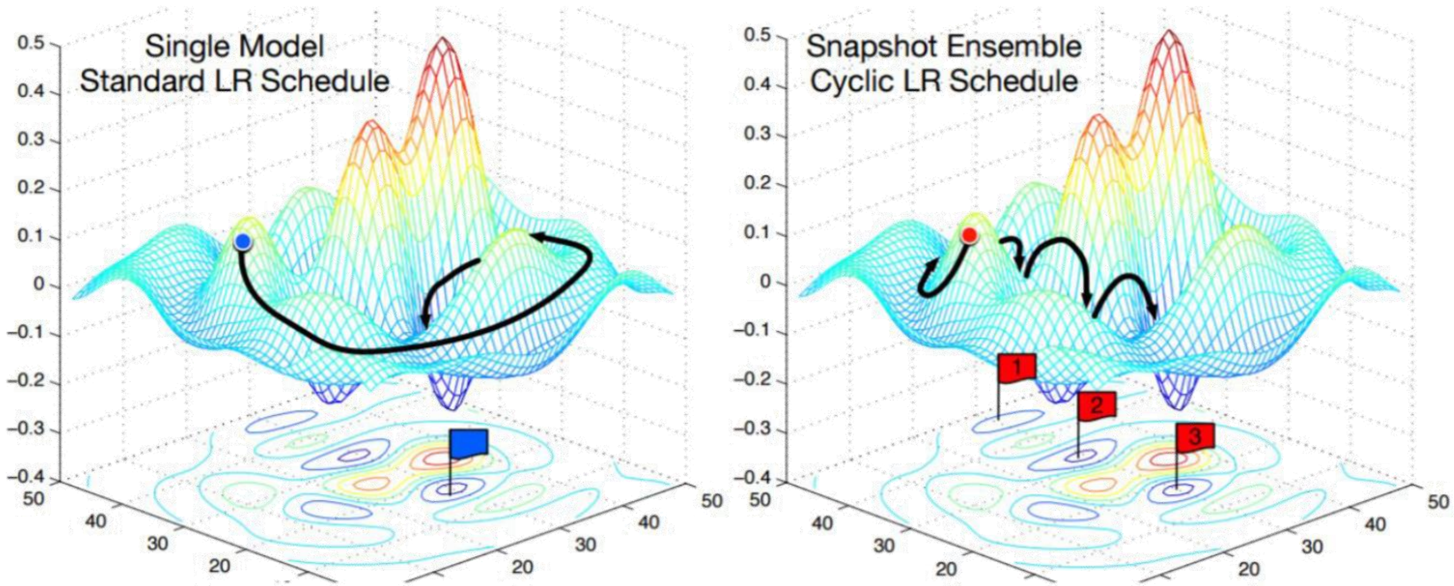


Model Ensembles

1. Same model, different initializations
   - Use cross-validation to determine the best hyperparameters
   - train multiple models with different random initialization
   - **Danger:** variety is only due to initialization.
2. Top models discovered during cross-validation.
   - Use cross-validation to determine the best hyperparameters
   - pick the top few (e.g. 10) models to form the ensemble
   - **Danger:** including suboptimal models

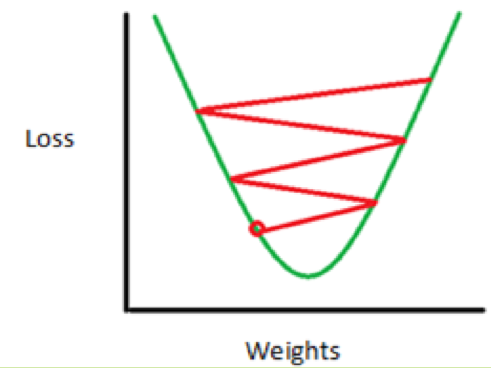
Model Ensembles

3. Different checkpoints of a single model
   - taking different checkpoints of a single network over time
     - when training is very expensive
   - **Danger:** lack of variety



Model Ensembles

4. Running average of parameters during training
   - **Averaging** the state of the network over last several iterations
     - Maintain a second copy of the network's weights with exponentially decaying sum of previous weights
   - **Smoothed version** of the weights over last few steps almost always achieves better validation error
   - **Why?**
   - Network is jumping around the mode
   - Higher chance of being nearer the mode

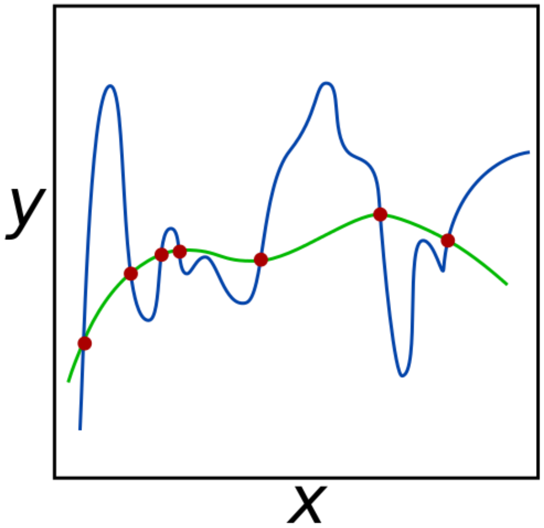


OUTLINE

- Model Ensembles
- **Regularization**
- Dropout
- Regularization: A common pattern

Regularization

- Definition
  - A process of introducing additional information in order to solve an ill-posed problem or to prevent overfitting.



- Usage
  - Learn simpler models
  - Induce models to be sparse
  - Introduce group structure into the learning problem
  - …

Regularization

- A regularization term (or regularizer) $R(f)$ is added to a loss function
  - V : loss function
  - f(x) : predicted value
  - λ : A parameter which controls the importance of the regularization term

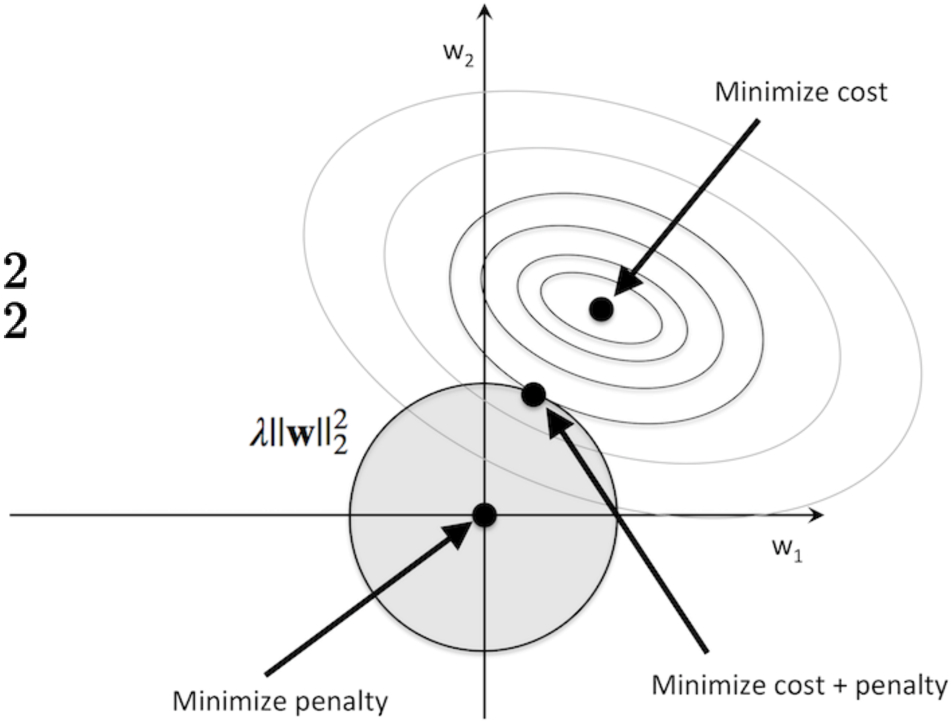
$$\min_f \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda R(f)$$

- Regularization introduces a penalty for exploring certain regions of the function space
  - used to build the model, which can improve generalization.

Controlling the capacity of Neural Networks to prevent overfitting

1. L2 regularization (Tikhonov regularization or Weight decay)
   – The most common form of regularization

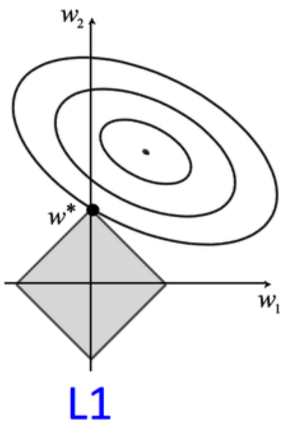
$$\min_f \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda \|w\|_2^2$$

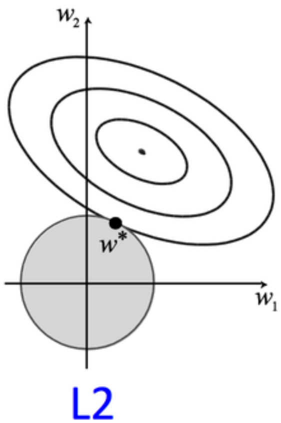


Controlling the capacity of Neural Networks to prevent overfitting

2. L1 regularization
   - Relatively common form of regularization
   - Leads the weight vectors to become sparse
     - Very close to exactly zero
   - Using only a sparse subset of their most important inputs

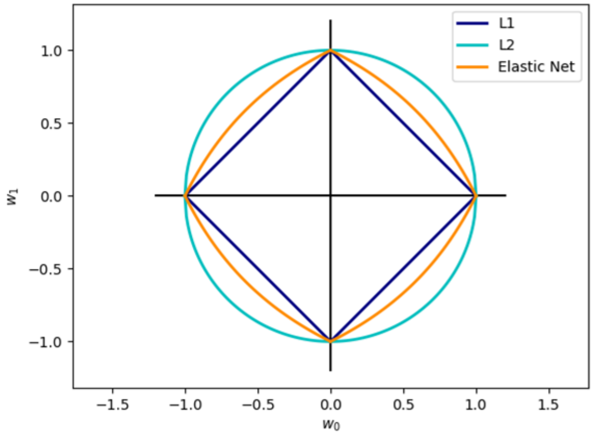
$$\min_f \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda \|w\|_1$$





Controlling the capacity of Neural Networks to prevent overfitting

3. **Elastic net regularization**
   - L1 + L2



4. **Max norm constraints**
   - Enforce an absolute upper bound on the magnitude of the weight vector for every neuron
   - Clamping the weight vector w of every neuron to satisfy $\| w \|_2 < c$
   - Network cannot "explode" even when the learning rates are set too high

5. **Dropout**

OUTLINE

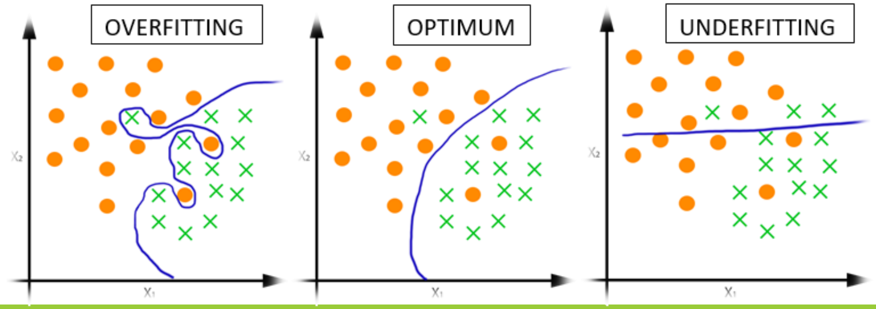
- Model Ensembles
- Regularization
- **Dropout**
- Regularization: A common pattern

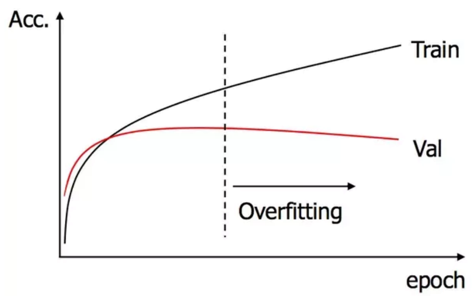
Dropout

- Dropout can be considered as a bagging technique
  – Averages over a large amount of models with tied parameters.
- Dropout can generate smoother objective surface
- A pretrain technique
  – we may pretrain a DNN using dropout to quickly find a relatively good initial point
  – Then fine-tune the DNN without using dropout

Dropout

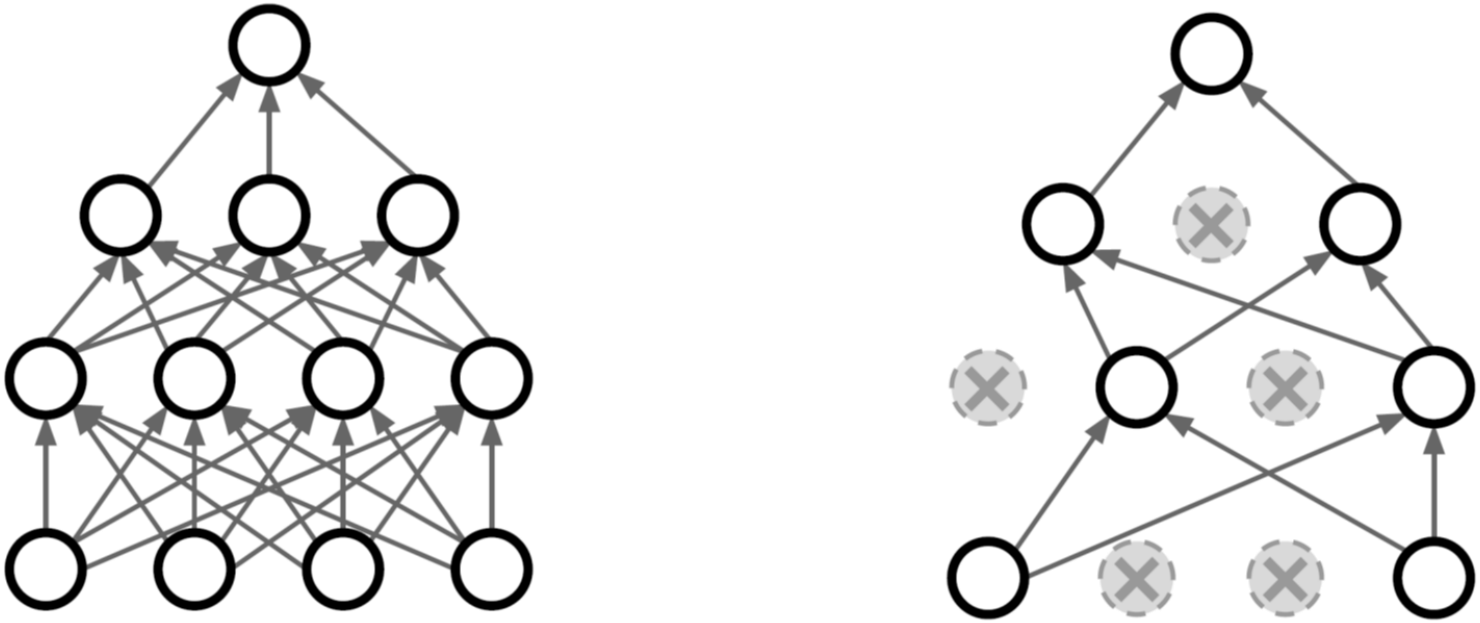
- Deep neural nets with a large number of parameters are very powerful machine learning systems
- Overfitting is a serious problem in Deep networks
- Large networks model ensembles are slow to use
  - Difficult to deal with overffitting by combining many different large neural nets
- Dropout is a technique for addressing this problem.





Dropout

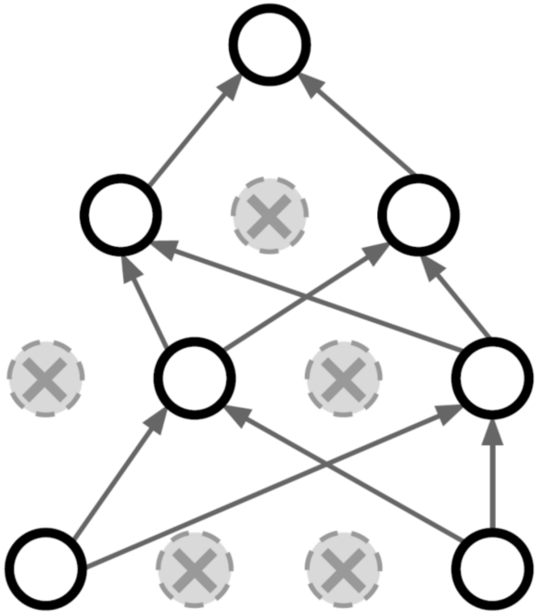
- The term **dropout** refers to dropping out units
  - Randomly set some neurons to zero
  - Probability of retaining is a hyperparameter
    - $p = 0.5$ is common

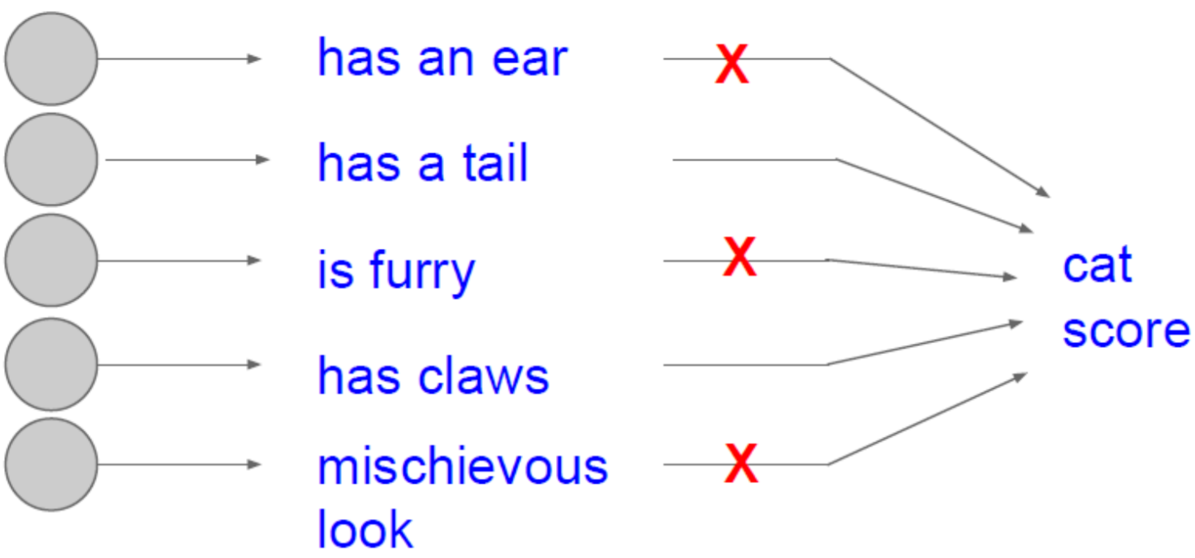


[Srivastava et al, 2014]

Dropout

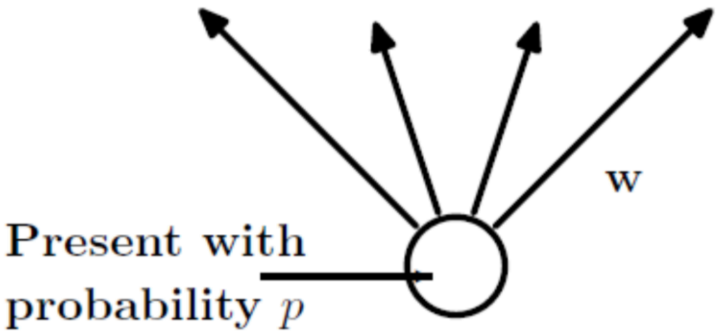
- How can this possibly be a good idea?
  - Forces the network to have a redundant representation
  - Prevents co-adaptation of features

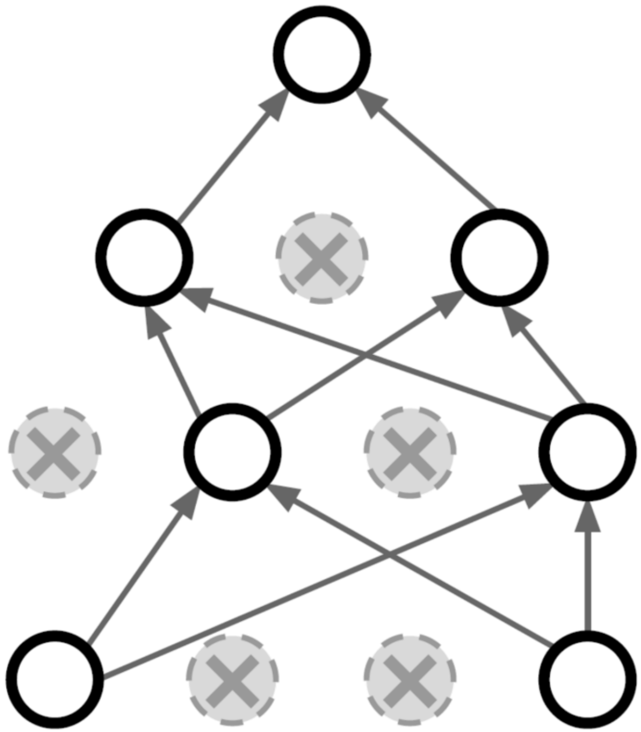




Dropout

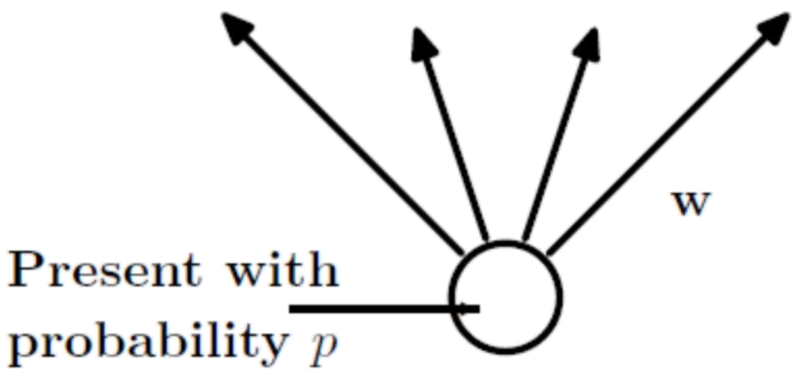
- How can this possibly be a good idea?
  - A neural net with n units, can be seen as a collection of 2^n possible thinned neural networks
    - A large ensemble of models
    - These networks all share weights
    - Each binary mask is one model
  - An FC layer with 4096 units
    - $2^{4096} \sim 10^{1233}$ possible masks





Dropout

- In the simplest case, each unit is retained with a fixed probability p independent of other units.
- p can be chosen using a validation set or can simply be set at 0.5.
- For the input units, however, the optimal probability of retention is usually closer to 1 than to 0.5.



Dropout

- At test time
  - It is not feasible to explicitly average the predictions from exponentially many thinned models

Output (label) Input (image)

$$\boxed{y} = f_W(\boxed{x}, \boxed{z})$$

Random mask

  - Want to "average out" the randomness at test-time

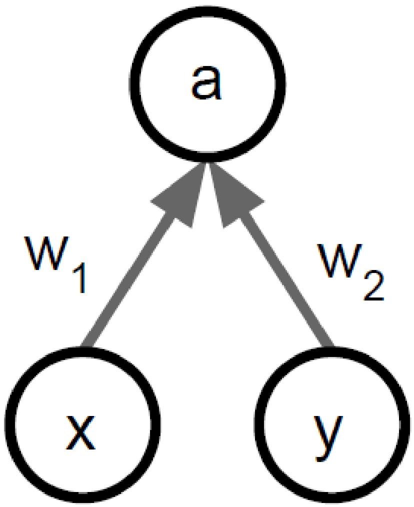
$$y = f(x) = E_z\big[f(x, z)\big] = \int p(z) f(x, z) dz$$

  - But this integral seems hard ...

Dropout

- Want to approximate the integral
  - Consider a single neuron

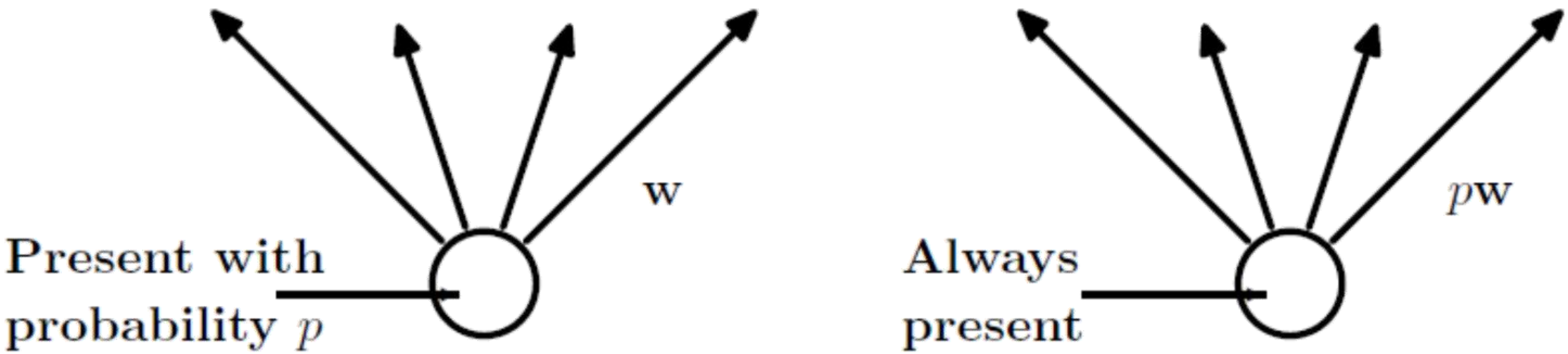
$$y = f(x) = E_z\big[f(x,z)\big] = \int p(z) f(x,z)\,dz$$



$$\begin{aligned} E\big[a\big] &= \frac{1}{4}(w_1 x + w_2 y) + \frac{1}{4}(w_1 x + 0y) \\ &\quad + \frac{1}{4}(0x + 0y) + \frac{1}{4}(0x + w_2 y) \\ &= \frac{1}{2}(w_1 x + w_2 y) \end{aligned}$$

Dropout

- Idea
  - Use a single neural net at test time without dropout
  - Multiply each weight by dropout probability



(a) At training time

(b) At test time

Dropout (MNIST)

Method	Unit Type	Architecture	Error %
Standard Neural Net (Simard et al., 2003)	Logistic	2 layers, 800 units	1.60
SVM Gaussian kernel	NA	NA	1.40
Dropout NN	Logistic	3 layers, 1024 units	1.35
Dropout NN	ReLU	3 layers, 1024 units	1.25
Dropout NN + max-norm constraint	ReLU	3 layers, 1024 units	1.06
Dropout NN + max-norm constraint	ReLU	3 layers, 2048 units	1.04
Dropout NN + max-norm constraint	ReLU	2 layers, 4096 units	1.01
Dropout NN + max-norm constraint	ReLU	2 layers, 8192 units	0.95
Dropout NN + max-norm constraint (Goodfellow et al., 2013)	Maxout	2 layers, (5×240) units	0.94
DBN + finetuning (Hinton and Salakhutdinov, 2006)	Logistic	500-500-2000	1.18
DBM + finetuning (Salakhutdinov and Hinton, 2009)	Logistic	500-500-2000	0.96
DBN + dropout finetuning	Logistic	500-500-2000	0.92
DBM + dropout finetuning	Logistic	500-500-2000	**0.79**

Dropout (TIMIT)

Method	Phone Error Rate%
NN (6 layers) (Mohamed et al., 2010)	23.4
Dropout NN (6 layers)	21.8
DBN-pretrained NN (4 layers)	22.7
DBN-pretrained NN (6 layers) (Mohamed et al., 2010)	22.4
DBN-pretrained NN (8 layers) (Mohamed et al., 2010)	20.7
mcRBM-DBN-pretrained NN (5 layers) (Dahl et al., 2010)	20.5
DBN-pretrained NN (4 layers) + dropout	**19.7**
DBN-pretrained NN (8 layers) + dropout	**19.7**

Table 7: Phone error rate on the TIMIT core test set.

OUTLINE

- Model Ensembles
- Regularization
- Dropout
- **Regularization: A common pattern**

Regularization: A common pattern

- **Training:** stochastic behavior in the forward pass
  - Add some kind of randomness

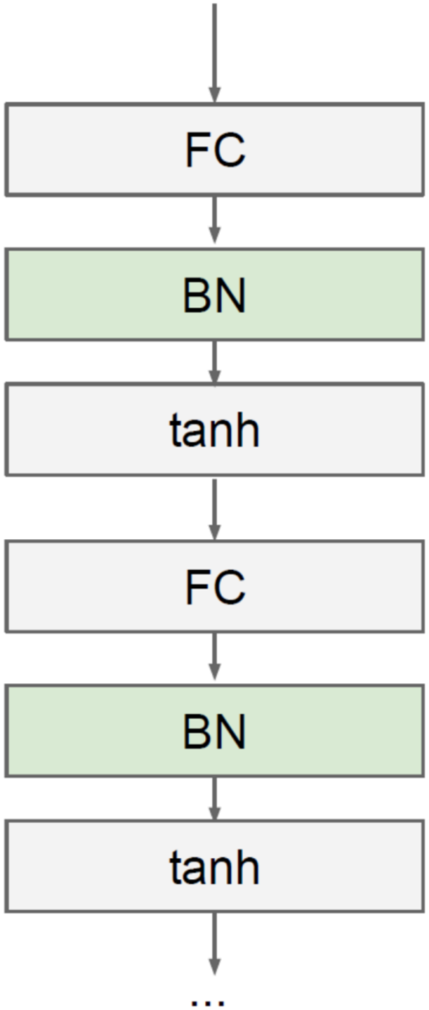
$$y = f_W(x, z)$$

- **Testing:** the noise is marginalized
  - Average out randomness
    - Analytically: as is the case with dropout when multiplying by p
    - Numerically: e.g. via sampling, by performing several forward passes with different random decisions and then averaging them

$$y = f(x) = E_z\left[f(x, z)\right] = \int p(z) f(x, z) dz$$

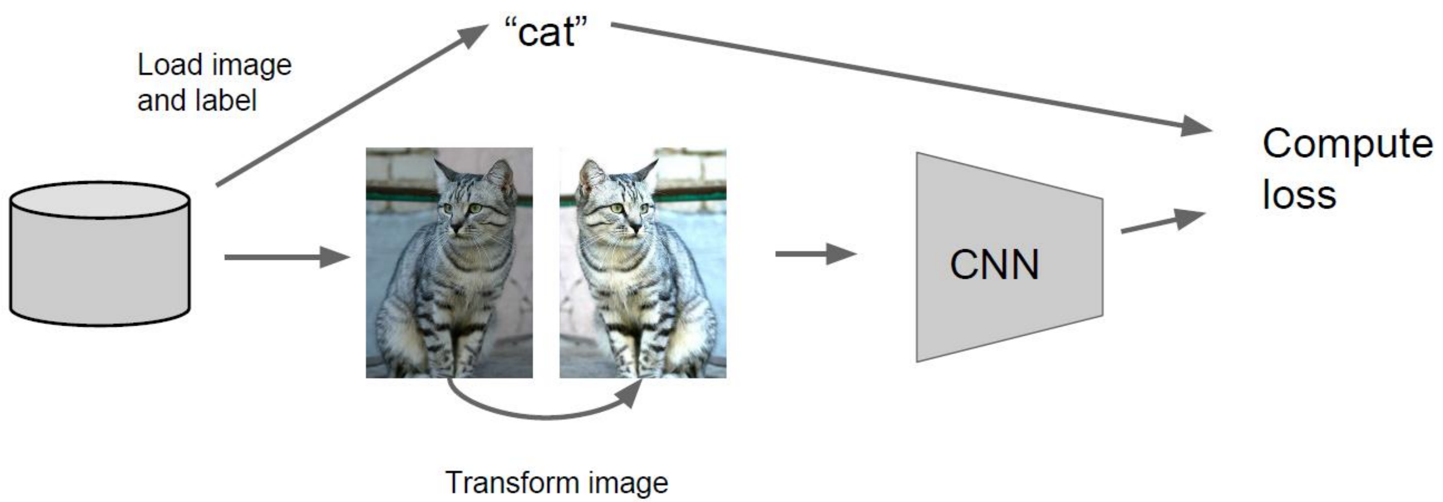
Regularization: A common pattern

- Example: Batch Normalization
  - Training (kind of randomness)
    - Normalize using stats from random minibatches
  - Testing (Average out randomness)
    - Use fixed stats to normalize



Regularization: A common pattern

- Example: Data Augmentation
  - Training (kind of randomness)
    - Transform image (Horizontal Flips, Random crops, …)
  - Testing (Average out randomness)
    - Sample random Transform



Regularization: A common pattern

- ResNet
  - Training : sample random crops / scales
    - Pick random L in range [256, 480]
    - Resize training image, short side = L
    - Sample random 224 x 224 patch
  - Testing : average a fixed set of crops
    - Resize image at 5 scales: {224, 256, 384, 480, 640}
    - For each size, use 10 224 x 224 crops: 4 corners + center, + flips

Regularization: A common pattern

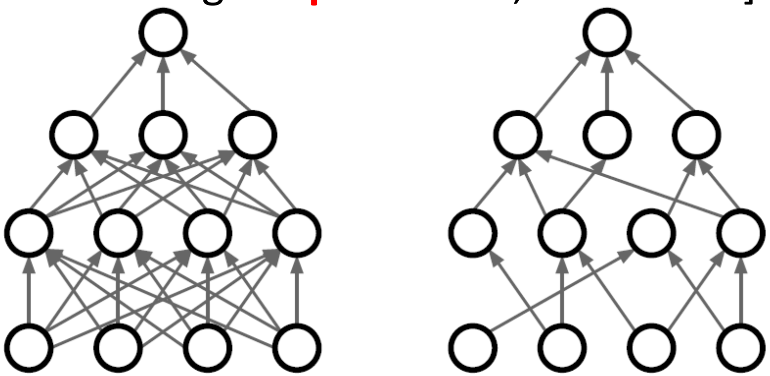
Get creative for your problem!

- Random mix/combinations of
  - Translation
  - contrast and brightness
  - rotation
  - stretching
  - shearing
  - lens distortions, …

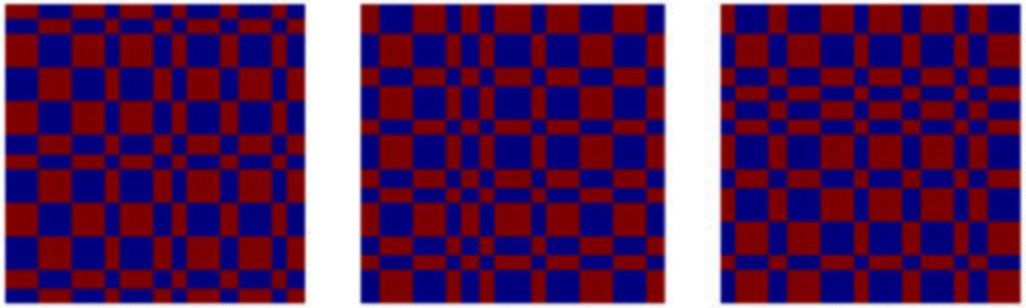
Regularization: A common pattern

- Other Examples

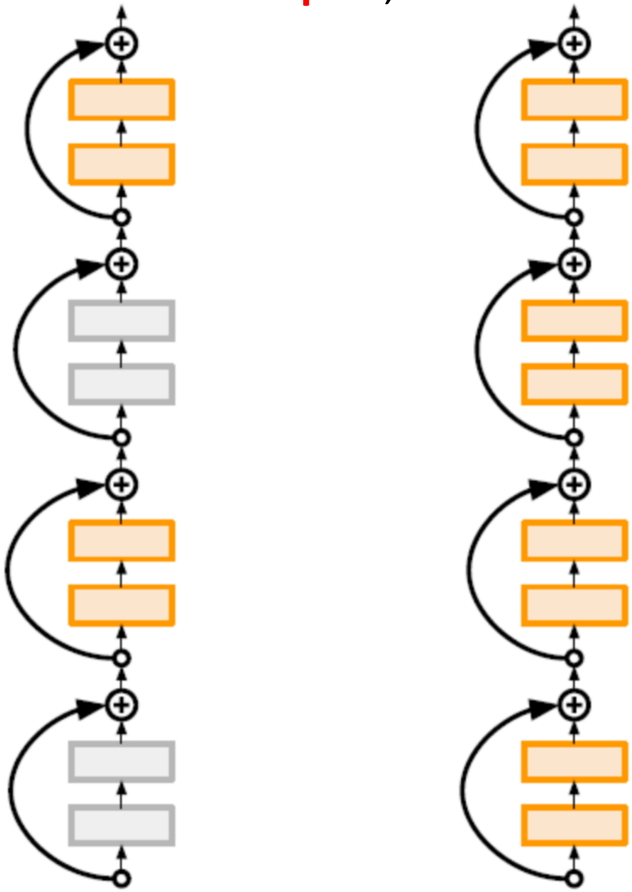
[Wan et al, "Regularization of Neural Networks using **DropConnect**", ICML 2013]



[Graham, "**Fractional Max Pooling**", arXiv 2014]



Huang et al, "Deep Networks with **Stochastic Depth**", ECCV 2016



References

- Stanford "Convolutional Neural Networks for Visual Recognition" course (Neural Nets notes 2)
- Stanford "Convolutional Neural Networks for Visual Recognition" course (Neural Nets notes 3)
- Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." Journal of machine learning research 15.1 (2014).
- https://en.wikipedia.org/wiki/Overfitting
- https://en.wikipedia.org/wiki/Regularization_(mathematics)

پیامبر اکرم (ص):

العبادَةُ سَبْعُونَ جُزْءاً أفضَلُها طَلَبُ الْحَلال.

عبادت هفتَاد قسمت دارد که برترین آنها طَلب روزی حلال است.

There are seventy branches of worship, the best of which is seeking for lawful sustenance.

تهذیب، ج ۶، ص ۳۲۴