Applied Deep Learning

Practical Tips

March 17th, 2020 **http://adl.miulab.tw**

國立臺灣大學
National Taiwan University

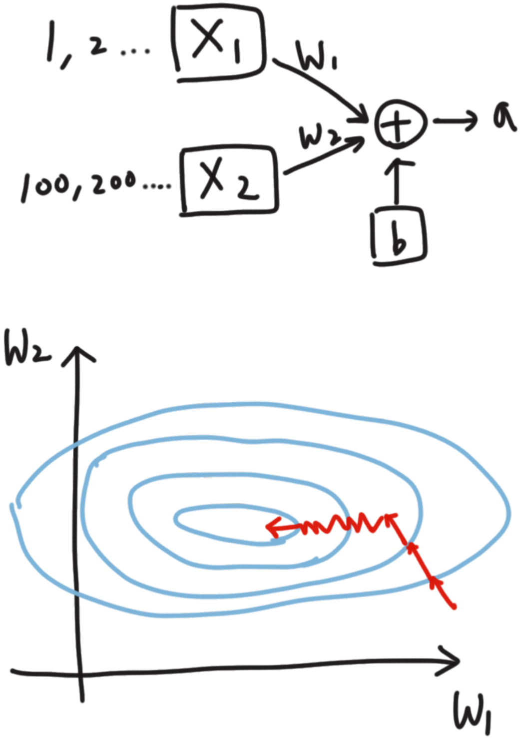
2

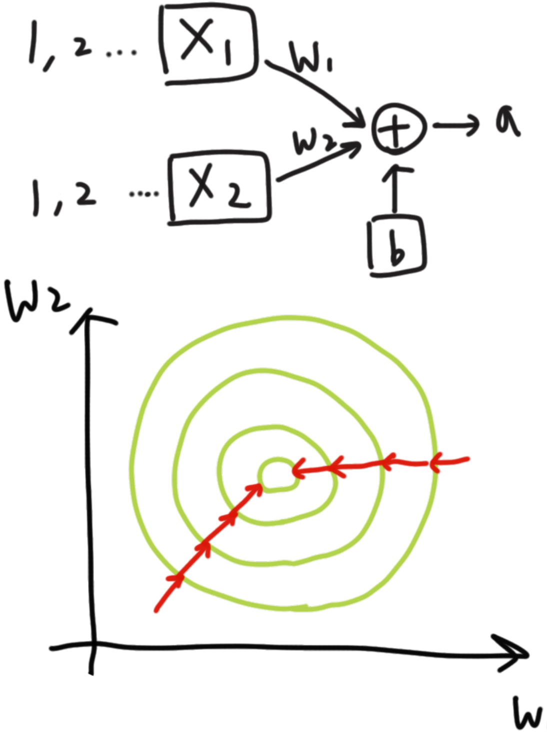
Mini-Batch Training

Feature Scaling

3

- Idea: make sure features are on the same scale

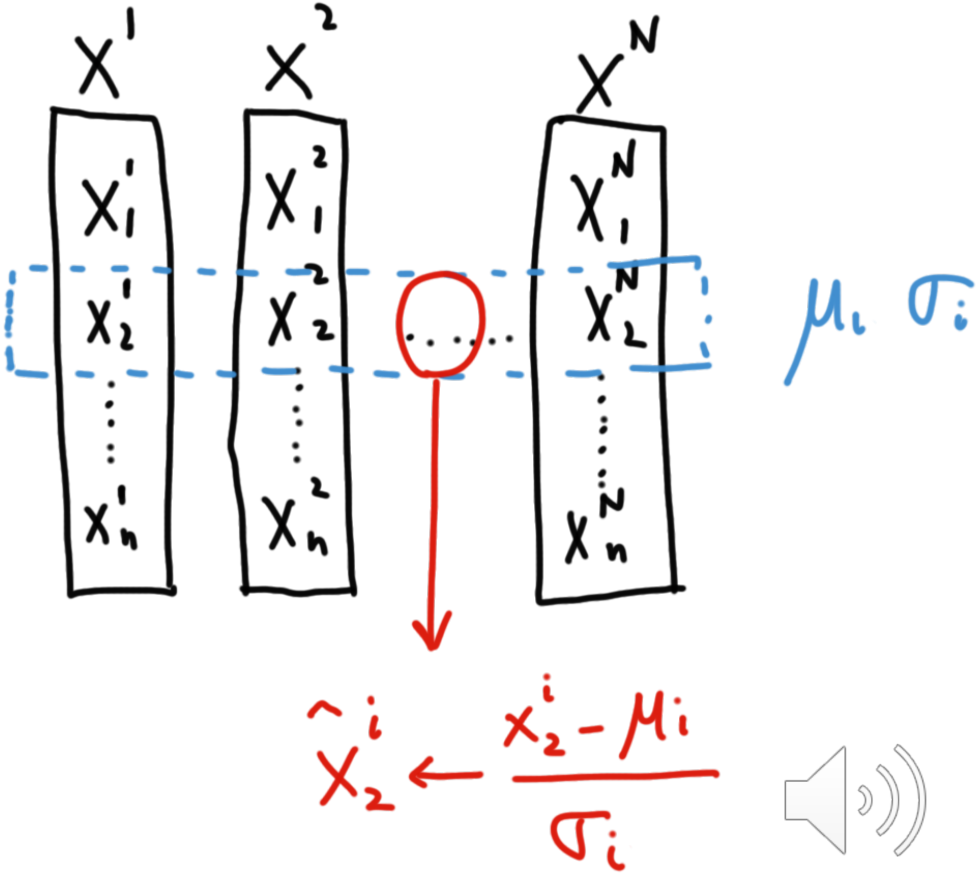




4

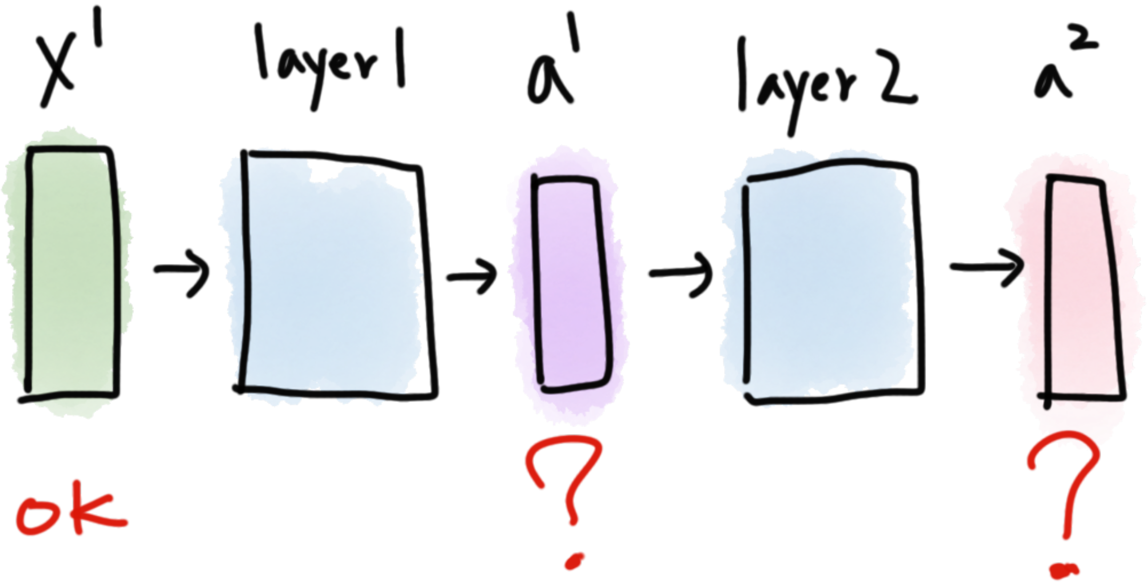
Feature Scaling

- for each dimension, compute mean and standard deviation
- the means of normalized feature vectors are all 0 and the variances are all 1

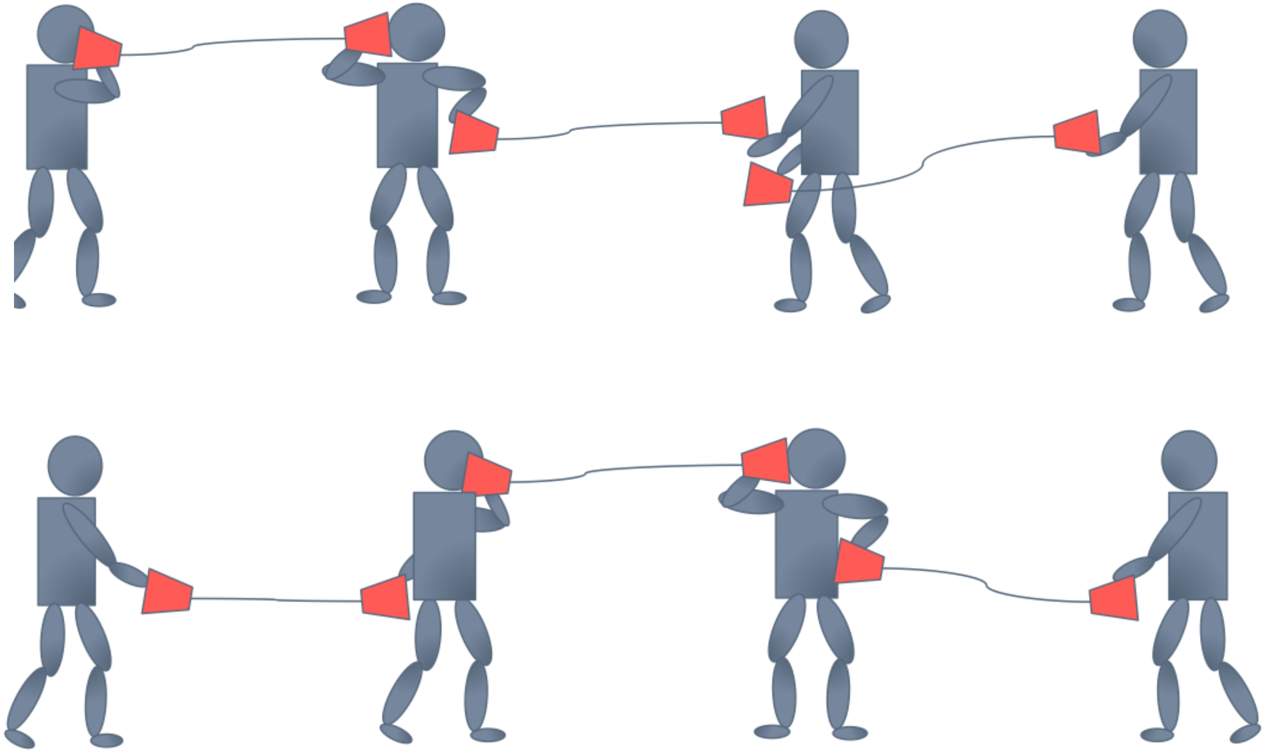


Hidden States as Features

- statistics of hidden states keep changing during training



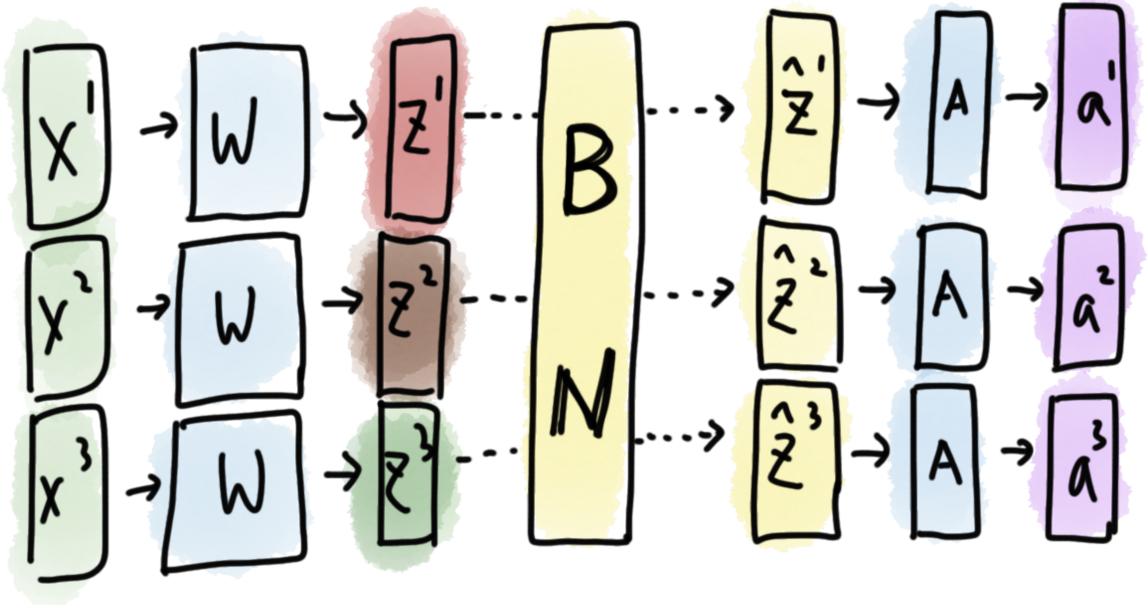
Internal Covariate Shift



7

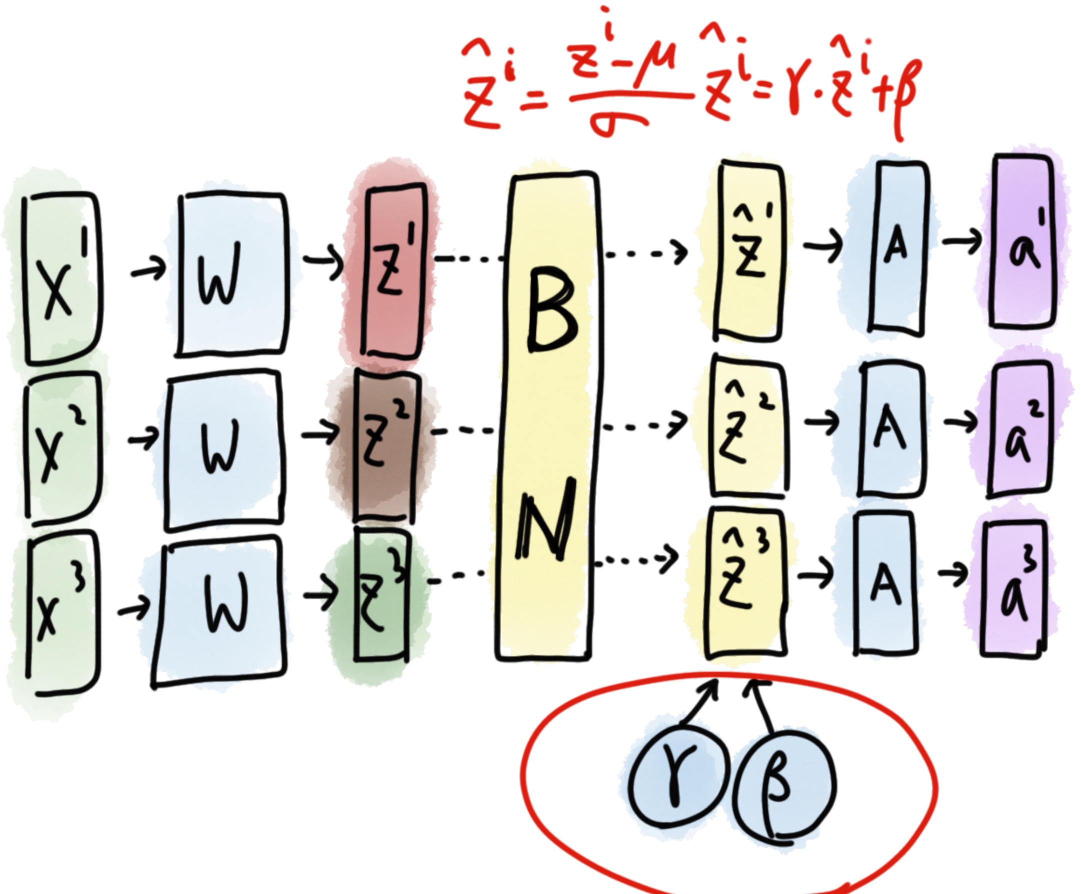
Batch Normalization

$$\hat{z}^i = \frac{z^i - \mu}{\sigma}$$



8

Batch Normalization

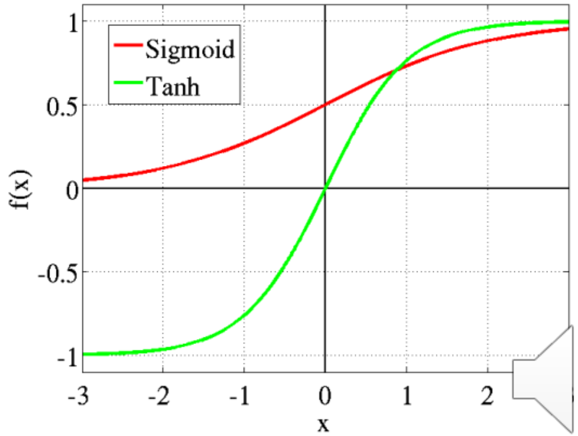


Batch Normalization

- learnable parameters γ and β to rescale and reshift distribution to preserve model capacity
- do not have “batch” in testing phase
- Ideal solution: computing mean and variance based on the whole training set
- practical solution: computing moving average of mean and variance of batches after convergence

Closer Look

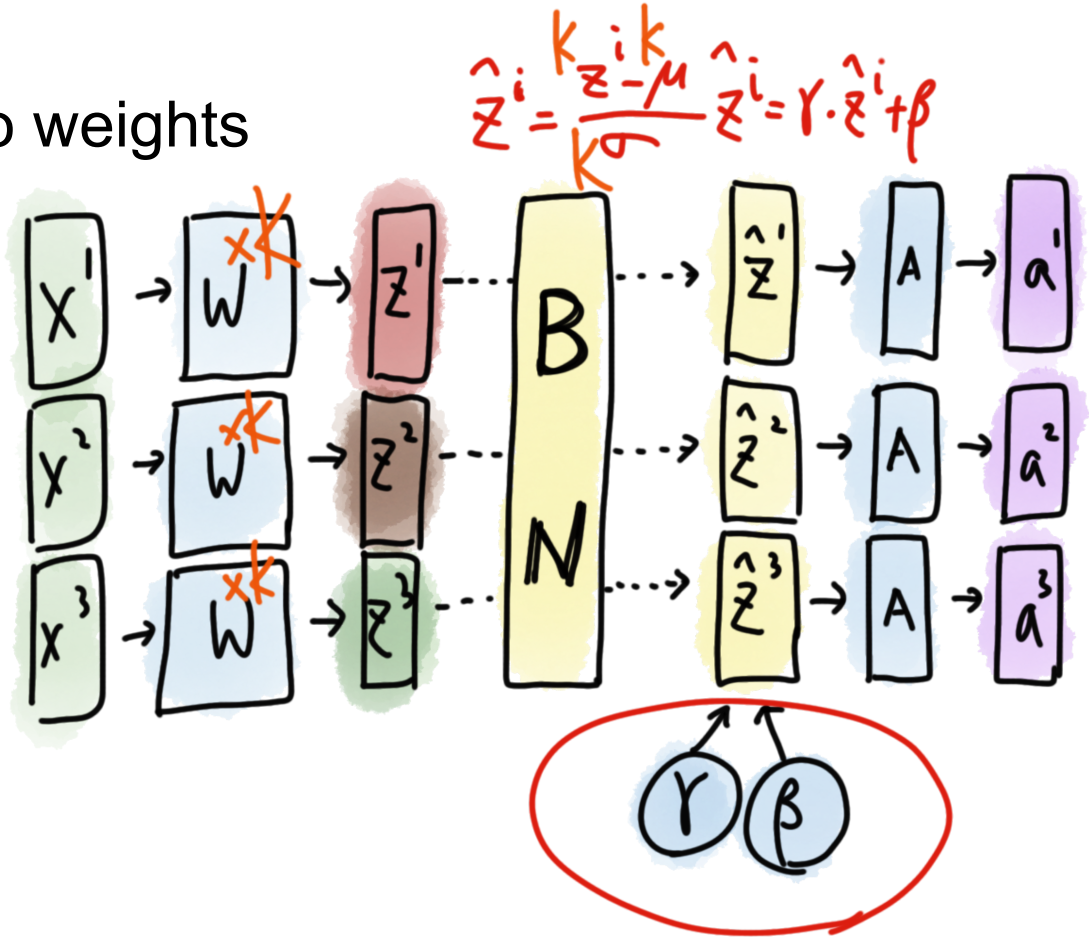
- Interval Covariate Shift?
- usually apply before activation function
- avoid exploding/vanishing gradients, especially for sigmoid and tanh activation functions
- batch size should be large
- not suitable for dynamic structure



11

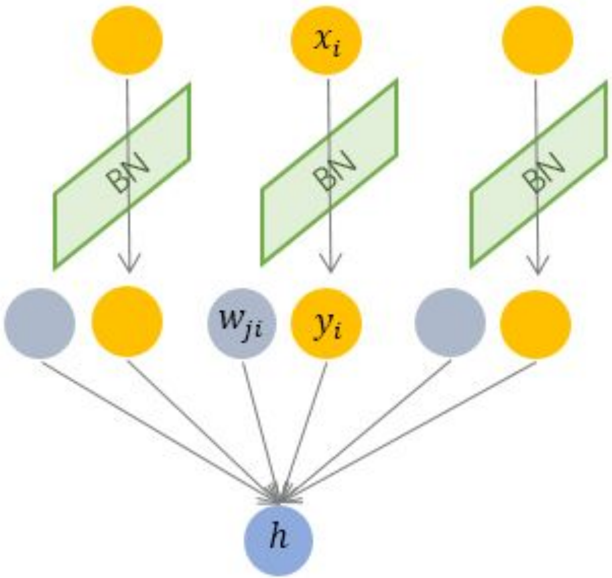
Closer Look

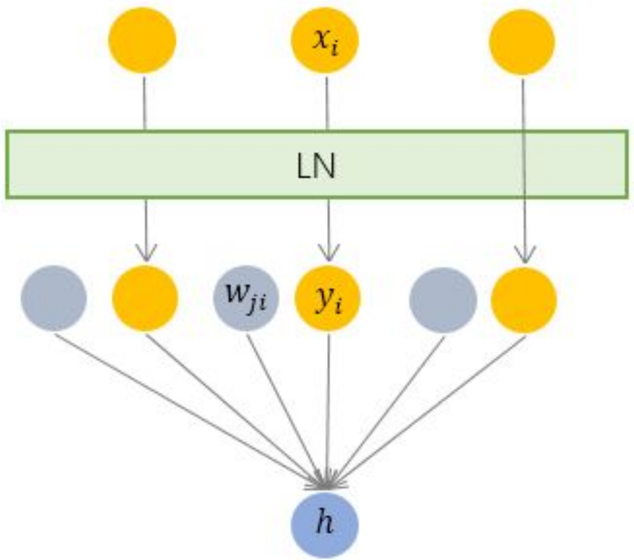
- Unsensitive to weights



12 Layer Normalization

- can be used in (1) small batch scenario, even a single data sample and (2) dynamic network structures like RNN





More Kinds of Normalization

- Weight Normalization
- Instance Normalization
- Group Normalization
- Spectral Normalization

How big is your batch size?

- **Intuitive idea: my GPU memory is enough → increase the batch size**
- …Is it correct?

How big is your batch size?

- The paper titled "*Revisiting Small Batch Training for Deep Neural Networks*"
- Quote from the paper: *"In all cases the best results have been obtained with batch sizes m= 32 or smaller, often as small as m= 2 or m= 4. With BN and larger datasets, larger batch sizes can be useful, up to batch size m= 32 or m= 64."*

Learning Rate

- Intuitive/simple idea: reduce the learning rate by some factor every few epochs.
  - At the beginning, we are far from the destination, so use a larger learning rate
  - After several epochs, as we get closer to the destination, reduce the learning rate
- Better idea: give different parameters different learning rates
  - Adaptive optimizers: Adagrad, RMSprop, Adam etc.

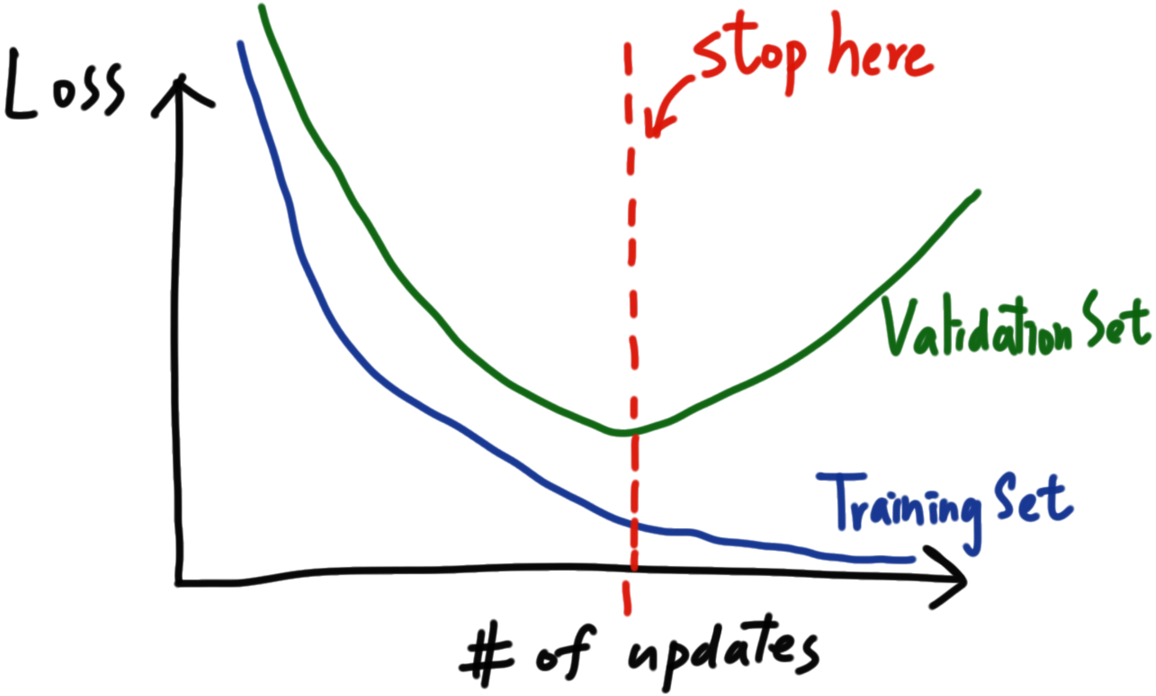
17

Generalization

To Prevent Overfitting

Early Stopping

- Q: how many epochs should we train the models?



Weight Decay

- Smaller weights are preferred. Why?
- (x, y) vs (x', y) $where\ x' = x + \varepsilon$
- $z = w \cdot x$
- $z' = w \cdot x' = w \cdot (x + \varepsilon) = z + w \cdot \varepsilon$
- To minimize the effect of noise, we want weights close to zero.

Regularization

- Add a weight constraint term into the objective

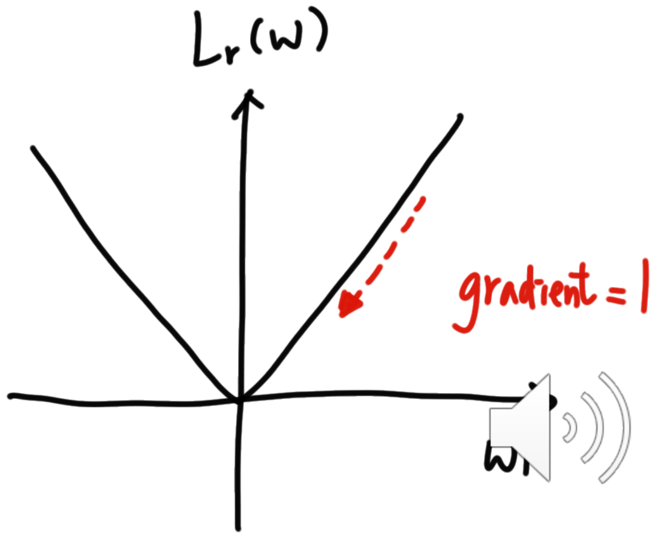
$$L' = L + L_r(w)$$

- By minimizing the loss, the weights will become smaller.

L1 Regularization

$$L_r(w) = \lambda \sum_{i=1}^{N} |w_i|$$

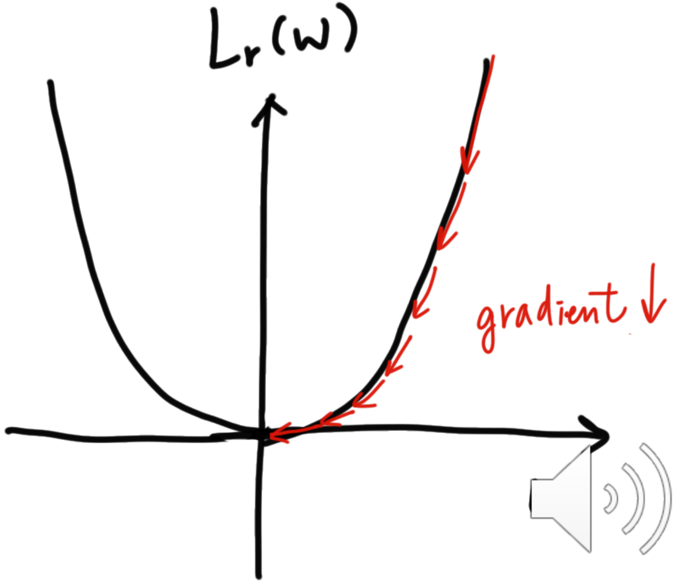
- feature selection/parameter sparsity



L2 Regularization

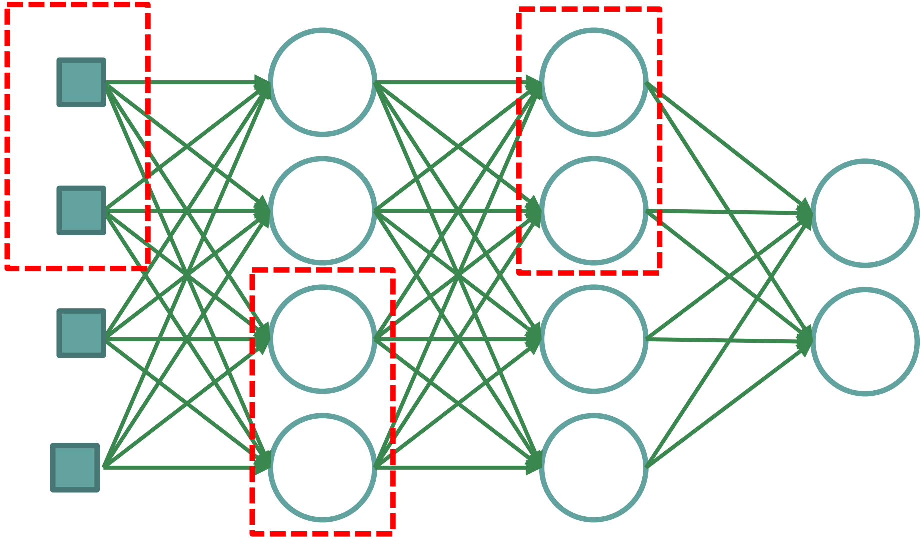
$$L_r(w) = \lambda \sum_{i=1}^{N} w_i^2$$

- "One should always try L2 first."
- encourage all weights to be small



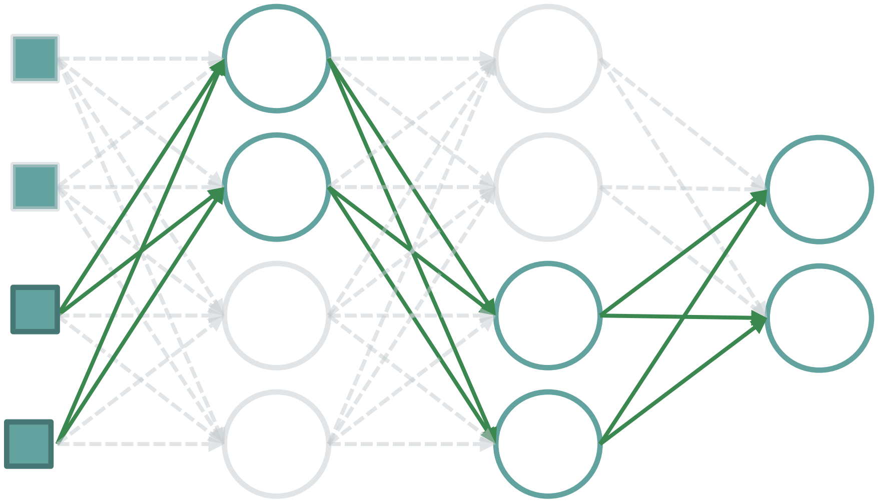
Dropout

- In each iteration of training, each neuron has p% probability to dropout



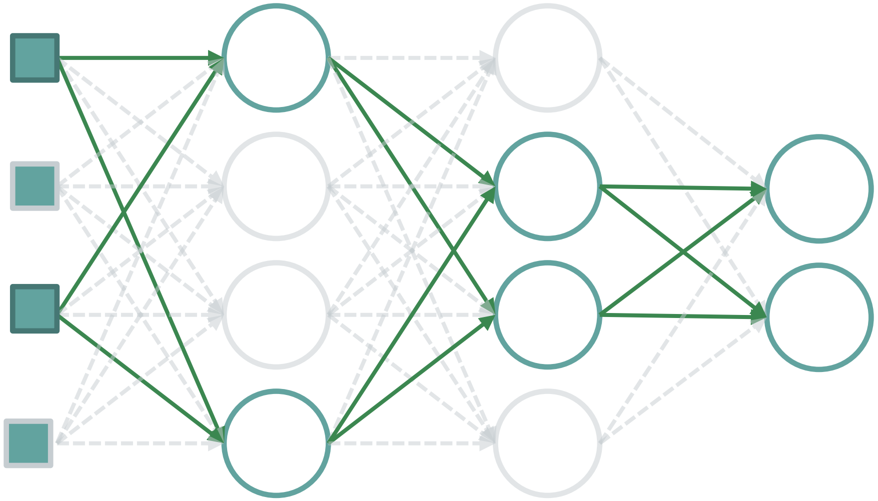
Dropout

- In each iteration of training, each neuron has p% probability to dropout



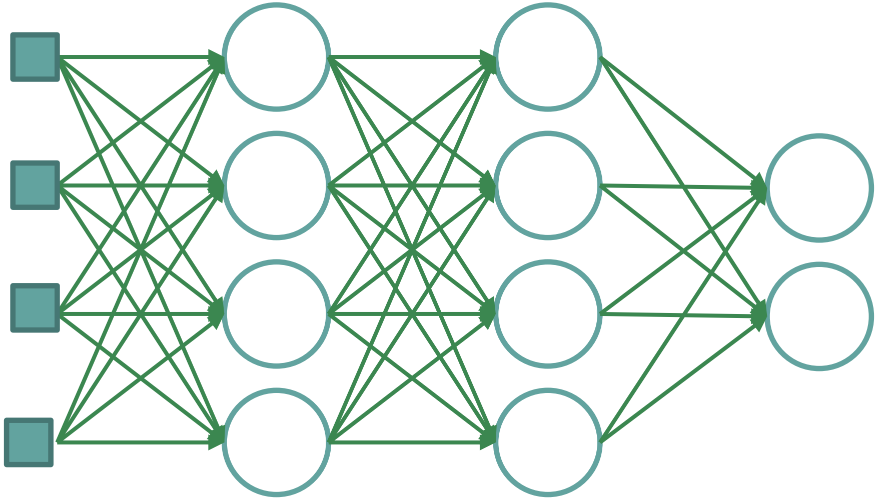
Dropout

- For each iteration, we resample the dropout neurons
- Using a new network for training



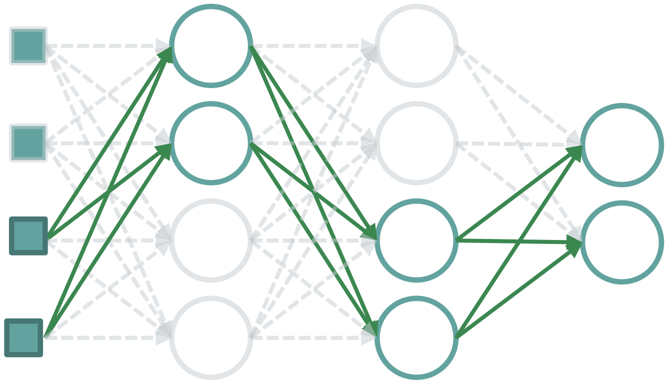
Dropout

- When testing, no dropout and all the weights times (100-p)%
- Why?

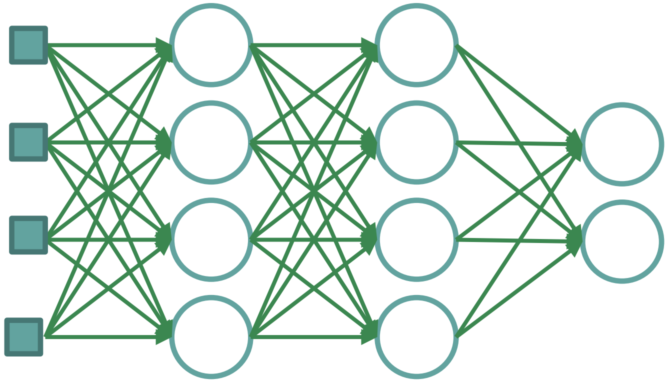


Dropout

- Assume p = 0.5



$$z = w \cdot x$$

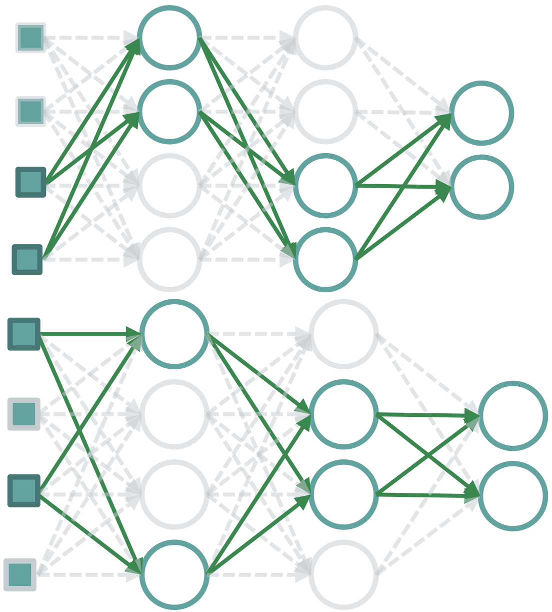


$$z' \approx 2z$$

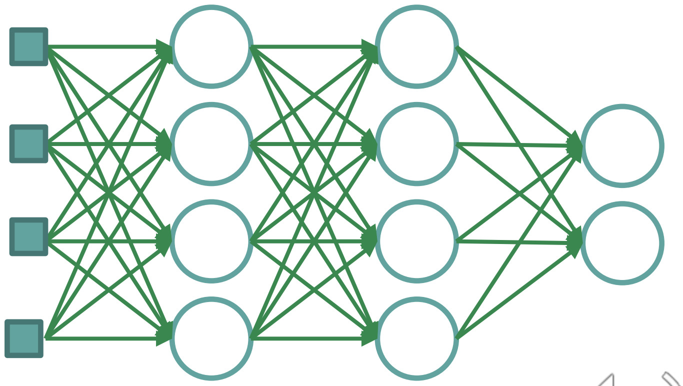
$$z' \cdot (1 - p) \approx z$$

Dropout

- Ensemble



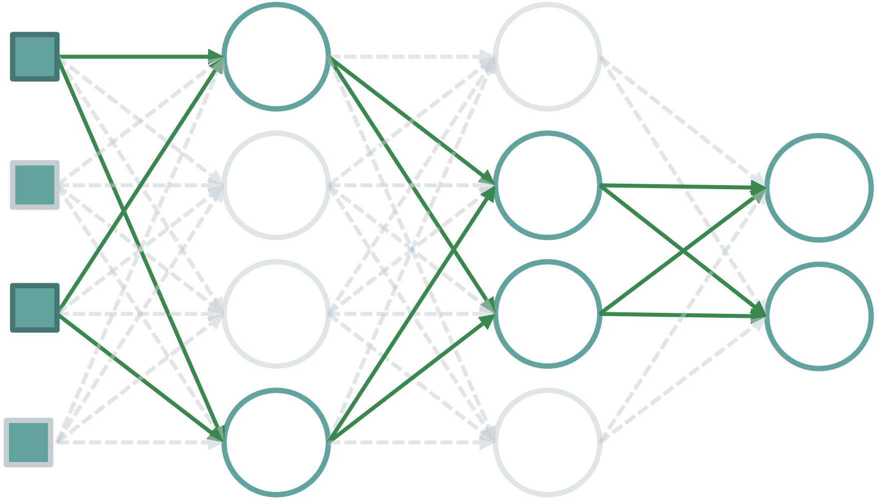
Train a bunch of networks



Average the results

Dropout

- depress the capacity → unleash the potential
- your teammate is a free rider → you need to work harder



References

- https://www.csie.ntu.edu.tw/~yvchen/f106-adl/doc/171116+171120_Tip.pdf
- https://zhuanlan.zhihu.com/p/33173246
- https://gab41.lab41.org/batch-normalization-what-the-hey-d480039a9e3b
- https://arxiv.org/pdf/1803.08494.pdf
- https://arxiv.org/pdf/1804.07612.pdf
- http://speech.ee.ntu.edu.tw/~tlkagk/courses/MLDS_2015_2/Lecture/Deep%20More%20(v2).pdf