Deep Learning Strategies

Mass Street University

PER EDUCATIONEM PROGRESSUS

Areas we need to improve in deep learning

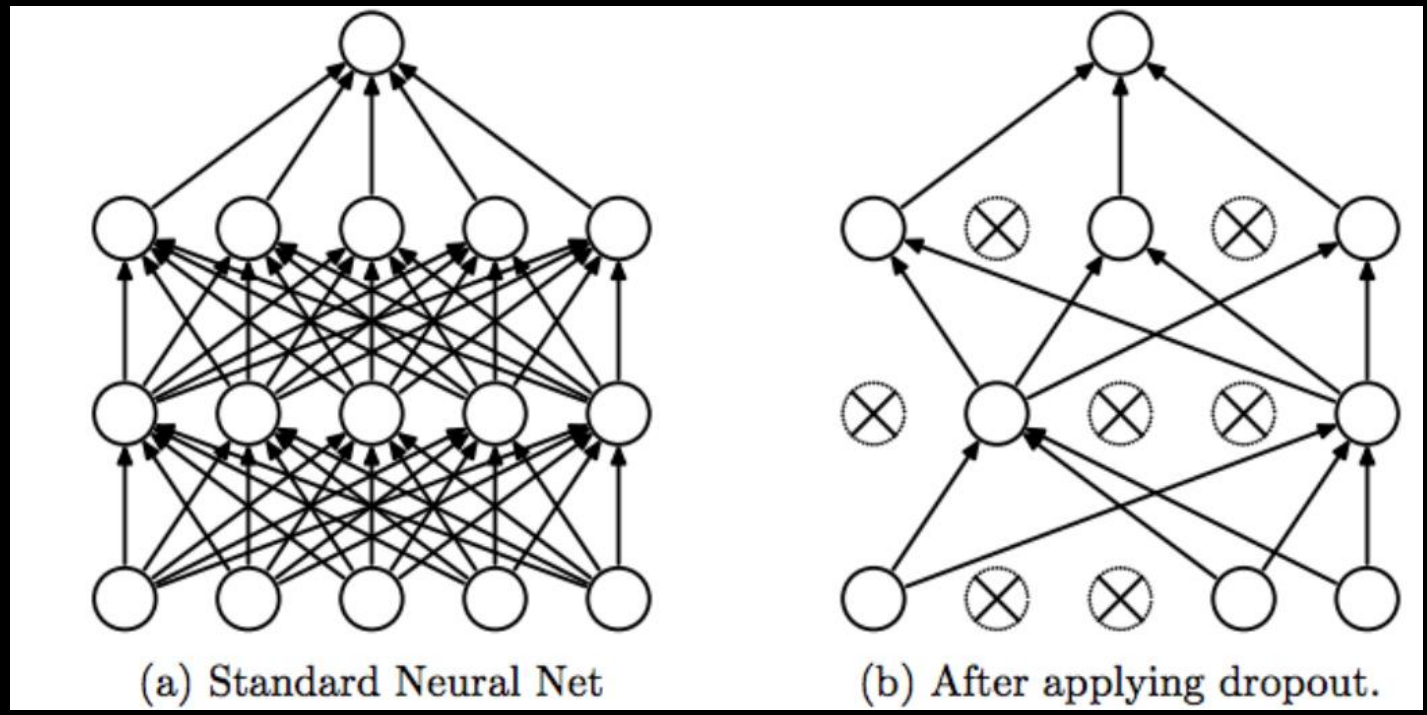
- Computation time and overfitting.
- Most deep learning models have cubic growth in complexity.
- Overfitting and exploding gradients.
- Large weight values.

Mass Street University

PER EDUCATIONEM PROGRESSUS

Dropout

- Randomly drops neurons in our network.
- Speeds up computation time.
- Keeps weights smaller by necessity.
- Reduces overfitting and improves generalization of our data.
- Creates new paths every epoch.
- A good dropout rate is between .3 and .8.



Mass Street University
PER EDUCATIONEM PROGRESSUS

Weight Constraints

- Restrict the size of weights in our network by force.
- Keeps our weights small and helps to generalize the data.
- Allows us to use larger learning rates which leads to faster convergence.
- Should only be used with other regularization methods.

Max Norm and L2 Norm

- Max norm clips values to a certain limit and scales gradients down to that limit.
- A good range for max norm is 3 to 4.
- L1 and L2 are used a lot in deep learning. They control how the optimization algorithm minimizes the loss function.
- L2 is a weight decay method.

L1 Regularization

$$\text{Cost} = \sum_{i=0}^{N} (y_i - \sum_{j=0}^{M} x_{ij} W_j)^2 + \lambda \sum_{j=0}^{M} |W_j|$$

L2 Regularization

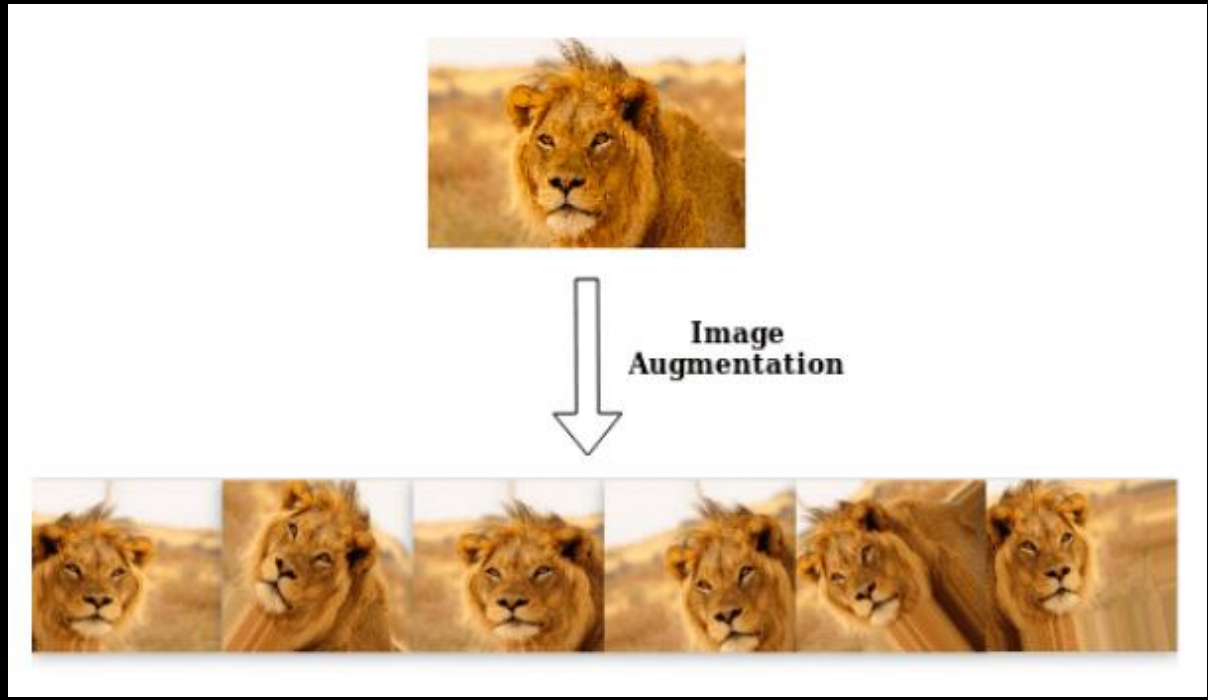
$$\text{Cost} = \underbrace{\sum_{i=0}^{N} (y_i - \sum_{j=0}^{M} x_{ij} W_j)^2}_{\text{Loss function}} + \underbrace{\lambda \sum_{j=0}^{M} W_j^2}_{\text{Regularization Term}}$$

Problems with Weight Decay

- It is a suggestion and doesn’t force weights to be small.
- Can’t use large learning rates.
- Is prone to becoming unstable if the network does too big of an update to the weights.

Data Augmentation

- Any change to data.
- Works best on images but still works with non image data.
- Applies small amounts of noise to the data to better generalize and stop overfitting.



Early Stopping

- We use validation data and monitor the error on it.
- When the error stagnates for a long period of time, we stop the model from training.
- Is usually used while automatic hyperparameter tuning.

Data Normalization

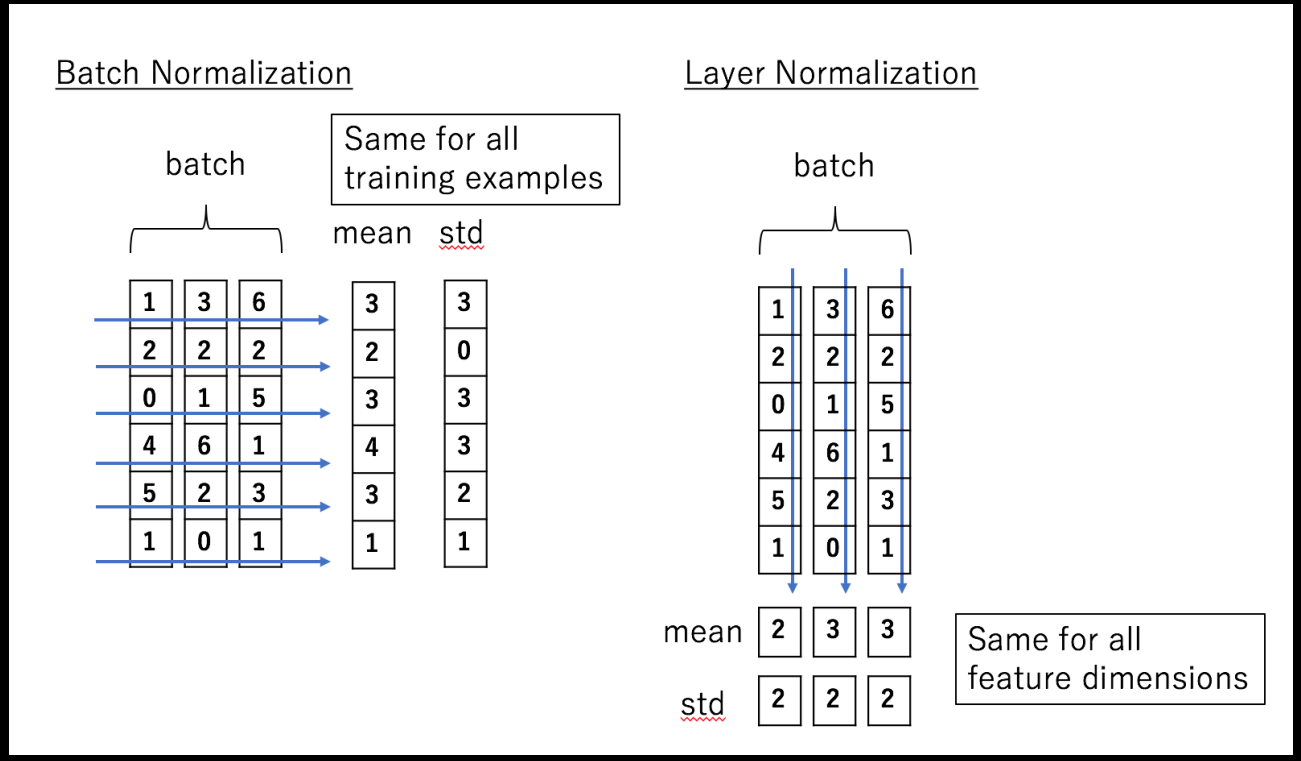
- Simply just scaling the values of numbers down to between 0 and 1.

Mass Street University

PER EDUCATIONEM PROGRESSUS

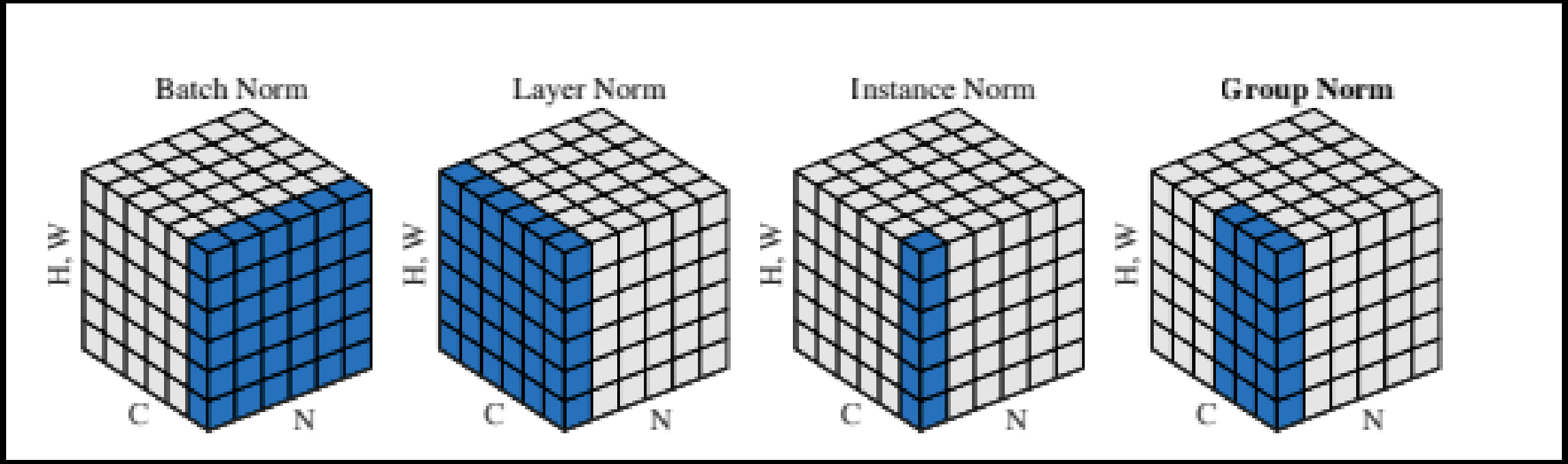
Batch Normalization

- Normalizes the activations in the network for each batch.
- Batch Norm calculates the variance and mean for each feature and subtracts the mean divides by the standard deviation.
- Use less dropout when using Batch Norm and use larger batch sizes.
- Can’t be used on RNNs.



Mass Street University
PER EDUCATIONEM PROGRESSUS

Normalization Techniques



Mass Street University
PER EDUCATIONEM PROGRESSUS

Initialization

- We set the starting weights and biases of our models to get them to train in different ways.
- If done well, it can greatly improve our model's performance.
- If you use ReLU you should use ‘he initialization’.
- If you are using TanH, use ‘Xavier initialization’.

Mass Street University

PER EDUCATIONEM PROGRESSUS

ReCap

- Dropout is the easiest, and often best, method to improve our networks.
- If you use Dropout, use max-norm with a value between 1 and 4.
- Regularize your network and normalize the input data.
- Use an initializer for the weights and biases.