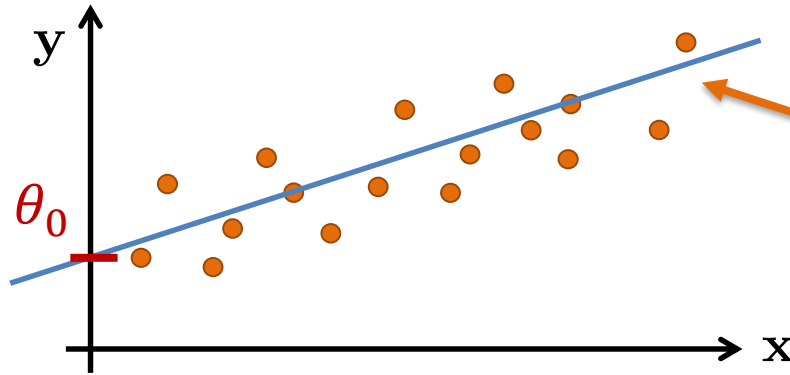# Introduction to Neural Networks

From Linear and Logistic Regression to Neural Networks

Linear Regression

= a **supervised learning** method to find a **linear model** of the form

$$\hat{y}_i = \theta_0 + \sum_{j=1}^{d} x_{ij}\theta_j = \theta_0 + x_{i1}\theta_1 + x_{i2}\theta_2 + \cdots + x_{id}\theta_d$$

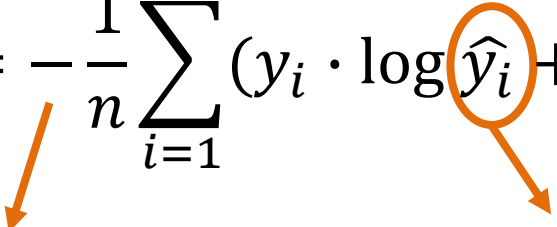Goal: find a model that explains a target y given the input x

Logistic Regression

- Loss per training sample

$$\mathcal{L}(\hat{y}_i, y_i) = -[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$$
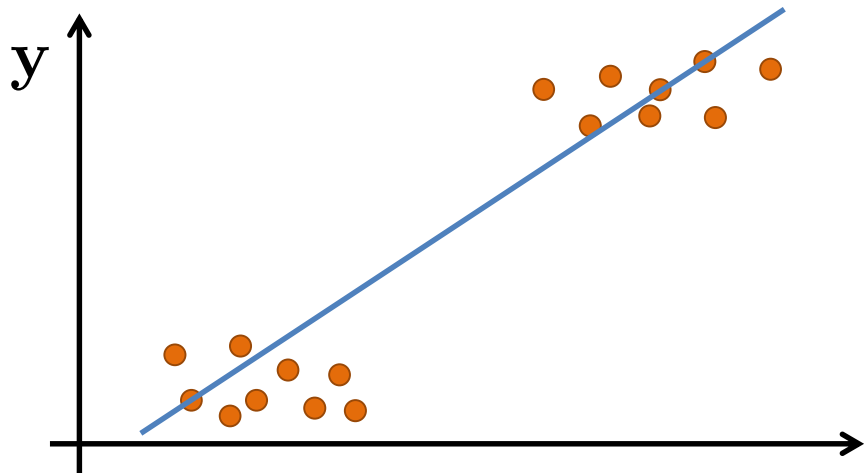
- Overall loss

$$\mathcal{C}(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} (y_i \cdot \log \widehat{y}_i + (1 - y_i) \cdot \log[1 - \widehat{y}_i])$$
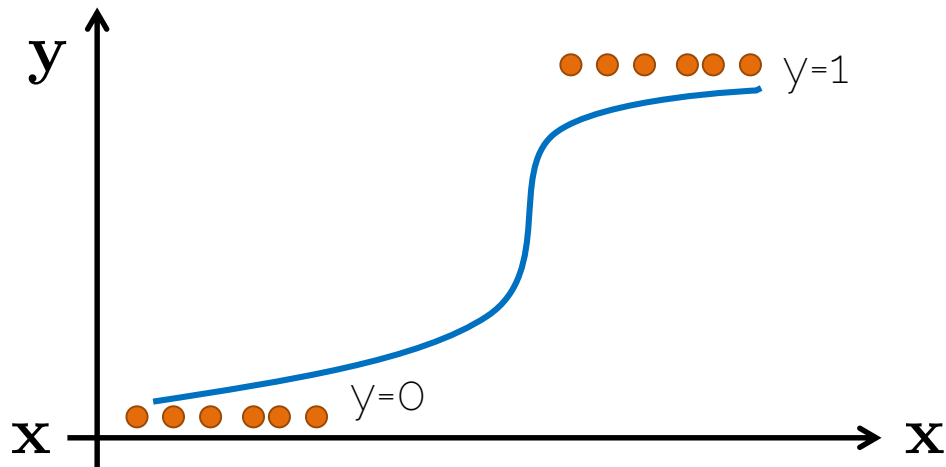
Minimization

$$\widehat{y}_i = \sigma(x_i \boldsymbol{\theta})$$
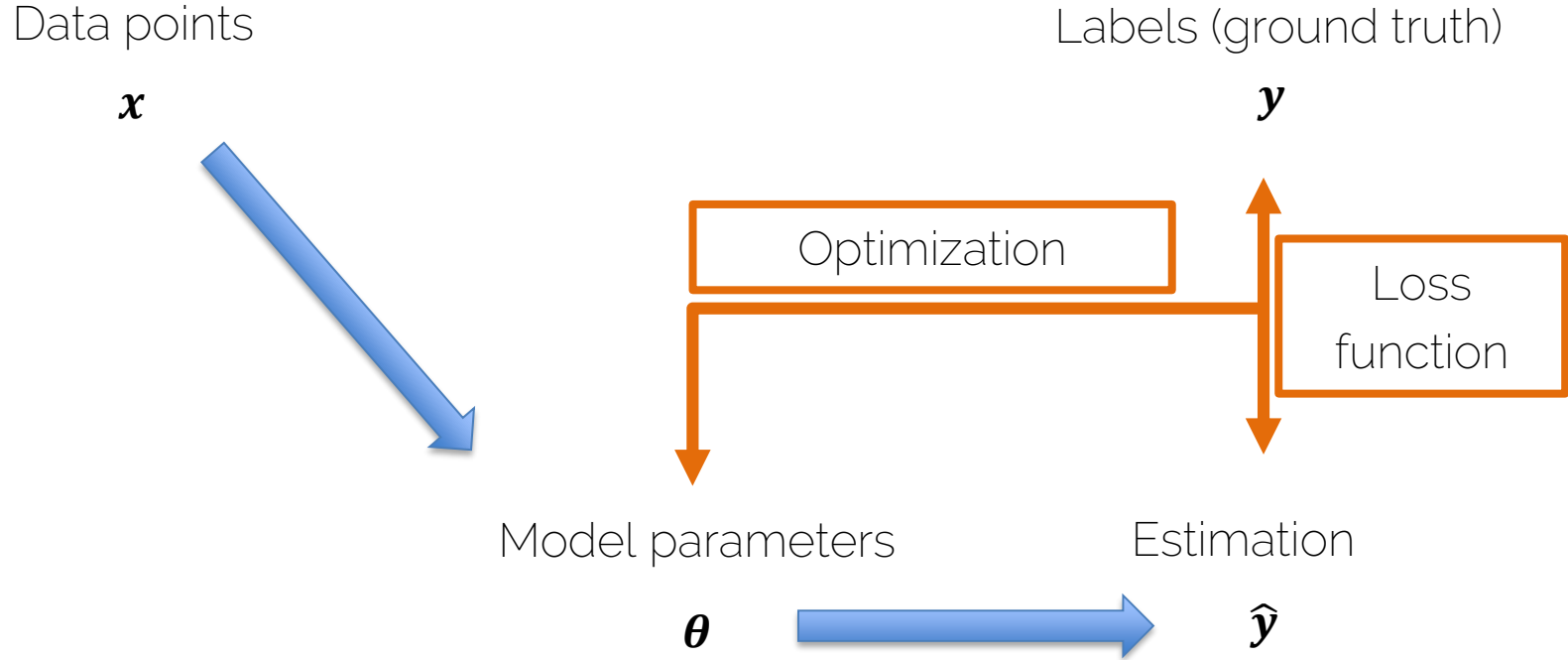
Linear vs Logistic Regression

Predictions can exceed the range of the training samples
→ in the case of classification [0;1] this becomes a real issue

Predictions are guaranteed to be within [0;1]

How to obtain the Model?

Data points

x

Labels (ground truth)

y

Optimization

Loss function

Model parameters

θ

Estimation

$\widehat{y}$

Linear Score Functions

- Linear score function as seen in linear regression

$$f_i = \sum_j w_{i,j}\, x_j$$
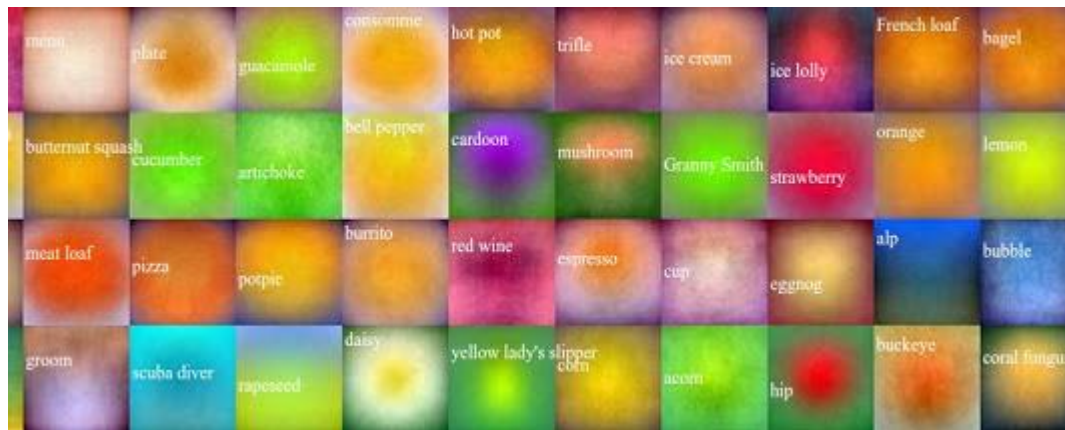
$$f = W\, x \qquad \text{(Matrix Notation)}$$

Linear Score Functions on Images

- Linear score function $f = Wx$
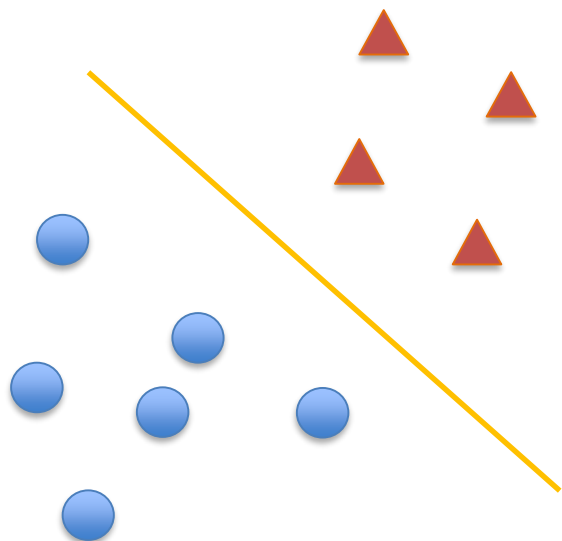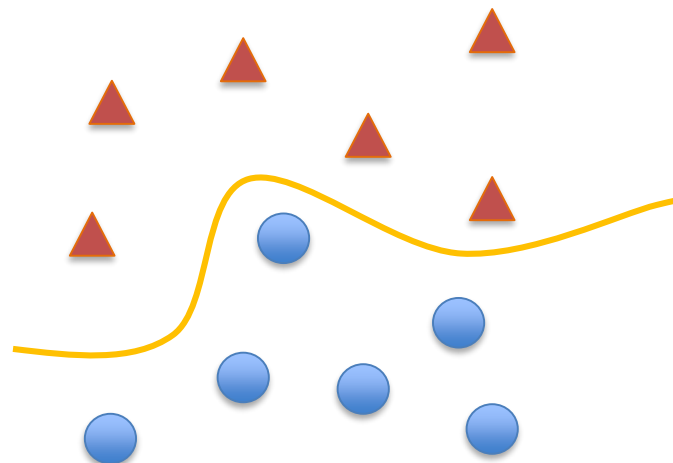

On CIFAR-10

On ImageNet

Source:: Li/Karpathy/Johnson

Linear Score Functions?

Logistic Regression

Linear Separation Impossible!

Linear Score Functions?

- Can we make linear regression better?
  - Naïve idea: Multiply with another weight matrix $\boldsymbol{W_2}$

$$\hat{\boldsymbol{f}} = \boldsymbol{W_2} \cdot \boldsymbol{W_1} \cdot \boldsymbol{x}$$

- Operation remains linear:

$$\boldsymbol{W} = \boldsymbol{W_2} \cdot \boldsymbol{W_1}$$
$$\hat{\boldsymbol{f}} = \boldsymbol{W}\boldsymbol{x}$$

- Solution $\rightarrow$ add non-linearity!!

Neural Network

- Linear score function $f = Wx$

- Neural network is a nesting of 'functions'
  - 2-layers: $f = W_2 \max(0, W_1 x)$
  - 3-layers: $f = W_3 \max(0, W_2 \max(0, W_1 x))$
  - 4-layers: $f = W_4 \tanh(W_3, \max(0, W_2 \max(0, W_1 x)))$
  - 5-layers: $f = W_5 \sigma(W_4 \tanh(W_3, \max(0, W_2 \max(0, W_1 x))))$
  - … up to hundreds of layers

Introduction to Neural Networks

Neural Network

Logistic Regression
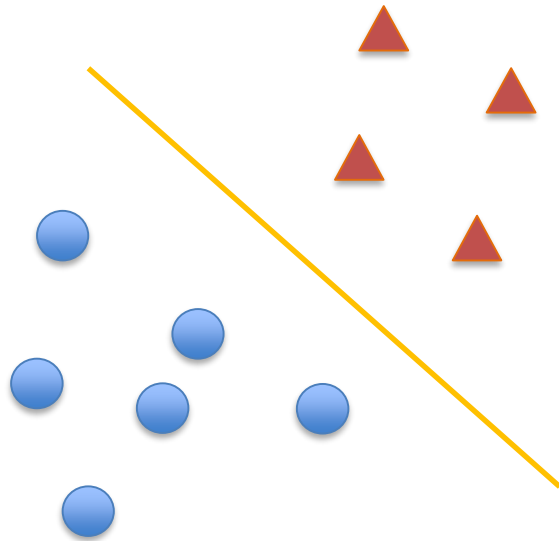
Neural Networks

Neural Network
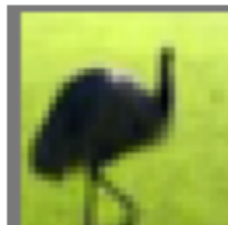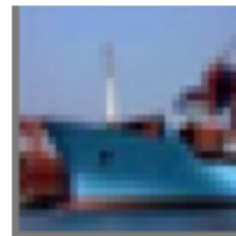
- Non-linear score function $f = \ldots (\max(\mathbf{0}, \boldsymbol{W_1 x}))$

On CIFAR-10
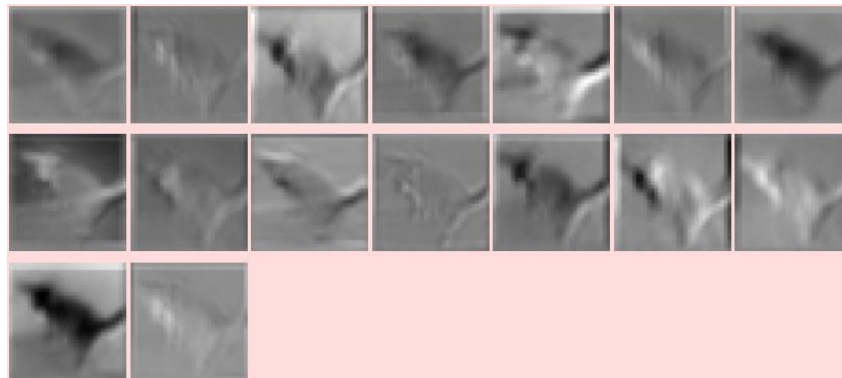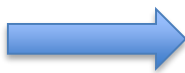
Visualizing activations of the first layer.

Source: ConvNetJS

Neural Network

1-layer network: $\boldsymbol{f} = \boldsymbol{Wx}$

2-layer network: $\boldsymbol{f} = \boldsymbol{W_2} \max(\boldsymbol{0}, \boldsymbol{W_1 x})$

$128 \times 128 = 16384$ 10

$128 \times 128 = 16384$ 1000   10
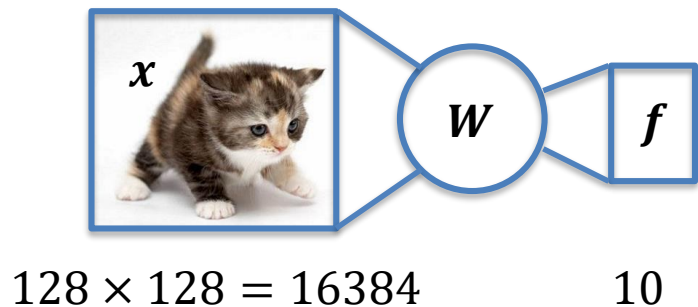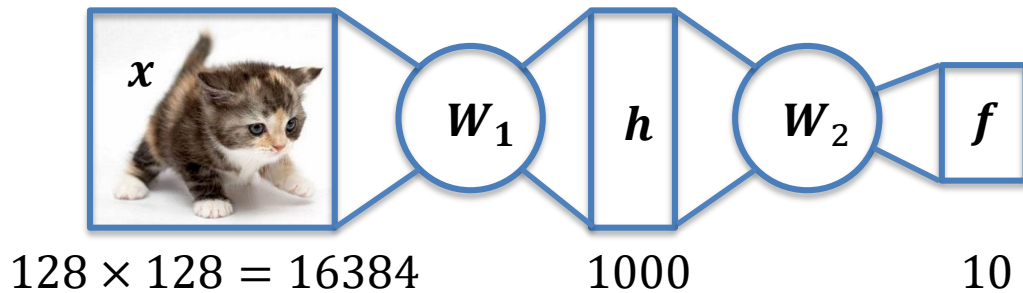
Why is this structure useful?

Neural Network

2-layer network: $f = W_2 \max(0, W_1 x)$

$128 \times 128 = 16384$ 1000 10

Input Layer   Hidden Layer   Output Layer
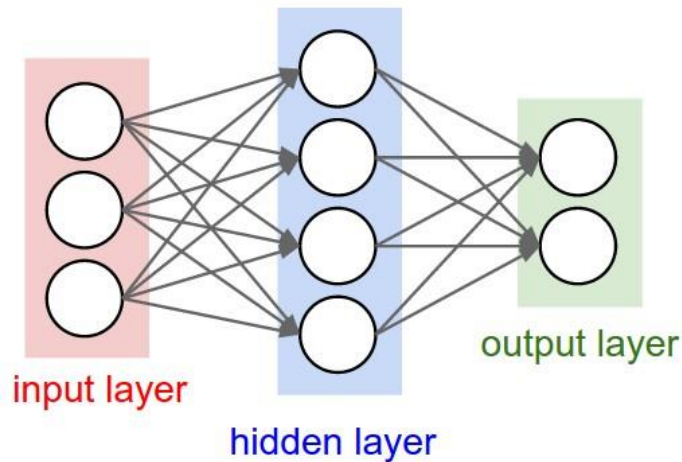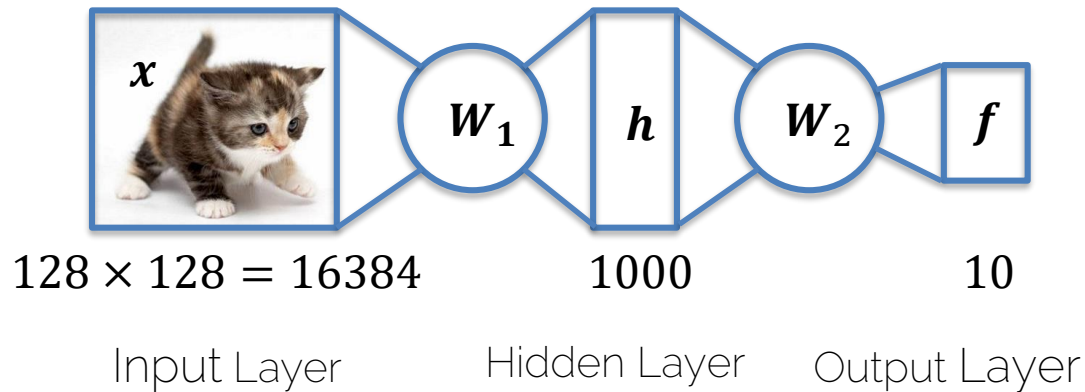
Net of Artificial Neurons

Neural Network

Source: https://towardsdatascience.com/training-deep-neural-networks-9fdb1964b964
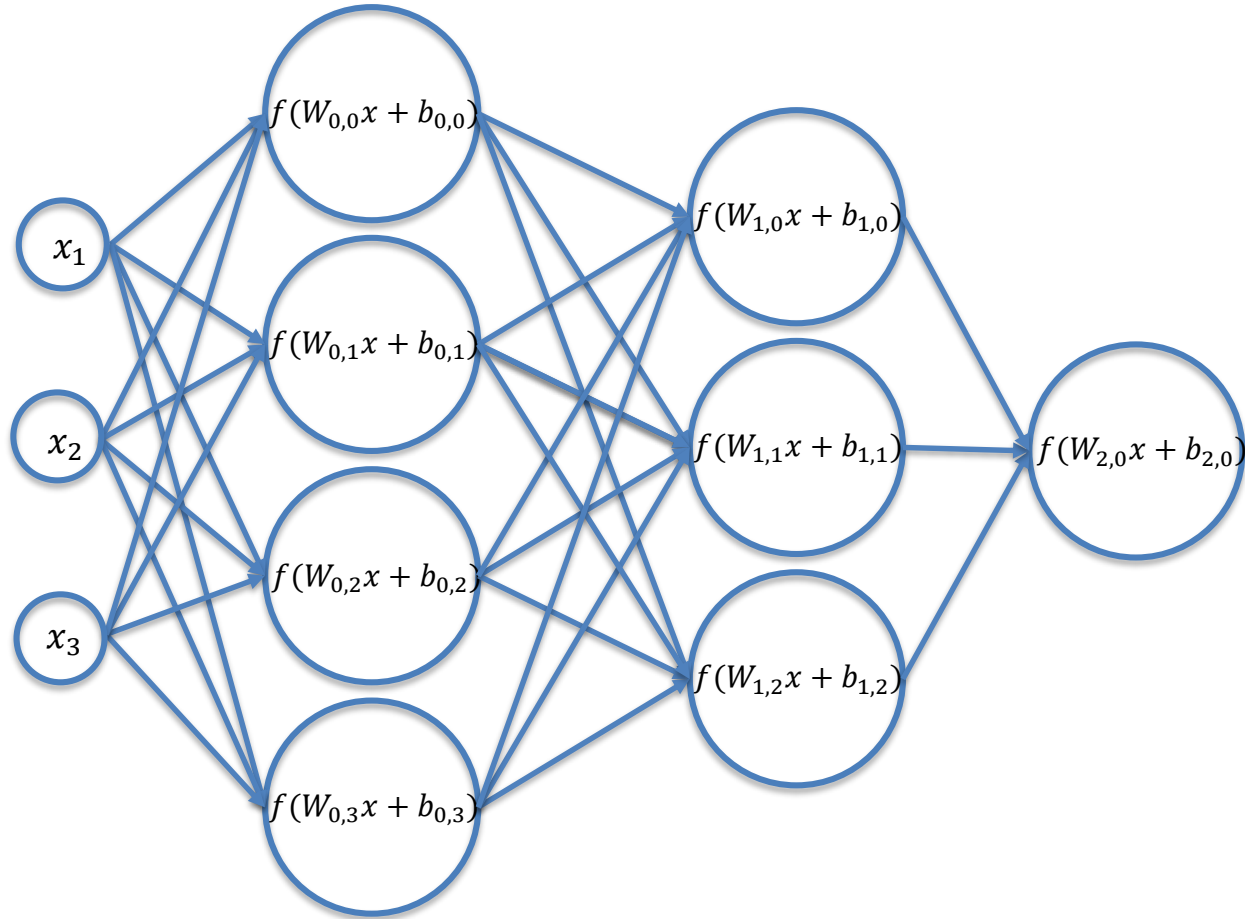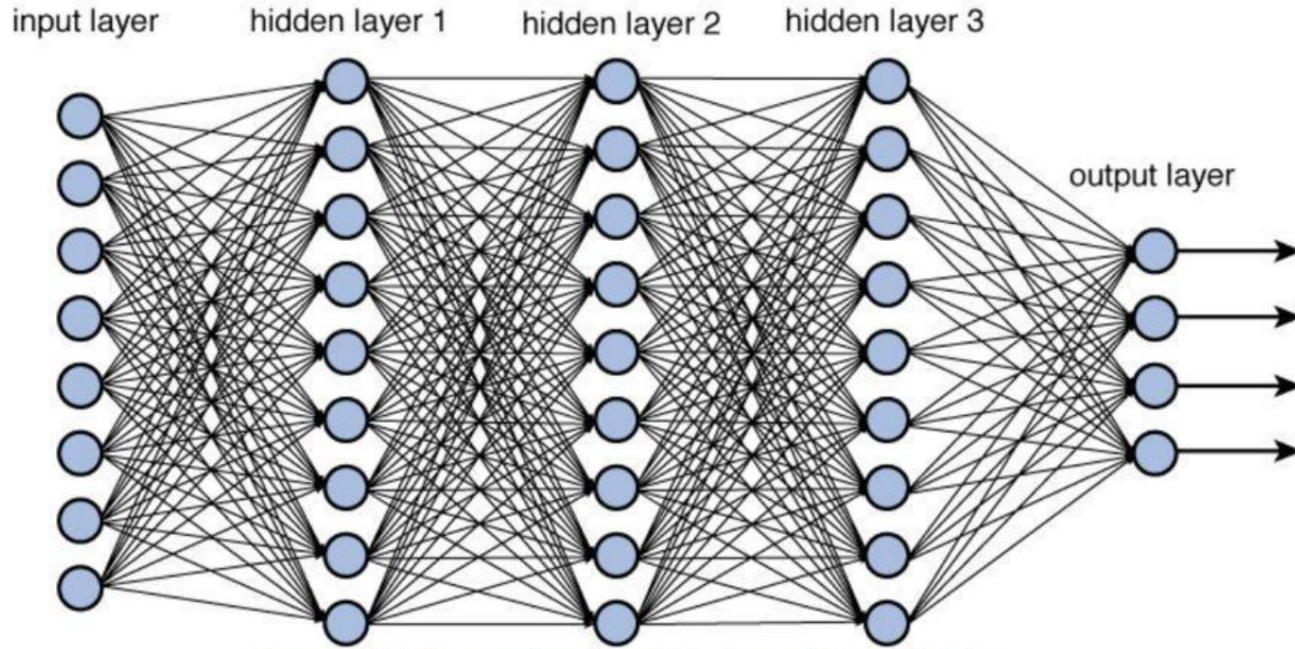
Activation Functions

Sigmoid: $\sigma(x) = \dfrac{1}{(1+e^{-x})}$
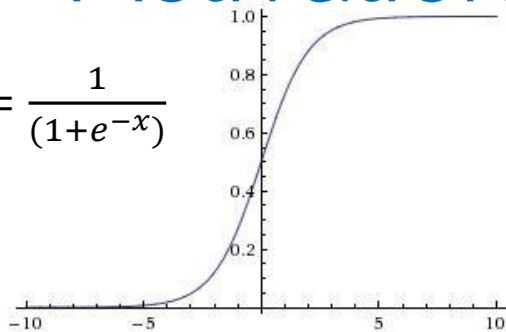
tanh: $\tanh(x)$
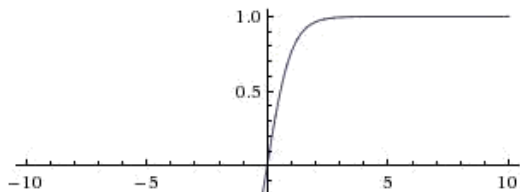
ReLU: $\max(0, x)$
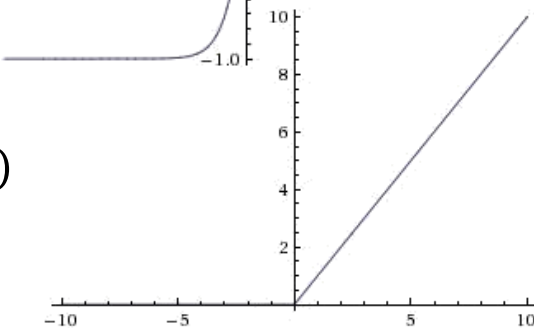
Leaky ReLU: $\max(0.1x, x)$
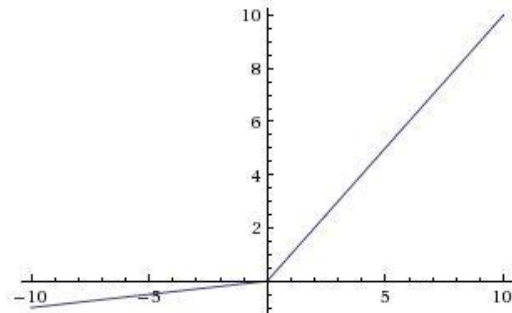
Parametric ReLU: $\max(\alpha x, x)$

Maxout $\max(w_1^T x + b_1, w_2^T x + b_2)$

ELU $f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha(e^x - 1) & \text{if } x \leq 0 \end{cases}$

Neural Network

$$f = W_3 \cdot (W_2 \cdot (W_1 \cdot x)))$$

Why activation functions?

Simply concatenating linear layers would be so much cheaper...

Neural Network

Why organize a neural network into layers?

input layer

hidden layer 1   hidden layer 2   hidden layer 3

output layer

Biological Neurons

impulses carried
toward cell body

dendrites

nucleus

cell body

branches
of axon

axon

impulses carried
away from cell body

axon
terminals
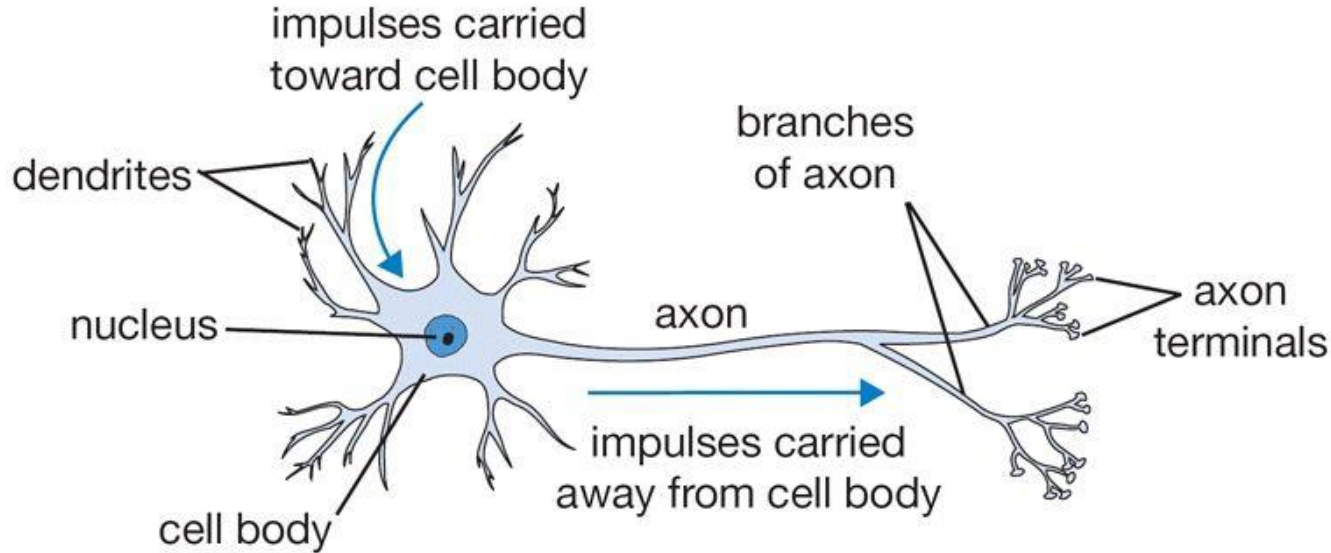
Credit: Stanford CS 231n

Biological Neurons
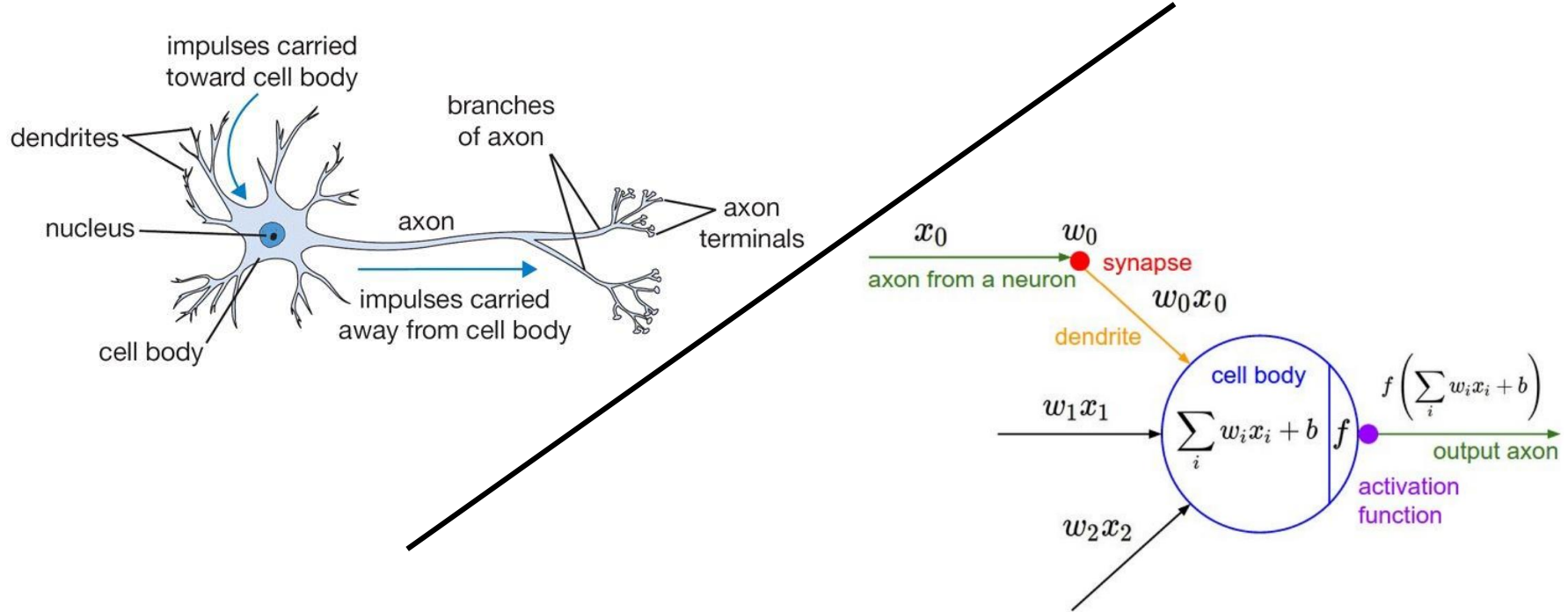
Credit: Stanford CS 231n

Artificial Neural Networks vs Brain

Artificial neural networks are **inspired** by the brain,
but not even close in terms of complexity!
The comparison is great for the media and news articles though… ☺

Artificial Neural Network

$f(W_{0,0}x + b_{0,0})$

$f(W_{0,1}x + b_{0,1})$

$f(W_{0,2}x + b_{0,2})$

$f(W_{0,3}x + b_{0,3})$

x_1

x_2

x_3

$f(W_{1,0}x + b_{1,0})$

$f(W_{1,1}x + b_{1,1})$

$f(W_{1,2}x + b_{1,2})$

$f(W_{2,0}x + b_{2,0})$
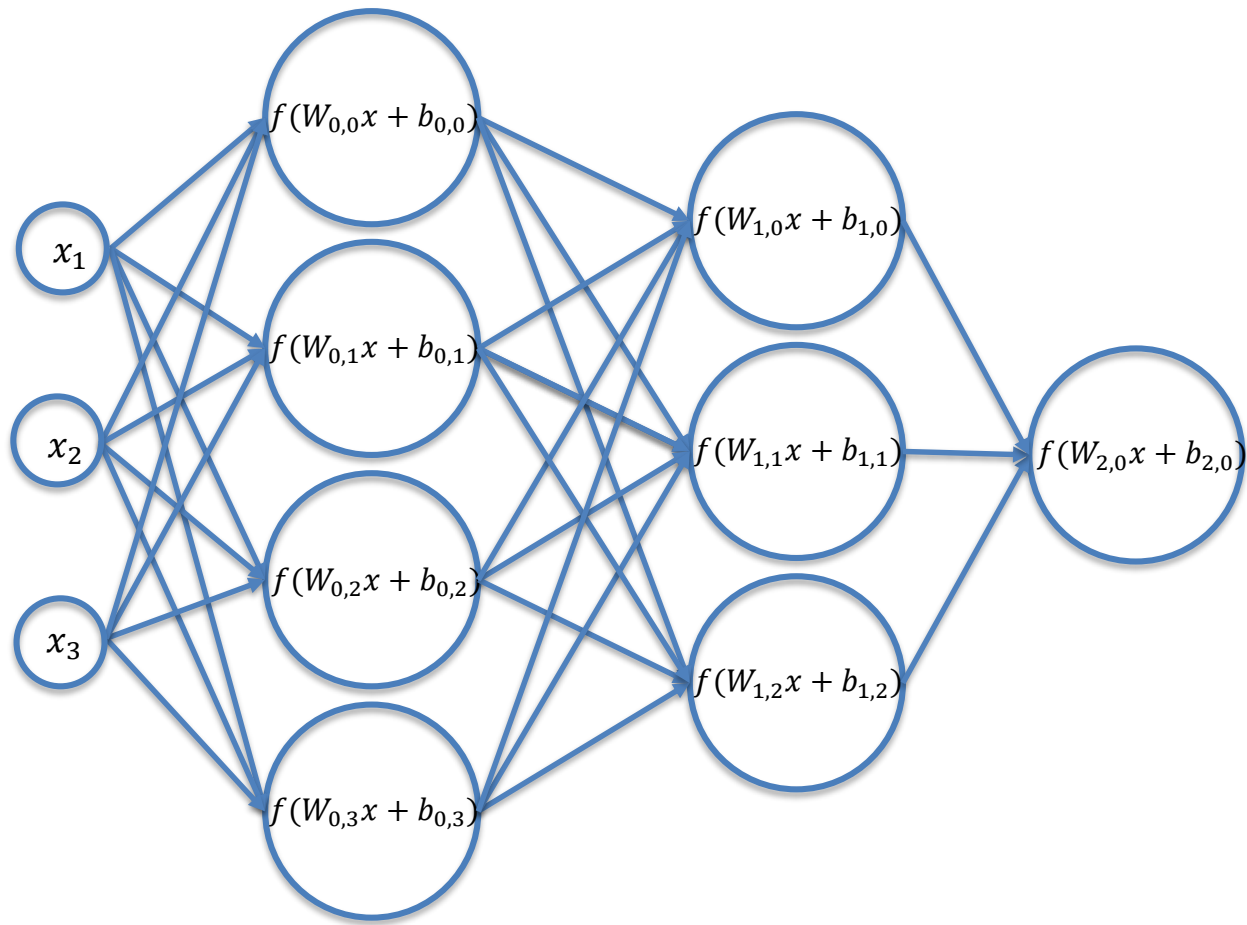
Neural Network

- Summary
  - Given a dataset with ground truth training pairs $[x_i; y_i]$,

  - Find optimal weights and biases W using stochastic gradient descent, such that the loss function is minimized
    - Compute gradients with backpropagation (use batch-mode; more later)
    - Iterate many times over training set (SGD; more later)

Computational Graphs

Computational Graphs

- Directional graph

- Matrix operations are represented as compute nodes.

- Vertex nodes are variables or operators like +, -, *, /, log(), exp() ...

- Directional edges show flow of inputs to vertices

Computational Graphs

- $f(x, y, z) = (x + y) \cdot z$
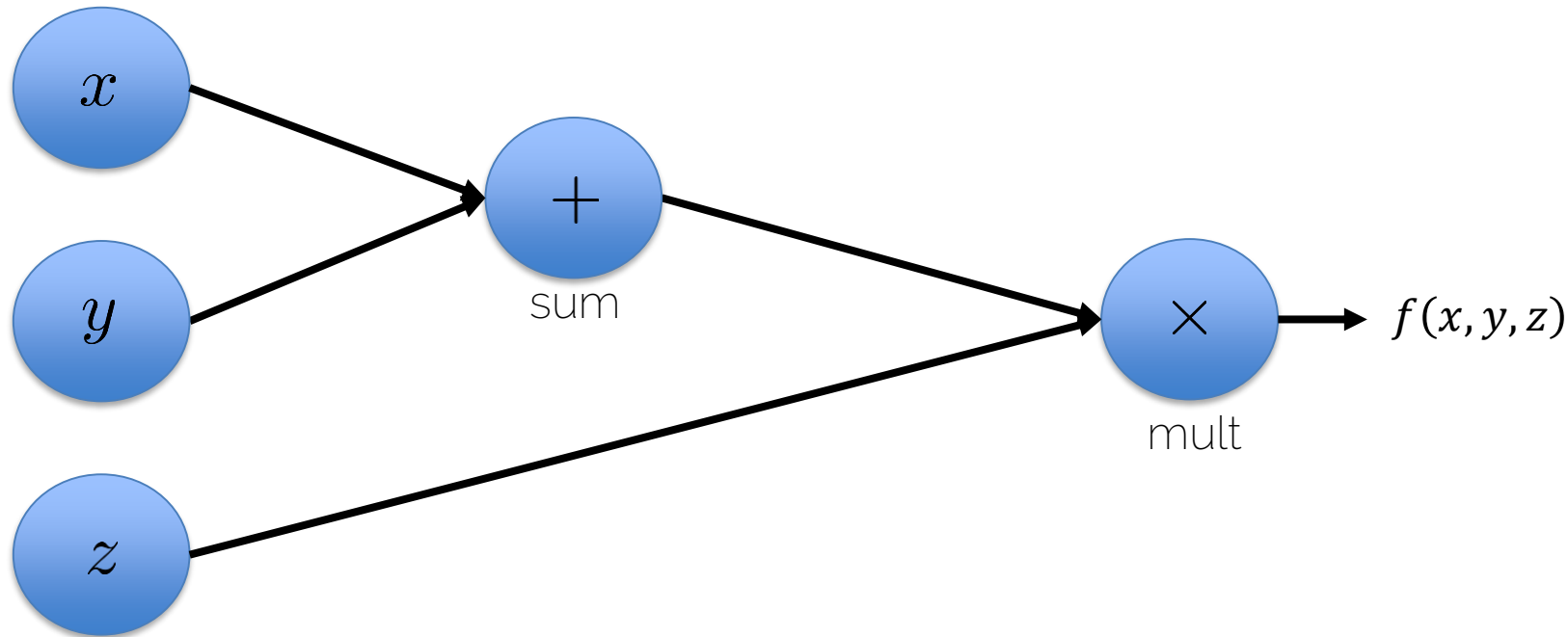
Evaluation: Forward Pass

- $f(x, y, z) = (x + y) \cdot z$

Initialization $x = 1, y = -3, z = 4$
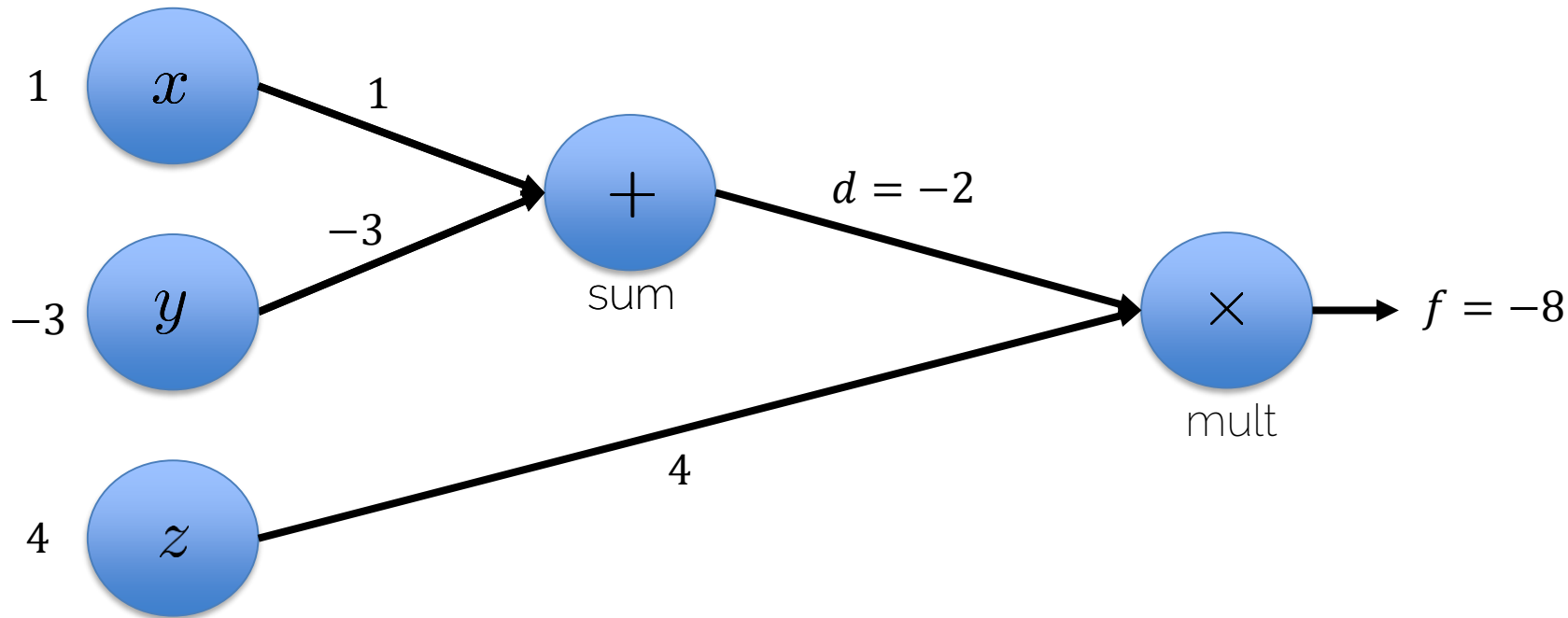
Computational Graphs

- Why discuss compute graphs?

- Neural networks have complicated architectures

$$f = W_5 \sigma(W_4 \tanh(W_3, \max(0, W_2 \max(0, W_1 x))))$$

- Lot of matrix operations!

- Represent NN as computational graphs!

Computational Graphs

A neural network can be represented as a computational graph...

- it has compute nodes (operations)

- it has edges that connect nodes (data flow)

- it is directional

- it can be organized into 'layers'

Computational Graphs
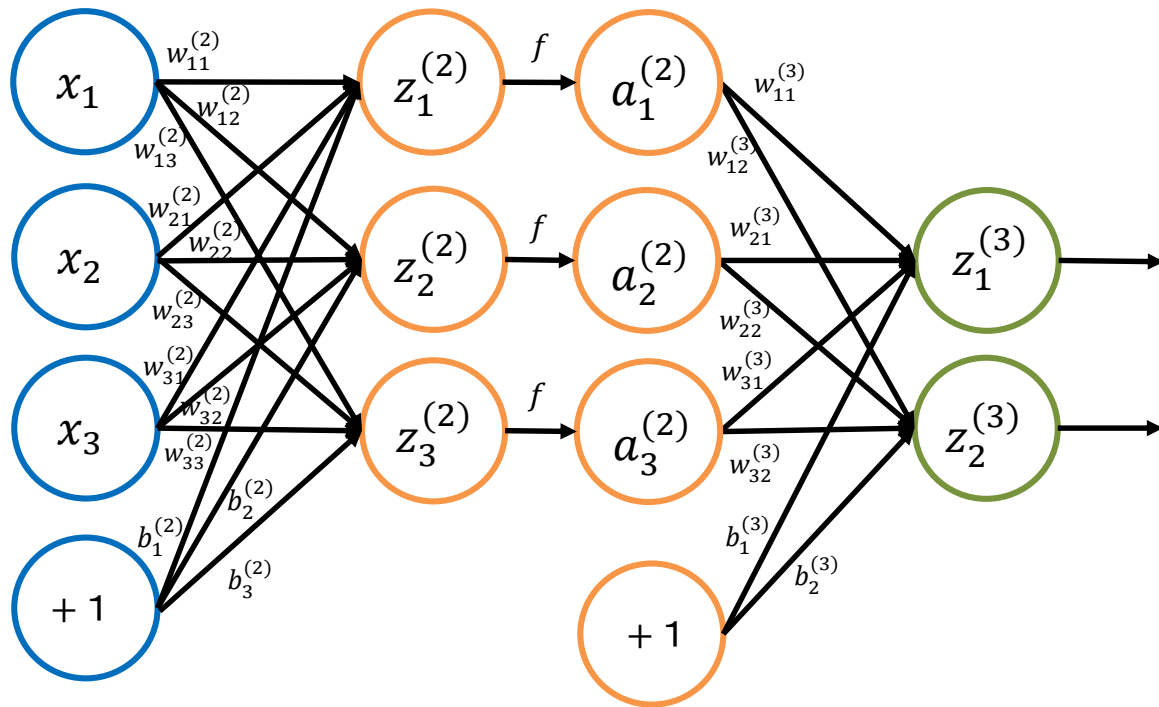
$$z_k^{(2)} = \sum_i x_i w_{ik}^{(2)} + b_k^{(2)}$$

$$a_k^{(2)} = f(z_k^{(2)})$$

$$z_k^{(3)} = \sum_i a_i^{(2)} w_{ik}^{(3)} + b_k^{(3)}$$

Computational Graphs

- From a set of neurons to a Structured Compute Pipeline

Convolution
AvgPool
MaxPool
Concat
Dropout
Fully connected
Softmax
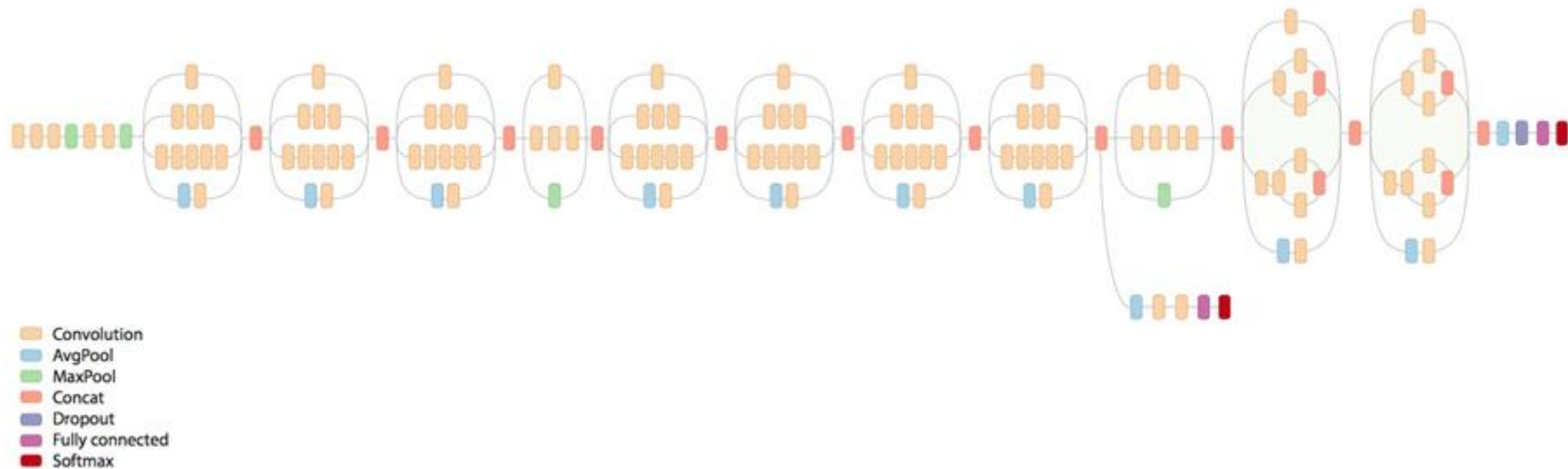
[Szegedy et al.,CVPR'15] Going Deeper with Convolutions

Computational Graphs

- The computation of Neural Network has further meanings:
  - The multiplication of W and x: encode input information
  - The activation function: select the key features
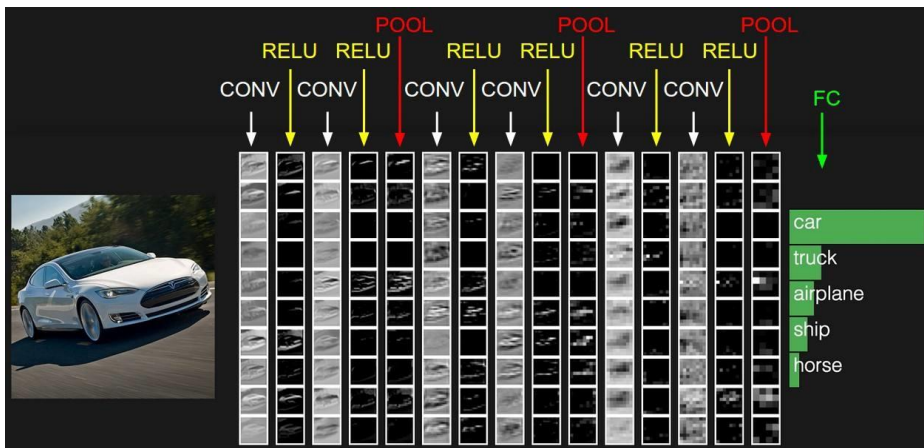
Source; https://www.zybuluo.com/liuhui0803/note/981434

Computational Graphs

- The computations of Neural Networks have further meanings:
  - The convolutional layers: extract useful features with shared weights

Kernel

Input Image

Output Image
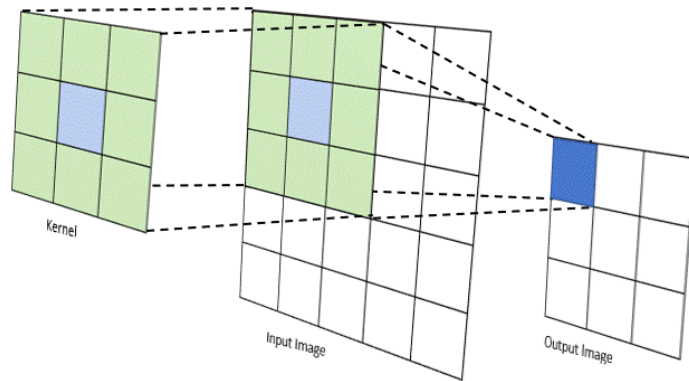
Source: https://medium.com/@timothy_terati/image-convolution-filtering-a54dce7c786b

Computational Graphs

- The computations of Neural Networks have further meanings:
  - The convolutional layers: extract useful features with shared weights

Source: https://www.zybuluo.com/liuhui0803/note/981434
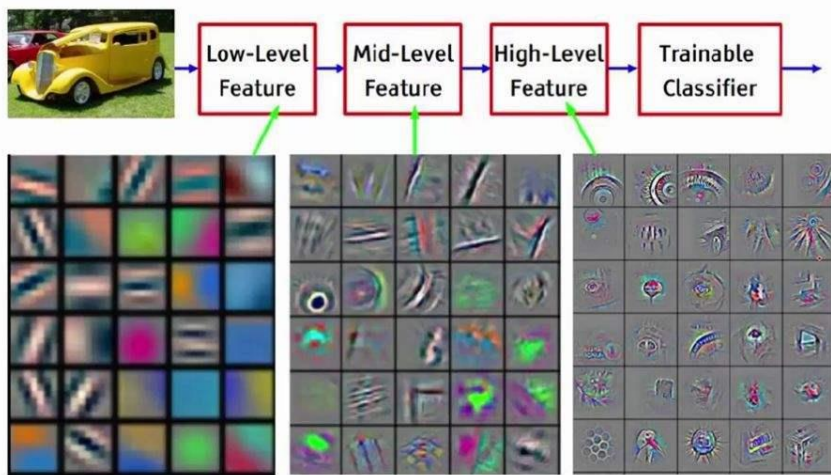
Loss Functions

What's Next?
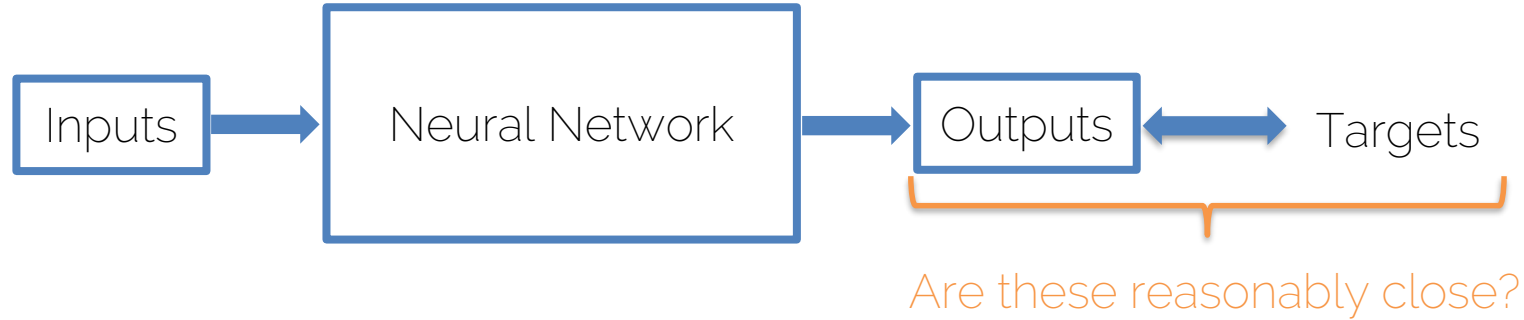
We need a way to describe how close the network's outputs (= predictions) are to the targets!

What's Next?

Idea: calculate a **'distance'** between prediction and target!
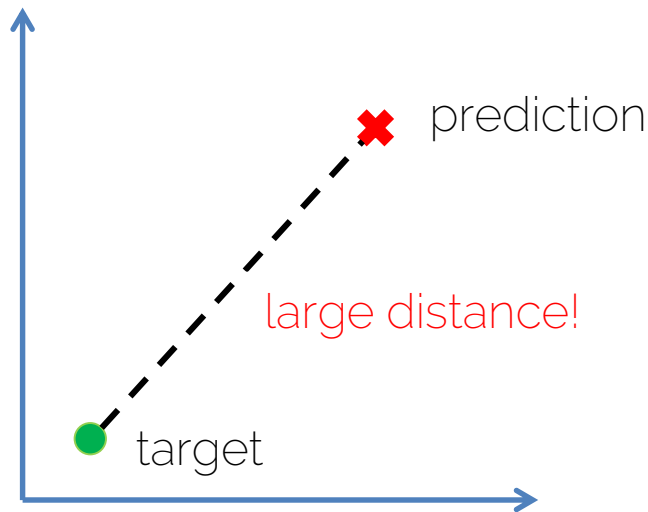
bad prediction
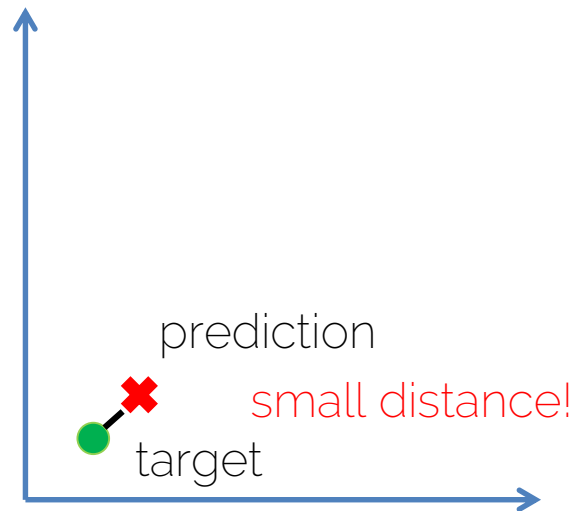
good prediction

Loss Functions

- A function to **measure the goodness of the predictions** (or equivalently, the network's performance)

Intuitively, ...

- a **large loss indicates bad predictions**/performance (→ performance needs to be improved by training the model)

- the choice of the loss function depends on the concrete problem or the distribution of the target variable

Regression Loss

- L1 Loss:

$$L(\boldsymbol{y}, \widehat{\boldsymbol{y}}; \boldsymbol{\theta}) = \frac{1}{n} \sum_{i}^{n} ||y_i - \widehat{y_i}||_1$$

- MSE Loss:

$$L(\boldsymbol{y}, \widehat{\boldsymbol{y}}; \boldsymbol{\theta}) = \frac{1}{n} \sum_{i}^{n} ||y_i - \widehat{y_i}||_2^2$$

Binary Cross Entropy

- Loss function for binary (yes/no) classification

$$L(\boldsymbol{y}, \widehat{\boldsymbol{y}}; \boldsymbol{\theta}) = -\frac{1}{n}\sum_i^n [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$$

Yes! (0.8)

No! (0.2)
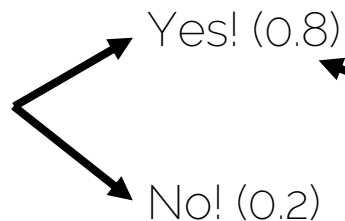
The network predicts the probability of the input belonging to the "yes" class!

Cross Entropy

Generalizes to multi-class classification:

$$y_{ik} = \begin{cases} 1 & if\ x_i \in class\ k \\ 0 & else \end{cases}$$

$$L(\boldsymbol{y}, \widehat{\boldsymbol{y}};\ \boldsymbol{\theta}) = -\sum_{i=1}^{n}\sum_{k=1}^{k}(y_{ik} \cdot \log\widehat{y}_{ik})$$

dog (0.1)

rabbit (0.2)

duck (0.7)

...
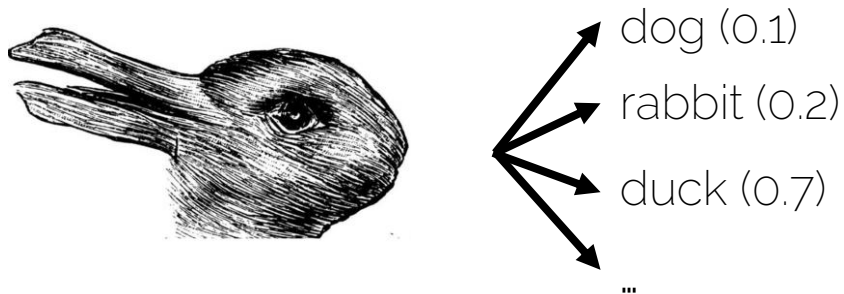
More General Case

- Ground truth: $\boldsymbol{y}$
- Prediction: $\widehat{\boldsymbol{y}}$
- Loss function: $L(\boldsymbol{y}, \widehat{\boldsymbol{y}})$
- Motivation:
  - minimize the loss <=> find better predictions
  - predictions are generated by the NN
  - find better predictions <=> find better NN

Initially

Prediction

Targets

Bad prediction!

Loss

Training time

t_1

During Training...

Prediction

Targets

Bad prediction!

Loss

t_2

Training time

During Training...

Prediction

Targets

Bad prediction!

Loss

t_3
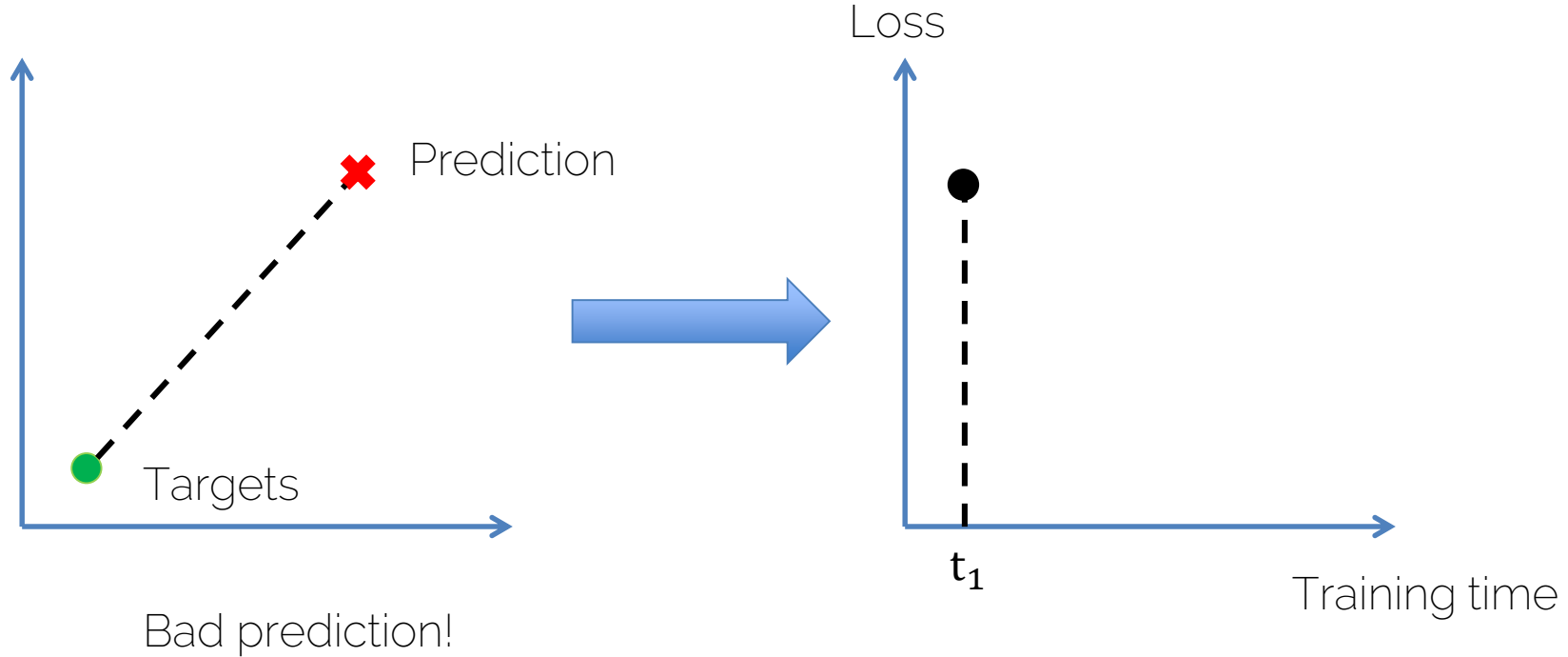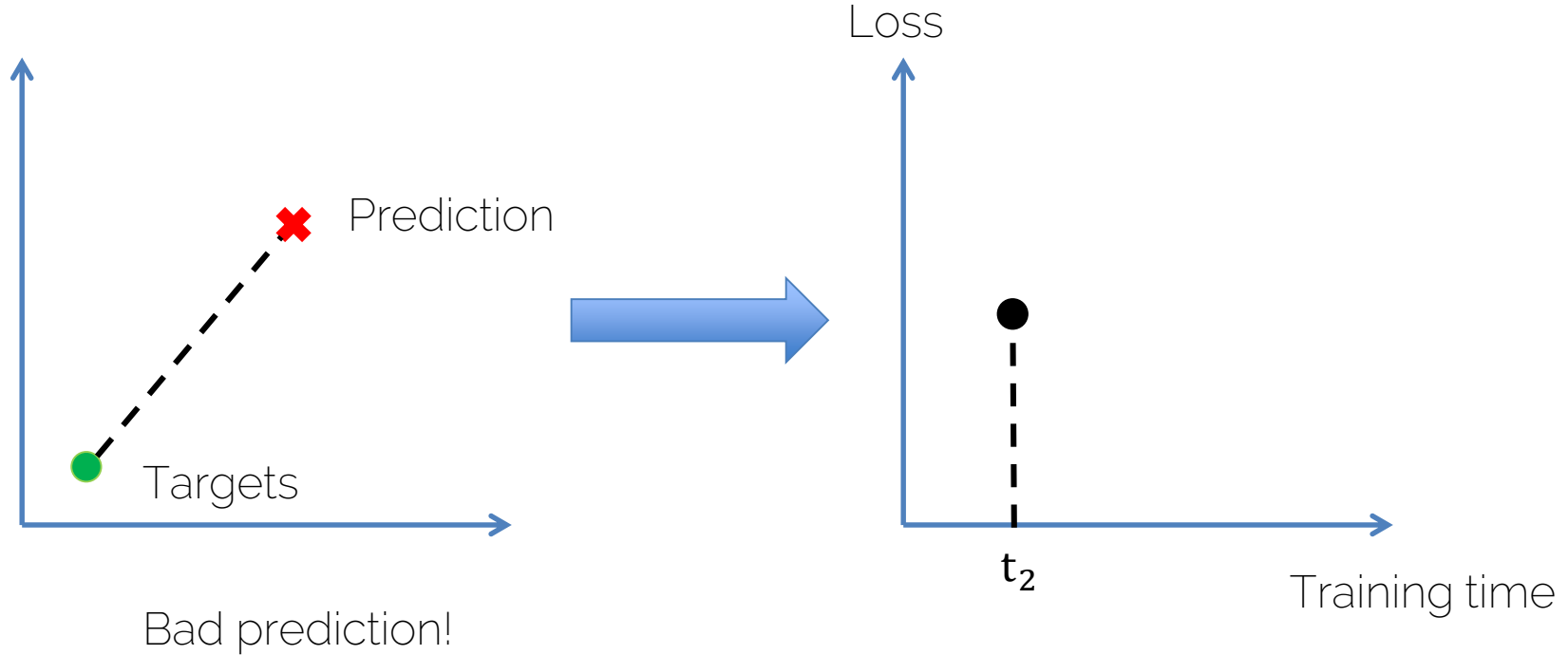
Training time
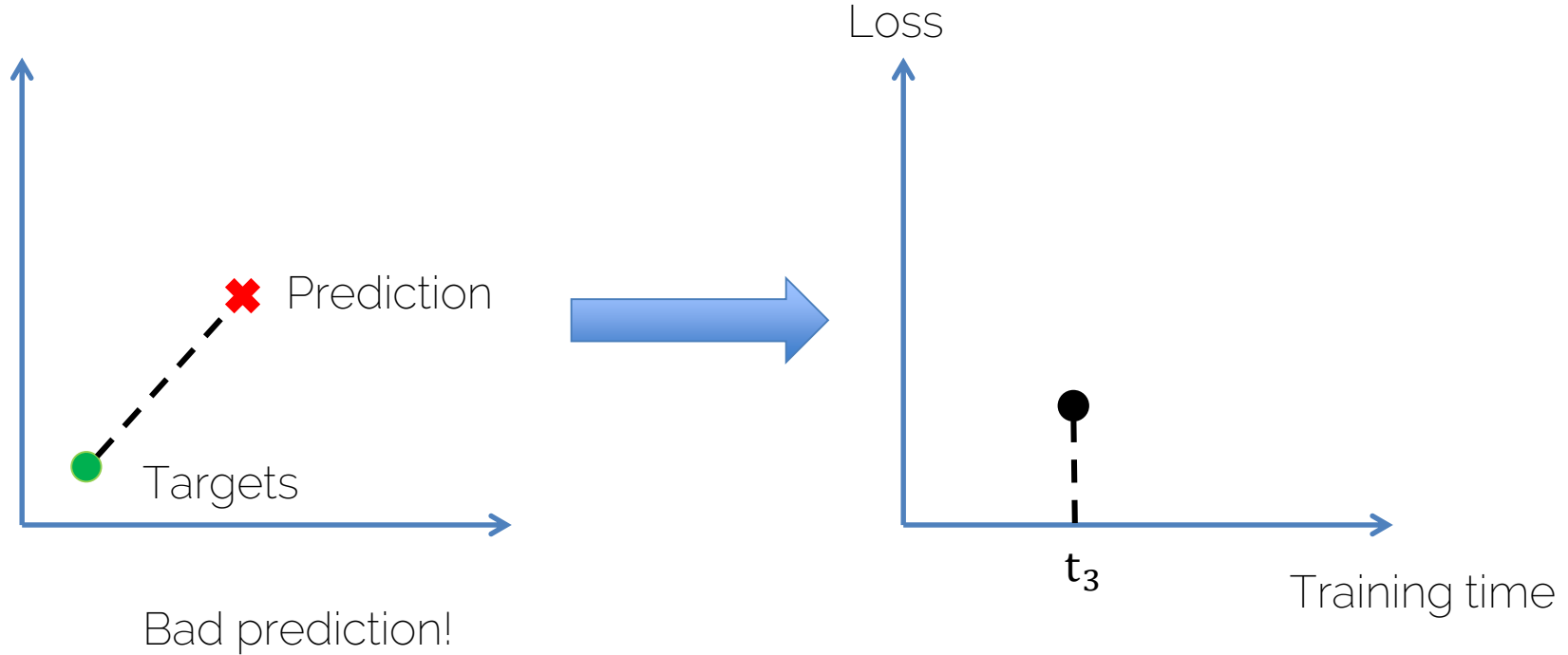
Training Curve

Loss

Training time
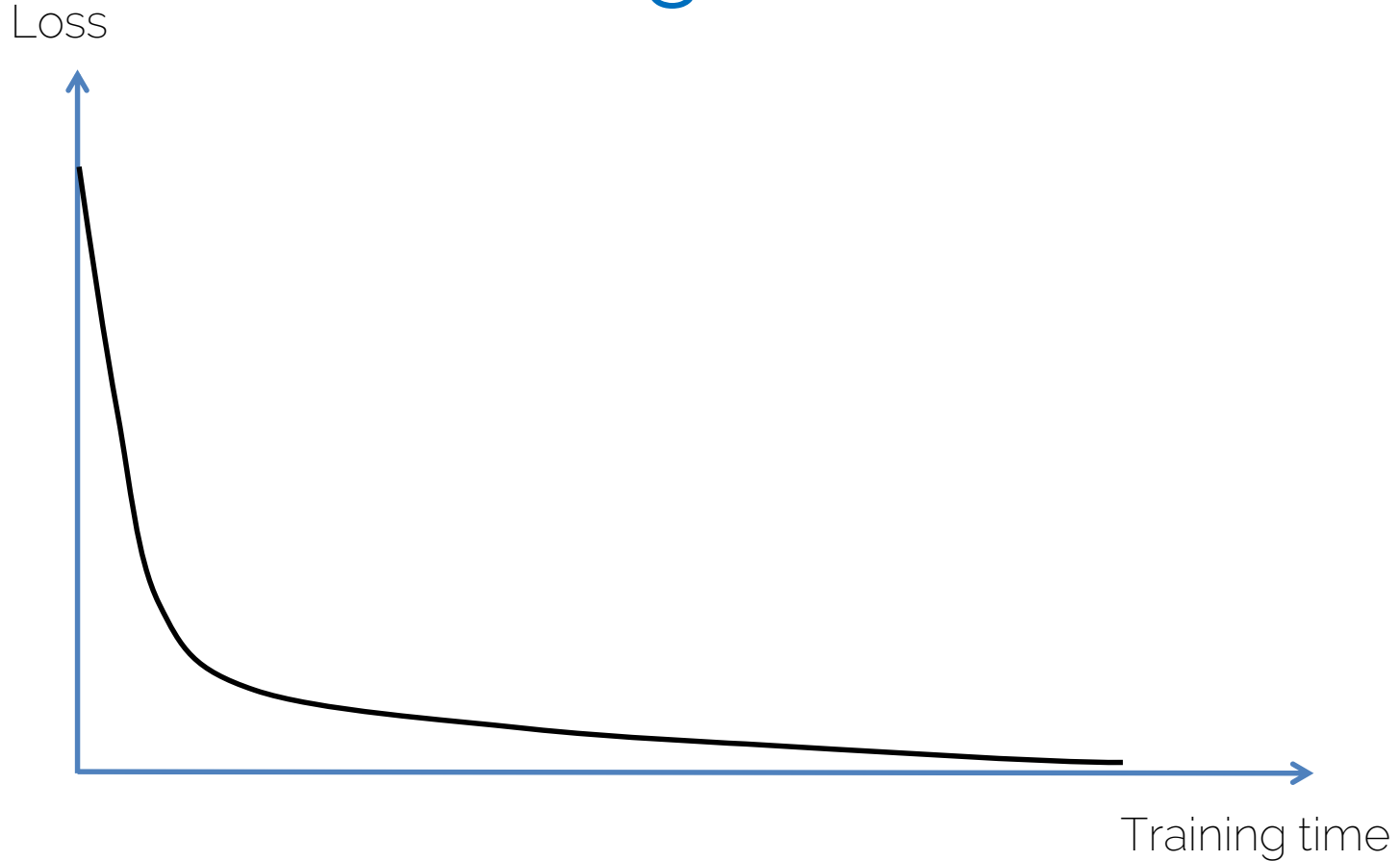
How to Find a Better NN?

Plotting loss curves against model parameters

Loss
$L(\boldsymbol{y}, f_{\boldsymbol{\theta}}(\boldsymbol{x}))$

$\boldsymbol{\theta}$

Parameters $\boldsymbol{\theta}$
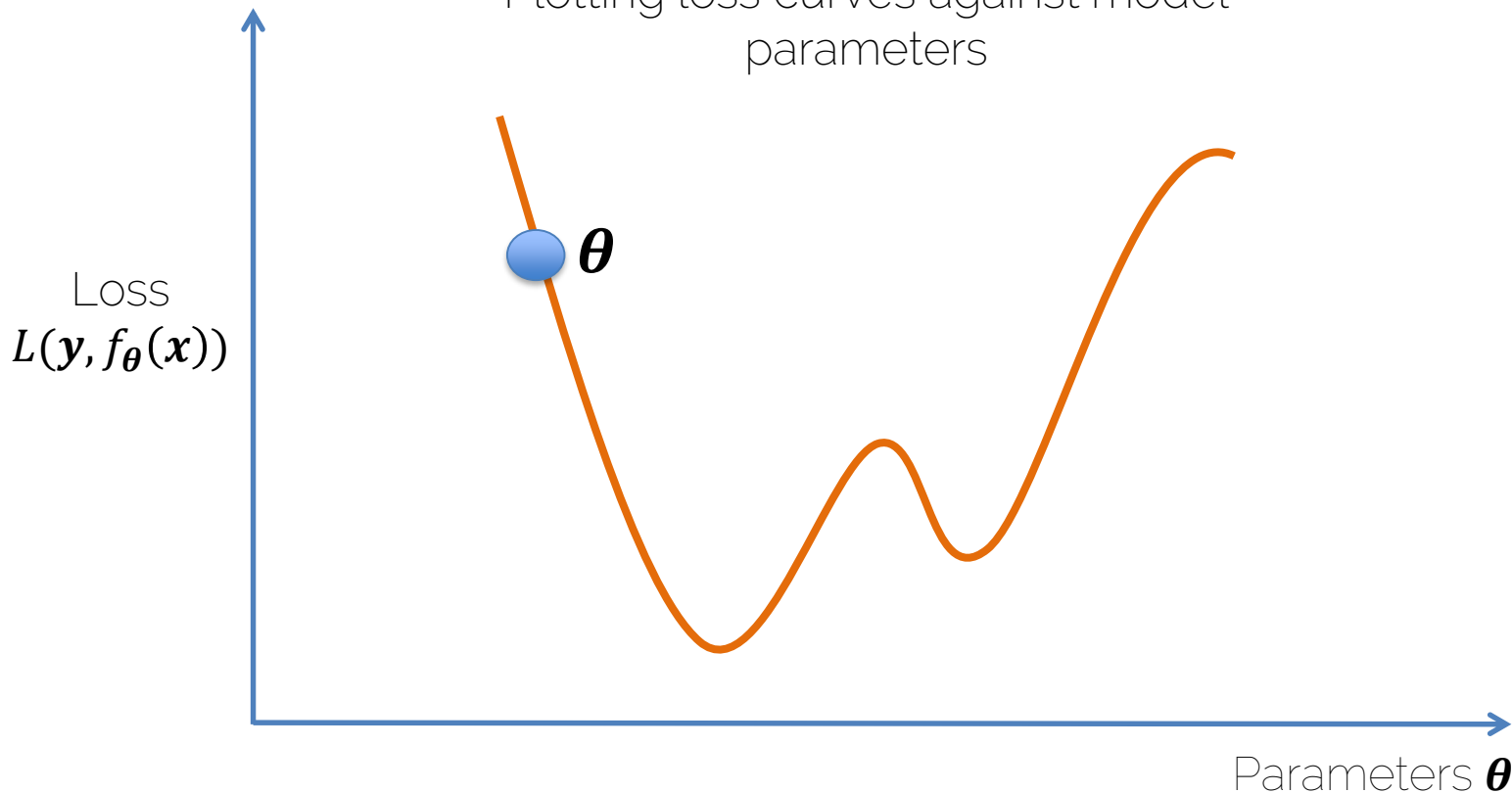
How to Find a Better NN?

- Loss function: $L(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = L(\boldsymbol{y}, f_{\boldsymbol{\theta}}(\boldsymbol{x}))$
- Neural Network: $f_{\boldsymbol{\theta}}(\boldsymbol{x})$
- Goal:
  - minimize the loss w. r. t. $\boldsymbol{\theta}$
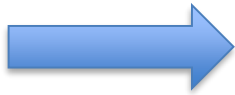
Optimization! We train compute graphs with some optimization techniques!

Gradient Descent

Minimize loss $L\big(\boldsymbol{y}, f_{\boldsymbol{\theta}}(\boldsymbol{x})\big)$ w.r.t. $\boldsymbol{\theta}$:
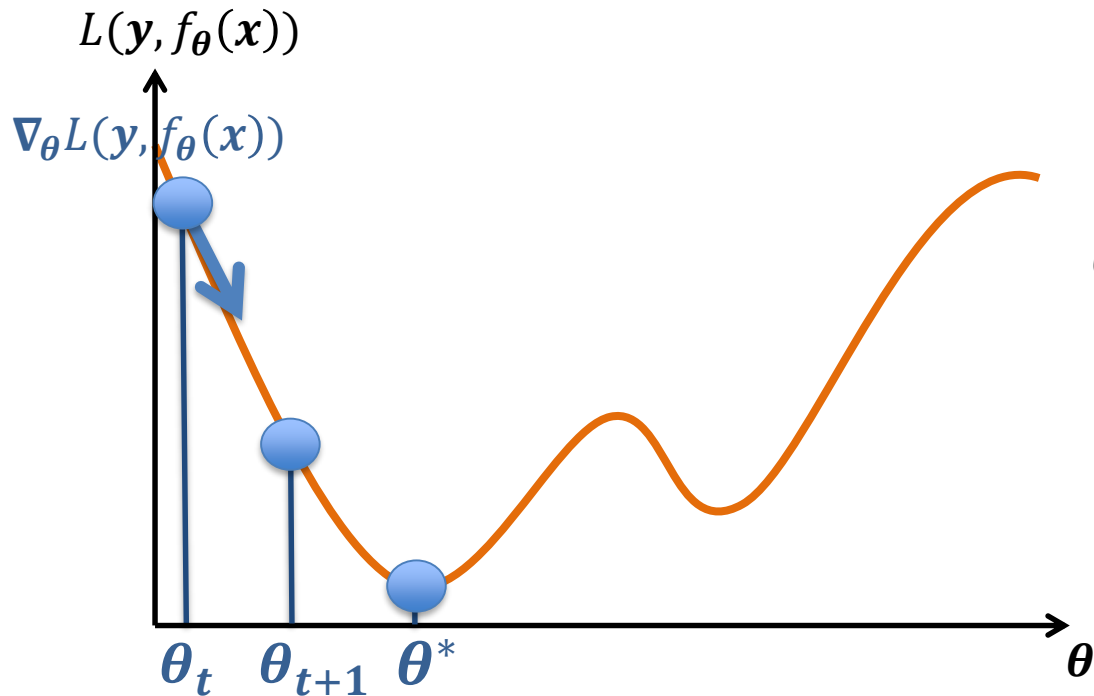
step size / learning rate

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha \, \nabla_{\boldsymbol{\theta}} L\big(\boldsymbol{y}, f_{\boldsymbol{\theta}}(\boldsymbol{x})\big)\Big|_{\boldsymbol{\theta}_t}$$

$t \to \infty$

$$\boldsymbol{\theta}^* = \arg\min L\big(\boldsymbol{y}, f_{\boldsymbol{\theta}}(\boldsymbol{x})\big)$$

How to Find a Better NN?

- Given inputs $\boldsymbol{x}$ and targets $\boldsymbol{y}$
- Given one layer NN with no activation function

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x}, \qquad \boldsymbol{\theta} = \boldsymbol{W}$$

Later $\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{b}\}$

- Given MSE Loss: $L(\boldsymbol{y}, \widehat{\boldsymbol{y}}; \boldsymbol{\theta}) = \frac{1}{n}\sum_{i}^{n} ||y_i - \widehat{y}_i||_2^2$

How to Find a Better NN?

- Given inputs $\boldsymbol{x}$ and targets $\boldsymbol{y}$

- Given one layer NN with no activation function

- Given MSE Loss: $L(\boldsymbol{y}, \widehat{\boldsymbol{y}}; \boldsymbol{\theta}) = \frac{1}{n} \sum_{i}^{n} ||y_i - \boldsymbol{W} \cdot x_i||_2^2$
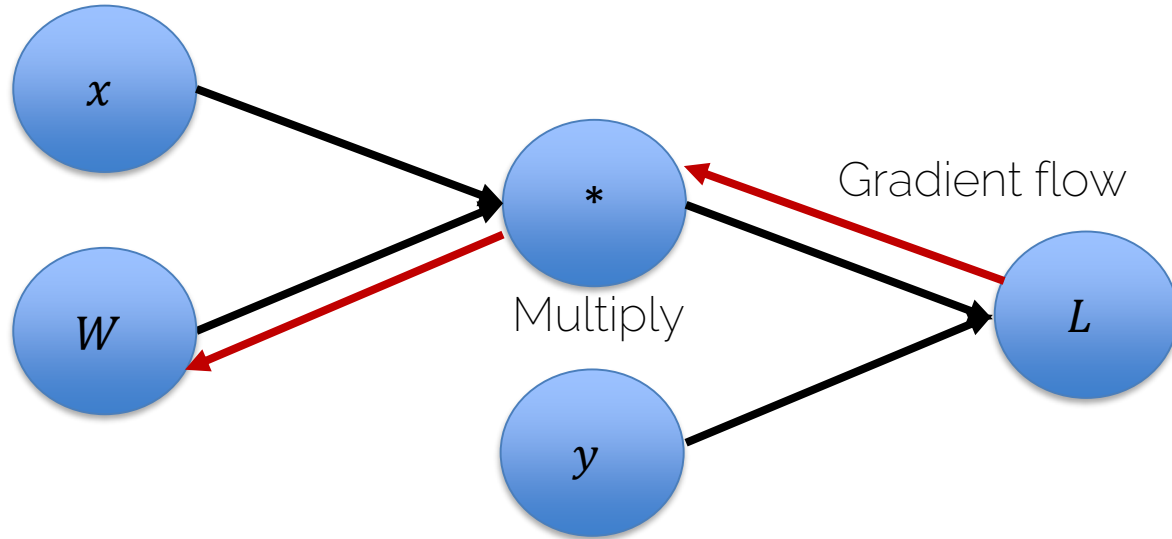
How to Find a Better NN?

- Given inputs $\boldsymbol{x}$ and targets $\boldsymbol{y}$

- Given one layer NN with no activation function

$$f_\theta(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x}, \qquad \boldsymbol{\theta} = \boldsymbol{W}$$

- Given MSE Loss: $L(\boldsymbol{y}, \widehat{\boldsymbol{y}}; \boldsymbol{\theta}) = \frac{1}{n}\sum_i^n ||\boldsymbol{W} \cdot x_i - y_i||_2^2$

- $\nabla_\theta L(\boldsymbol{y}, f_\theta(\boldsymbol{x})) = \frac{2}{n}\sum_i^n (\boldsymbol{W} \cdot x_i - y_i) \cdot x_i^T$

How to Find a Better NN?

- Given inputs $\boldsymbol{x}$ and targets $\boldsymbol{y}$

- Given a multi-layer NN with many activations

$$f = W_5 \sigma(W_4 \tanh(W_3, \max(0, W_2 \max(0, W_1 x))))$$

- Gradient descent for $L\big(\boldsymbol{y}, f_{\boldsymbol{\theta}}(\boldsymbol{x})\big)$ w. r. t. $\boldsymbol{\theta}$

  - Need to propagate gradients from end to first layer (W_1).

How to Find a Better NN?

- Given inputs x and targets y
- Given multi-layer NN with many activations
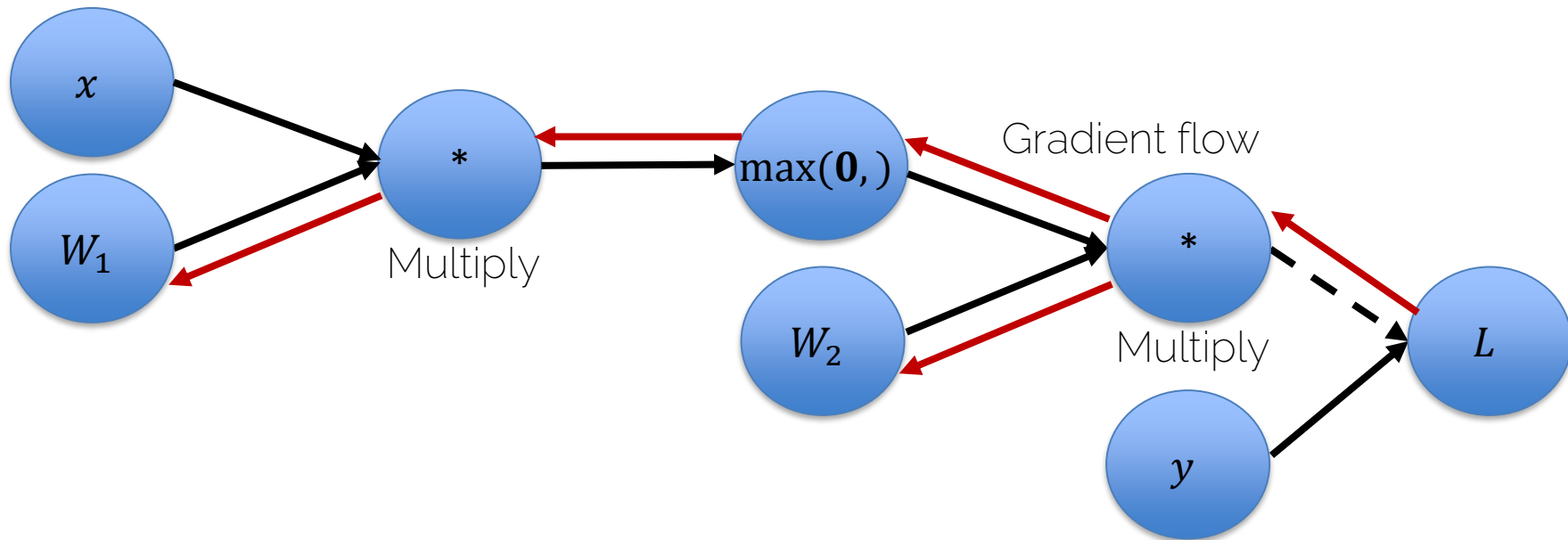
How to Find a Better NN?

- Given inputs $\boldsymbol{x}$ and targets $\boldsymbol{y}$

- Given multilayer layer NN with many activations

$$f = W_5 \sigma(W_4 \tanh(W_3, \max(0, W_2 \max(0, W_1 x))))$$

- Gradient descent solution for $L\big(\boldsymbol{y}, f_{\boldsymbol{\theta}}(\boldsymbol{x})\big)$ w. r. t. $\boldsymbol{\theta}$

  – Need to propagate gradients from end to first layer (W_1)

- Backpropagation: Use chain rule to compute gradients

  – Compute graphs come in handy!

How to Find a Better NN?

- Why gradient descent?
  - Easy to compute using compute graphs
- Other methods include
  - Newtons method
  - L-BFGS
  - Adaptive moments
  - Conjugate gradient

Summary

- Neural Networks are computational graphs
- Goal: for a given train set, find optimal weights

- Optimization is done using gradient-based solvers
  – Many options (more in the next lectures)

- Gradients are computed via backpropagation
  – Nice because can easily modularize complex functions

Next Lectures

- **Next Lecture:**
  - Backpropagation and optimization of Neural Networks

- **Check for updates** on website/piazza regarding exercises

See you next week ☺