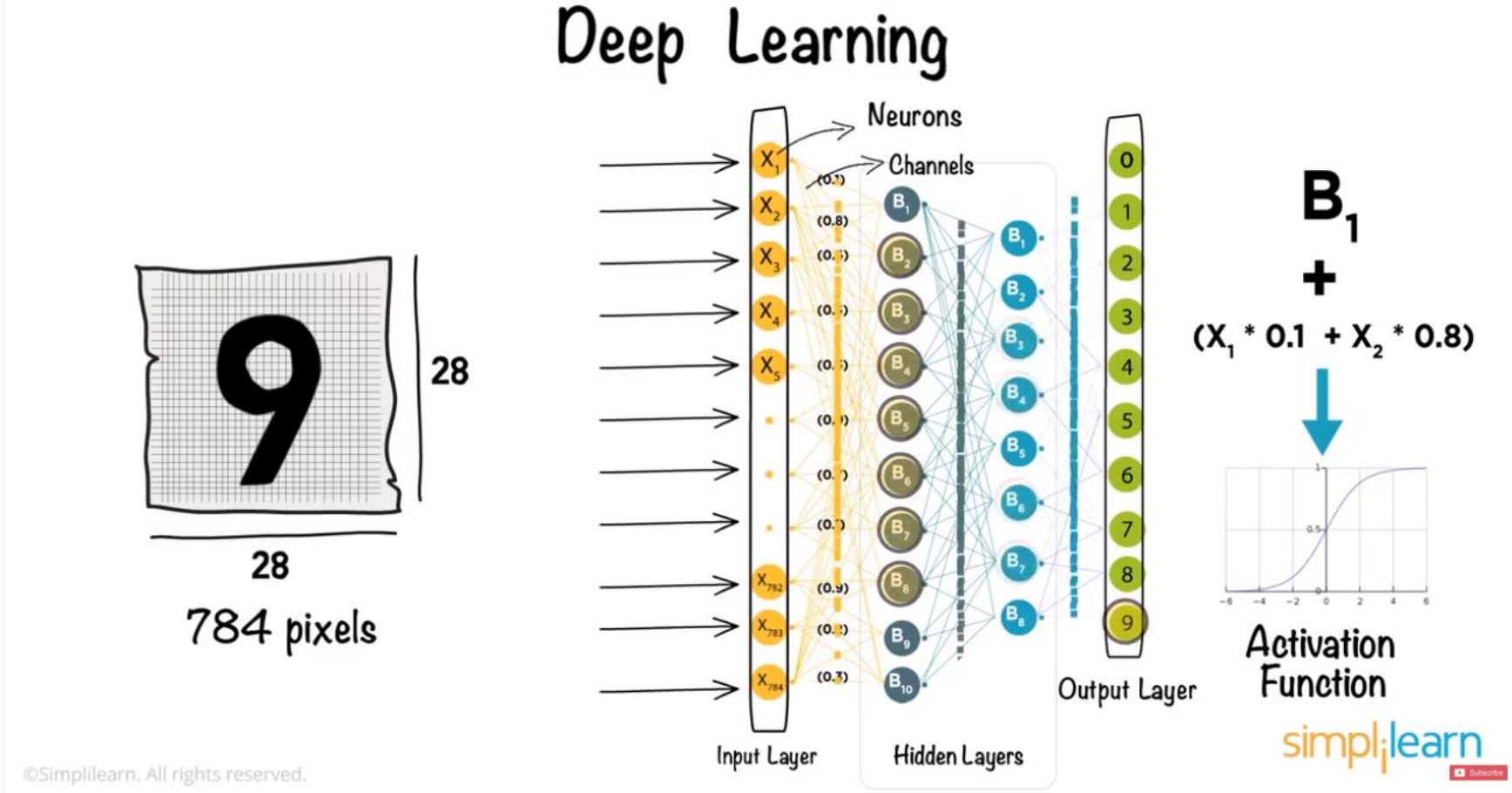
# Today: Outline

- *Pre-lecture Material*
- *Revision*

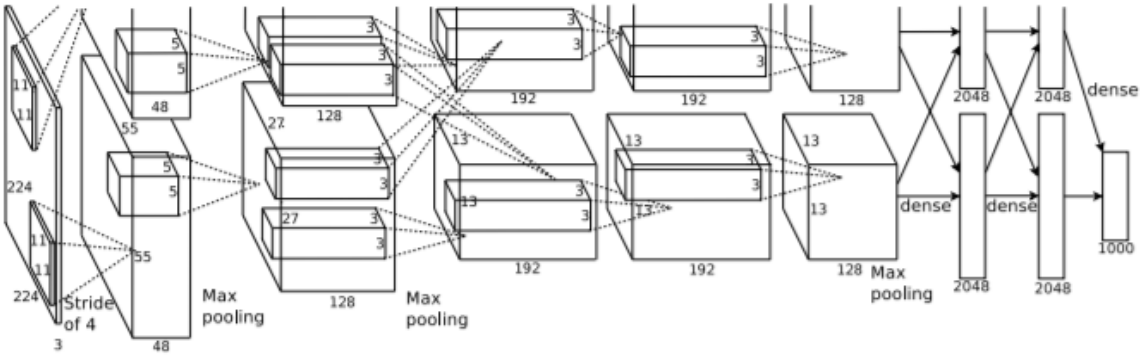
- Reminders:
  - ***Tue Jun 22:*** *Exam during class time*
    *(I sent an email for different timezones)*

  - Practice problems available on Resources

# Pre-lec Material 1: Introduction to Deep Learning

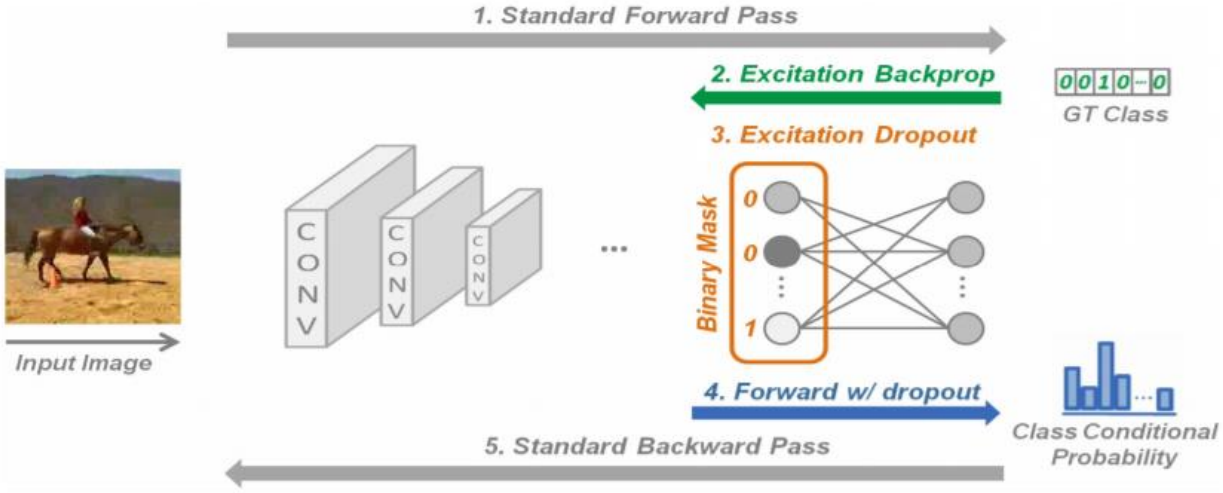


# Pre-lec Material 2: The AlexNet




- ImageNet
- CNN
- Data augmentation for regularization
- Dropout for regularization
- Reporting accuracy (top1/top5)

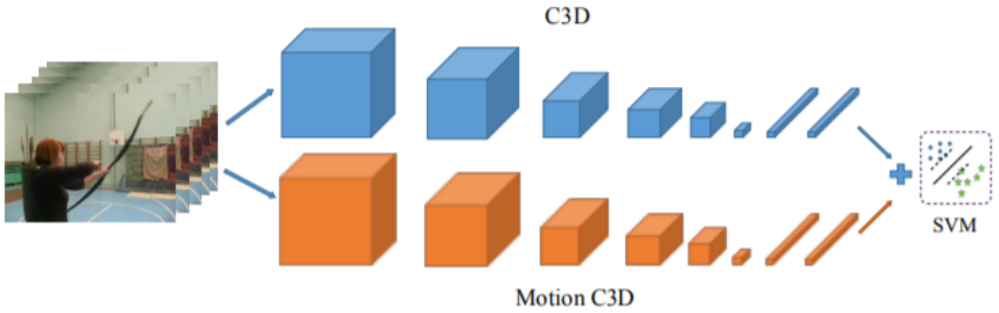
# Pre-lecture 3: Excitation Dropout

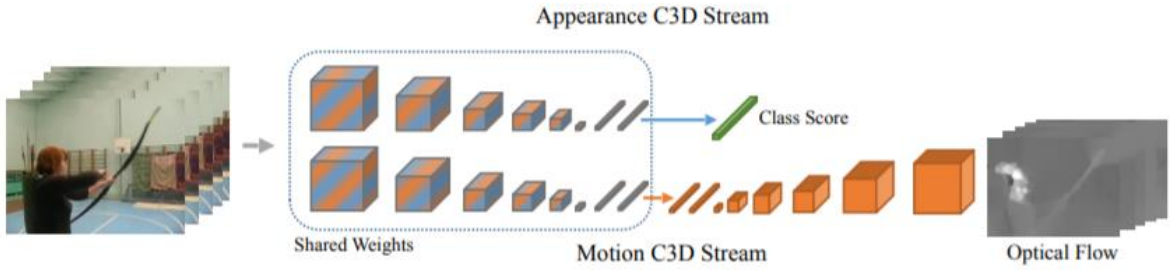


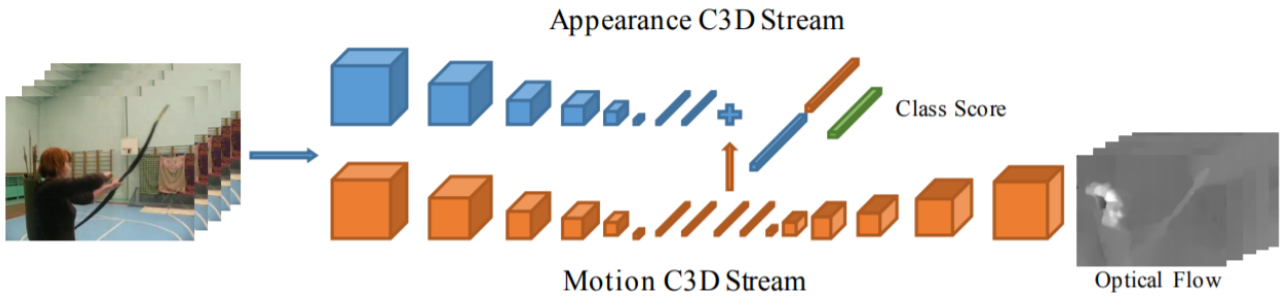

- Leads to:

  1. Better generalization on test data.
  2. Higher utilization of network neurons.
  3. Resilience to network compression.

# Pre-lec Material 4: Multiple Modalities & Auxiliary Tasks



# Pre-lec 5: Grad-CAM

- ***Goal:*** Making CNNs more transparent

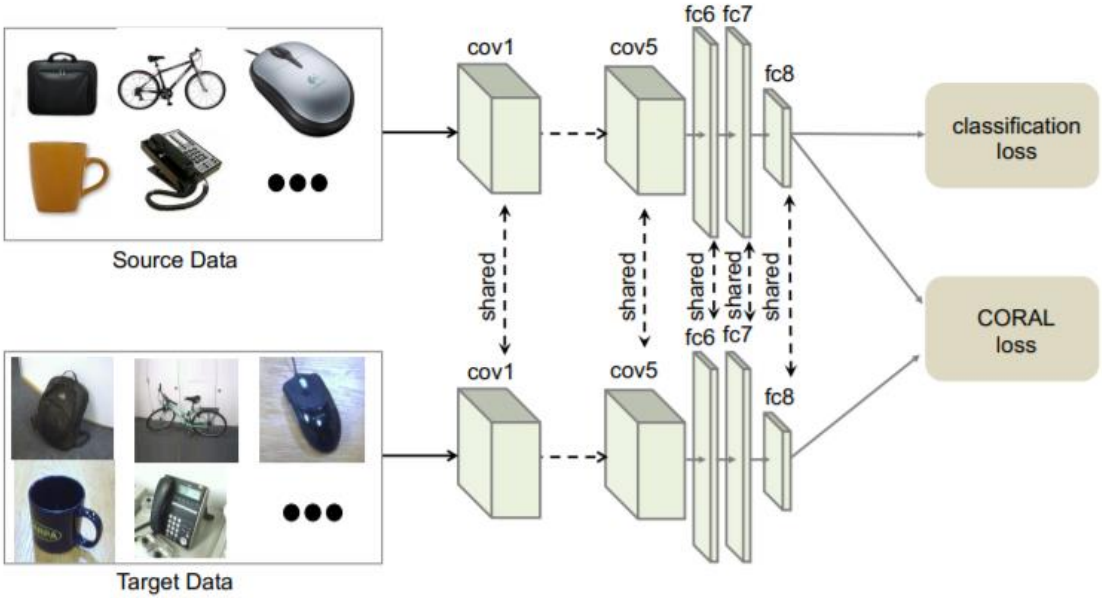
- ***How?*** Uses the gradients of any target concept (say logits for ‘dog’ or even a caption) flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept.

- ***Does it need model parameters?***

# Contrast GradCAM with Input Occlusion

- Grad-CAM uses model parameters -> ***white-box model***
- Input occlusion does not use model parameters -> ***black-box model***

# Pre-lec 6: Domain Adaptation




$$\ell = \ell_{CLASS.} + \sum_{i=1}^{t} \lambda_i \ell_{CORAL}$$

# *Why Now?*

1952 Stochastic Gradient Descent

1958 Perceptron
- Learnable Weights

1986 Backpropagation
- Multi-Layer Perceptron

1995 Deep Convolutional NN
- Digit Recognition

Neural Networks date back decades, so why the resurgence?

## 1. Big Data
- Larger Datasets
- Easier Collection & Storage

IMAGENET

WIKIPEDIA
The Free Encyclopedia

## 2. Hardware
- Graphics Processing Units (GPUs)
- Massively Parallelizable

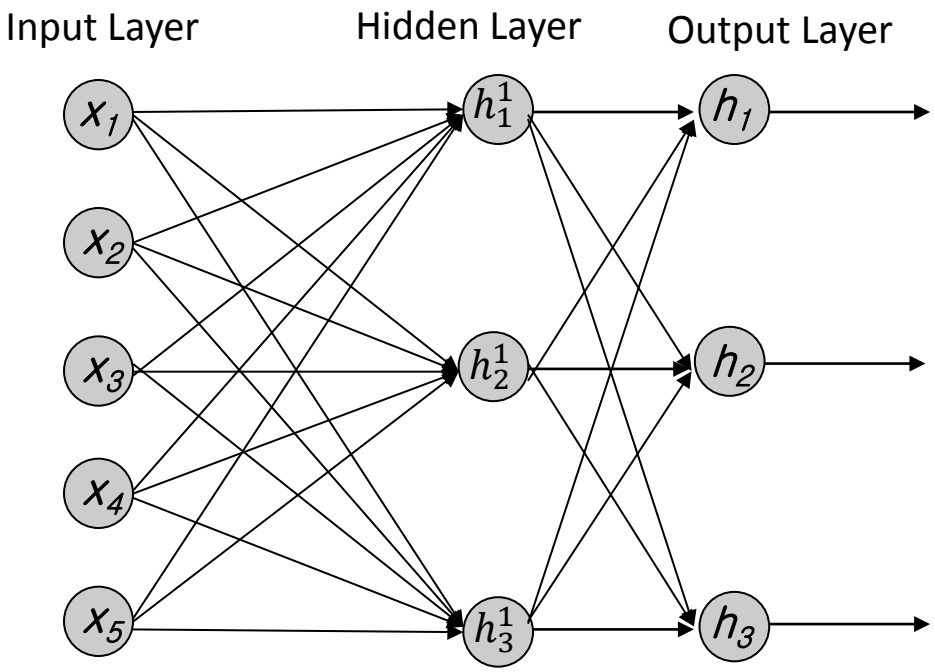
## 3. Software
- Improved Techniques
- New Models
- Toolboxes

TensorFlow

# Artificial Neural Network:

## general notation

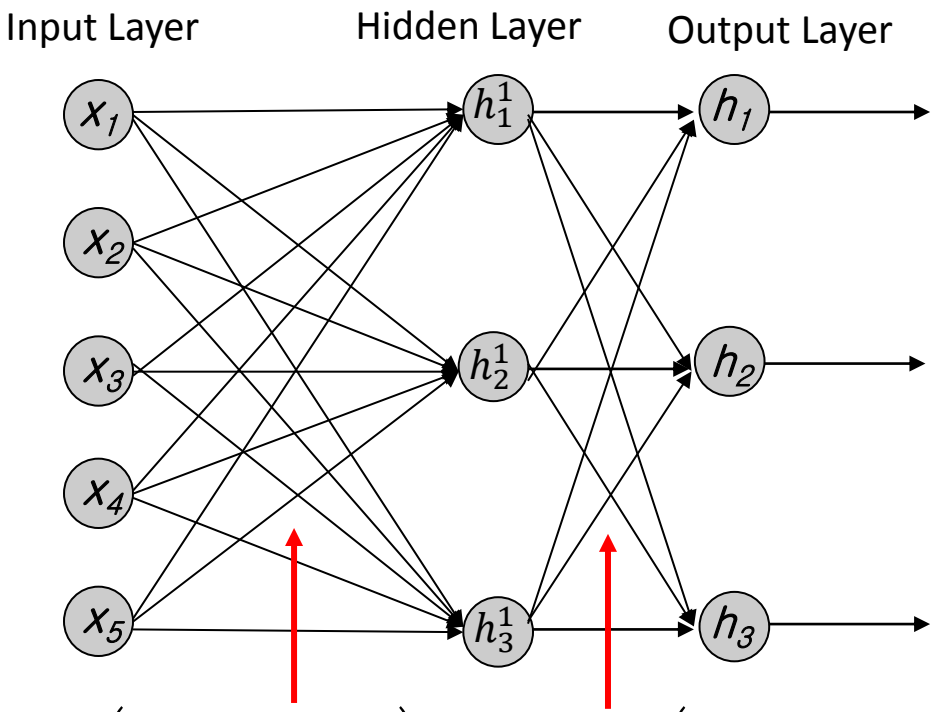
- *How many parameters are presented on this neural network depiction?*



# Artificial Neural Network:

## general notation

- *How many parameters are presented on this neural network depiction?*




15+ 9 = 24 weights

$$\Theta^{(1)} = \begin{pmatrix} \theta_{11} & \cdots & \theta_{15} \\ \vdots & \ddots & \vdots \\ \theta_{31} & \cdots & \theta_{35} \end{pmatrix} \qquad \Theta^{(2)} = \begin{pmatrix} \theta_{11} & \cdots & \theta_{13} \\ \vdots & \ddots & \vdots \\ \theta_{31} & \cdots & \theta_{33} \end{pmatrix}$$

# Cost function

- *Mark the terms responsible for the task error, and the terms responsible for regularization.*
- *What task is this Cost function modeling?*

Neural network: $h_\Theta(x) \in \mathbb{R}^K \; (h_\Theta(x))_i = i^{th} \text{ output}$

$$J(\Theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log(h_\Theta(x^{(i)}))_k + (1 - y_k^{(i)}) \log(1 - (h_\Theta(x^{(i)}))_k)\right]$$
$$+\frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\Theta_{ji}^{(l)})^2$$

# Cost function

Neural network: $h_\Theta(x) \in \mathbb{R}^K \ (h_\Theta(x))_i = i^{th}$ output

task error

$$J(\Theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log(h_\Theta(x^{(i)}))_k + (1 - y_k^{(i)}) \log(1 - (h_\Theta(x^{(i)}))_k)\right]$$

$$+\frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\Theta_{ji}^{(l)})^2$$

regularization

# Cost function

- *How is  this function used to help the network Learn the task?*

Neural network: $h_\Theta(x) \in \mathbb{R}^K \ (h_\Theta(x))_i = i^{th} \text{ output}$

$$J(\Theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log(h_\Theta(x^{(i)}))_k + (1 - y_k^{(i)}) \log(1 - (h_\Theta(x^{(i)}))_k)\right]$$
$$+\frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\Theta_{ji}^{(l)})^2$$

$$J(\Theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log h_\theta(x^{(i)})_k + (1 - y_k^{(i)}) \log(1 - h_\theta(x^{(i)})_k)\right]$$

$$+\frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\Theta_j^{(l)})^2$$

$$\min_{\Theta} J(\Theta)$$

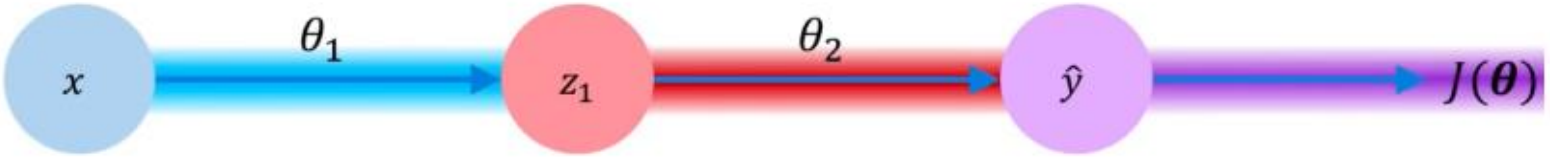
Need code to compute:

- $J(\Theta)$
- $\frac{\partial}{\partial \Theta_{ij}^{(l)}} J(\Theta)$

***Backpropagation***

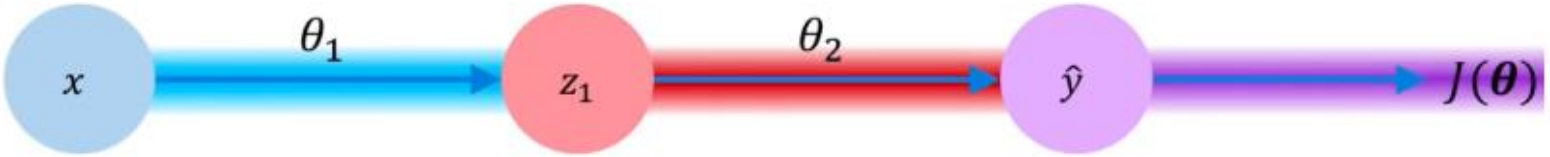
# Computing Gradients: Backpropagation

- How does a small change in $\theta_1$ affect the final loss $J(\theta)$?



# Computing Gradients: Backpropagation

- How does a small change in $\theta_1$ affect the final loss $J(\theta)$?



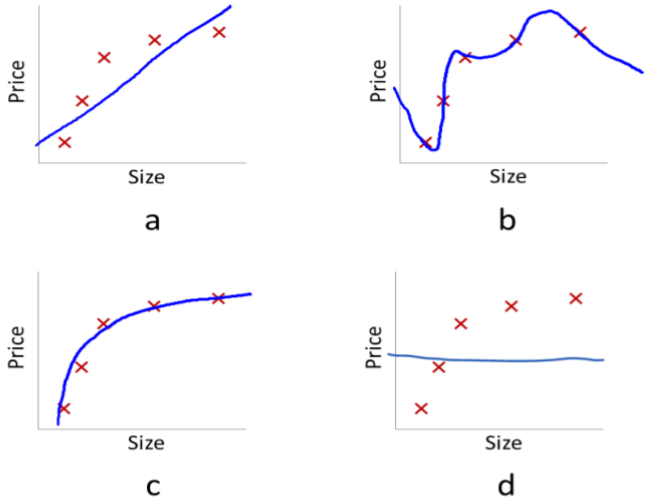

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_1} = \frac{\partial J(\boldsymbol{\theta})}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial z_1} * \frac{\partial z_1}{\partial \theta_1}$$

# Which Lambda goes with each plot?

Alice is trying to fit a linear regression model to predict house price based on size using polynomial features. Since her training dataset is very small, she is applying regularization. She fit several models by minimizing the cost function

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n}\theta_j^2\right]$$

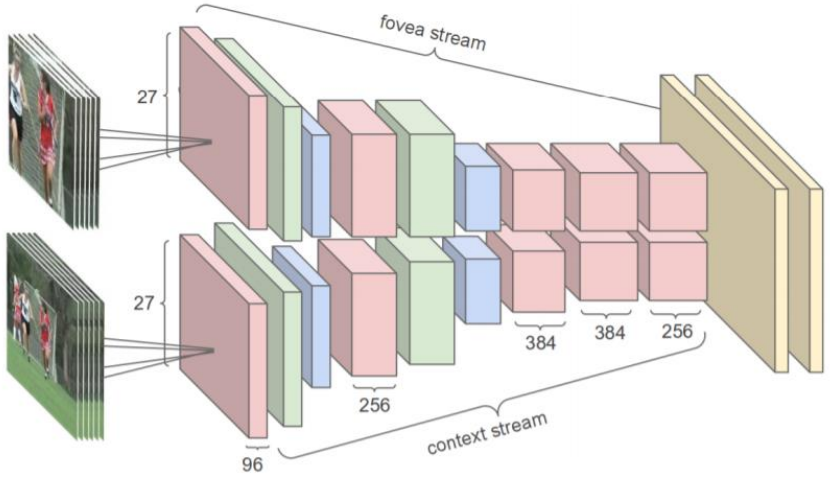
for $\lambda = 10^0, 10^1, 10^2, 10^3$. The following are sketches of the resulting models.




$a$: $10^2$ $b$: $10^0$ $c$: $10^1$ $d$: $10^3$

# Design Question

- Human eyes focus on the central part of the field of view (fovea), but also have peripheral vision (context). How would you design a DNN to incorporate this functionality?

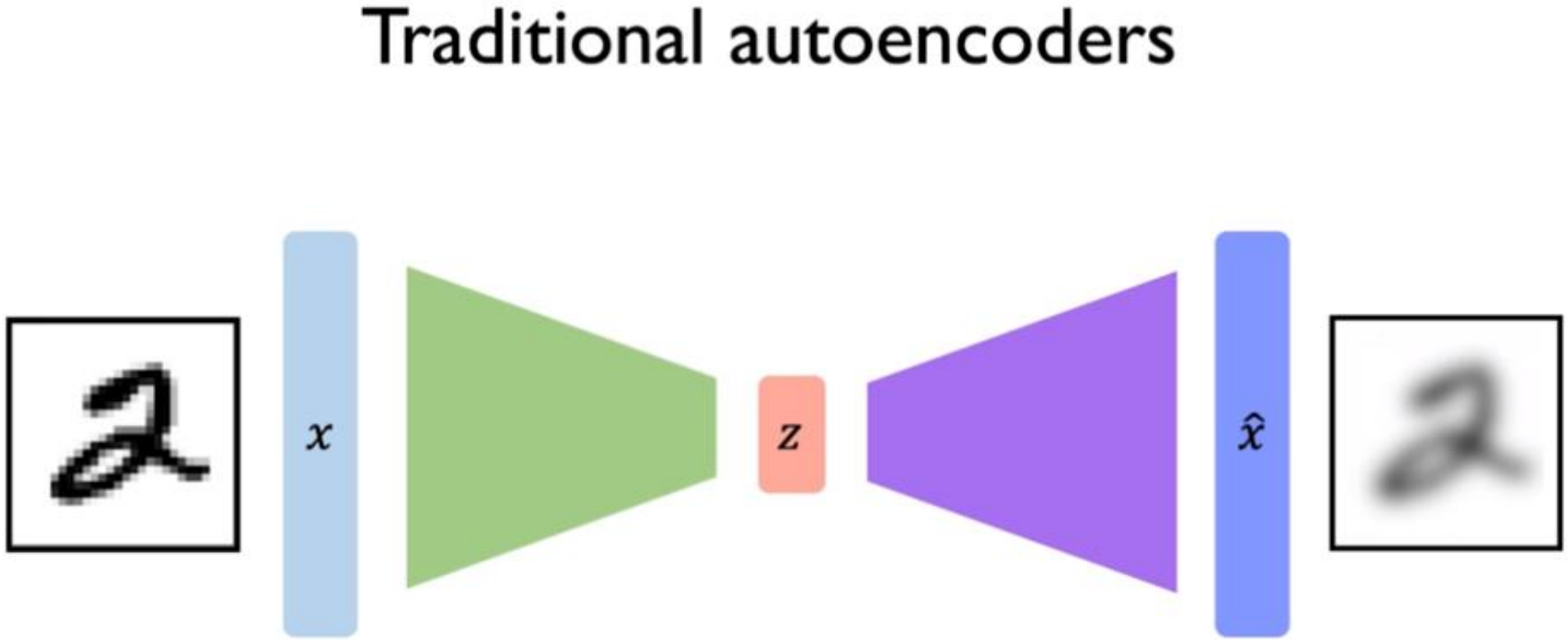


# Density Estimation

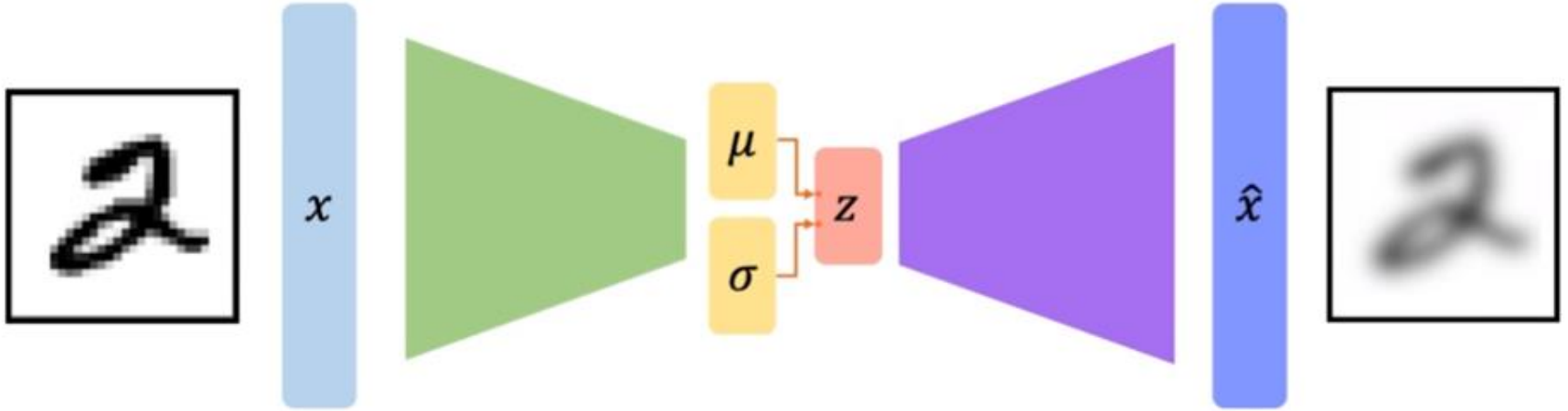
- Give three applications of models that are able to perform density estimation:

  - Anomaly Detection
  - De-noising Input
  - Generation of new samples
    (e.g.: can be used for de-biasing Models)

# What is the difference between a VAE and a traditional autoencoder?

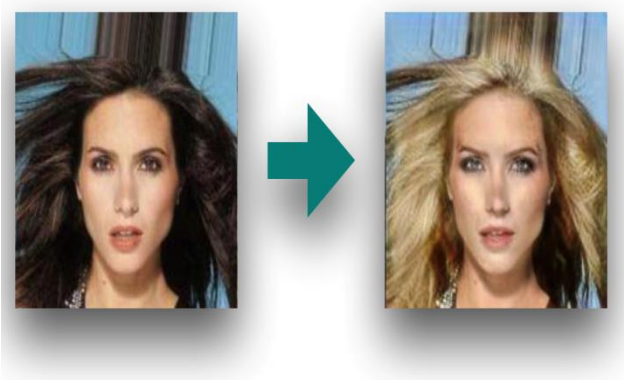


# VAEs: key difference with traditional autoencoder

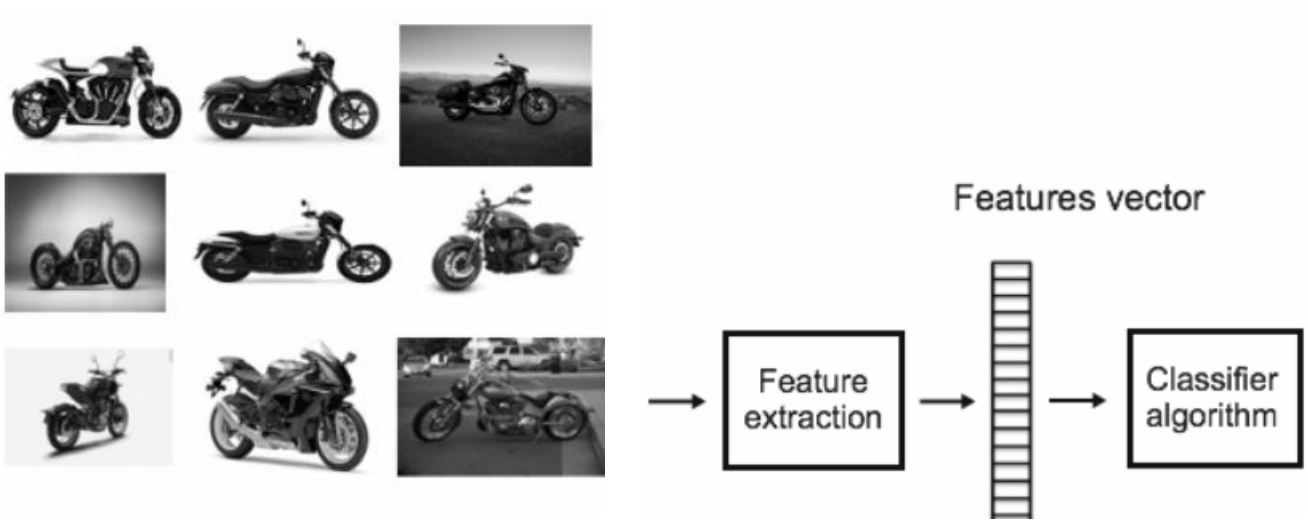


# What is a Deepfake?

- A manipulation of a person's image or video, for example: modifying expressions, poses, age, hair color, gender, or other attributes of the person.
- Given the widespread availability of these systems, malicious actors can modify images of a person without their consent which raises many ethical concerns.

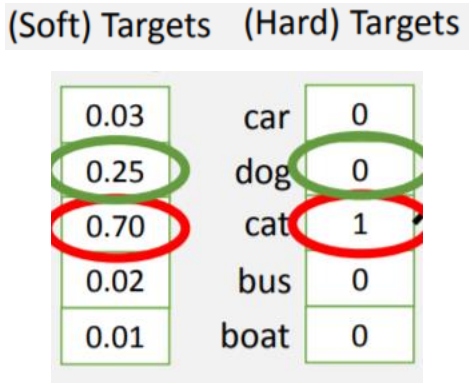


# How can we extract features for the following motorcycle images?

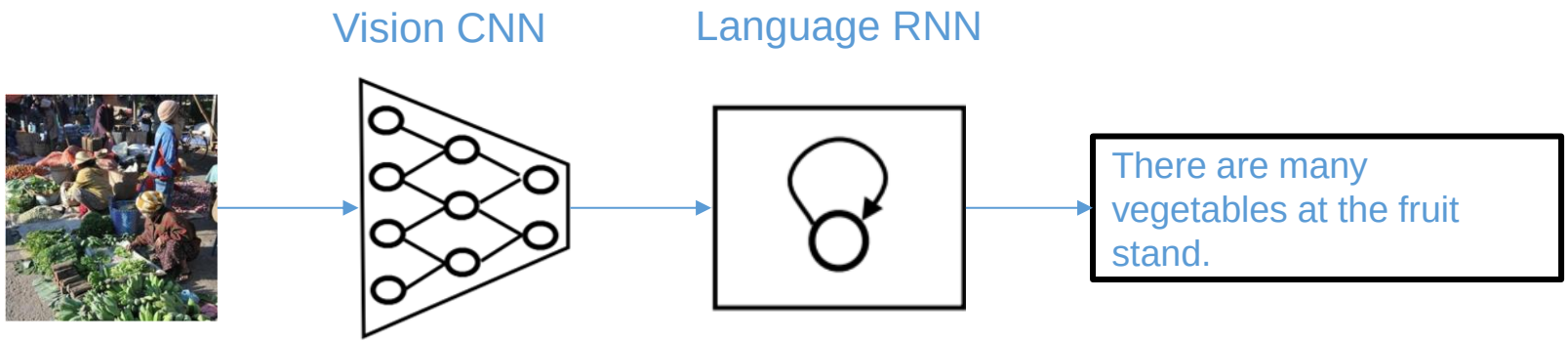


# Describe the benefit of using soft over hard targets in knowledge distillation

- Hard targets have no information about wrong classes
- Soft targets have information about wrong classes



# What is the name of the task the following architecture addresses?




- Image Captioning

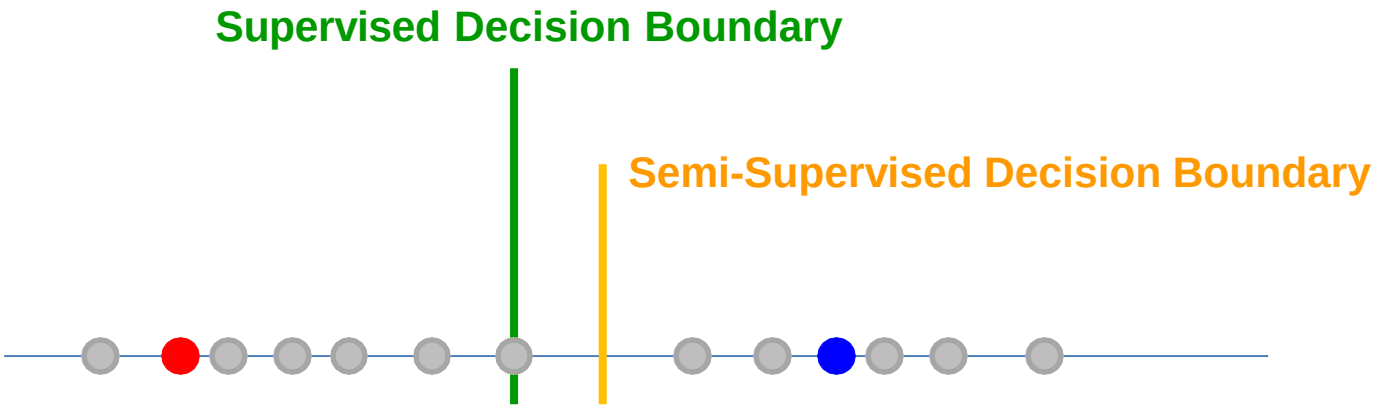
# Semi-supervised Learning

- Can adding unlabeled data improve supervised performance?

Positive labeled data

Negative labeled data

Unlabeled data



# Contrast Self-Training and Co-Training

- Self-Training

Input: labeled data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$, unlabeled data $\{\mathbf{x}_j\}_{j=l+1}^{l+u}$.

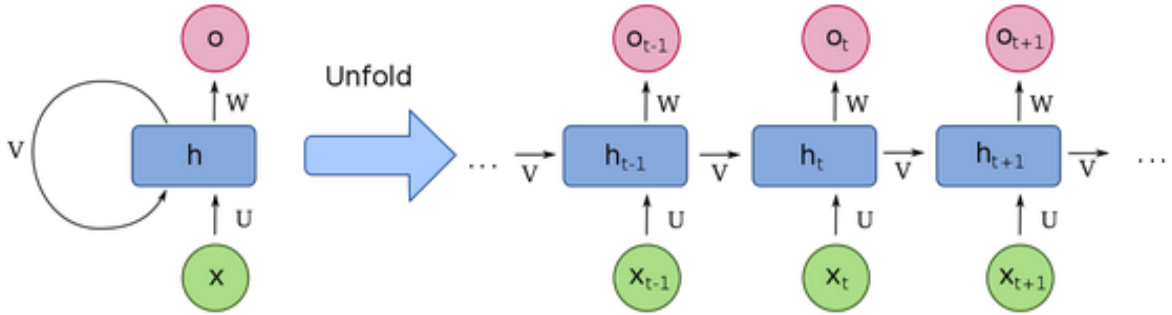
1. Initially, let $L = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ and $U = \{\mathbf{x}_j\}_{j=l+1}^{l+u}$.
2. Repeat:
3. Train $f$ from $L$ using supervised learning.
4. Apply $f$ to the unlabeled instances in $U$.
5. Remove a subset $S$ from $U$; add $\{(\mathbf{x}, f(\mathbf{x}))|\mathbf{x} \in S\}$ to $L$.

- Co-Training

1. let $L_1 = L_2 = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}$.
2. Repeat until unlabeled data is used up:
3. Train view-1 $f^{(1)}$ from $L_1$, view-2 $f^{(2)}$ from $L_2$.
4. Classify unlabeled data with $f^{(1)}$ and $f^{(2)}$ separately.
5. Add $f^{(1)}$'s top $k$ most-confident predictions $(\mathbf{x}, f^{(1)}(\mathbf{x}))$ to $L_2$.
   Add $f^{(2)}$'s top $k$ most-confident predictions $(\mathbf{x}, f^{(2)}(\mathbf{x}))$ to $L_1$.
   Remove these from the unlabeled data.

# How is backpropagation different in RNN compared to an ANN?

In Recurrent Neural Networks, we have an additional loop at each node:

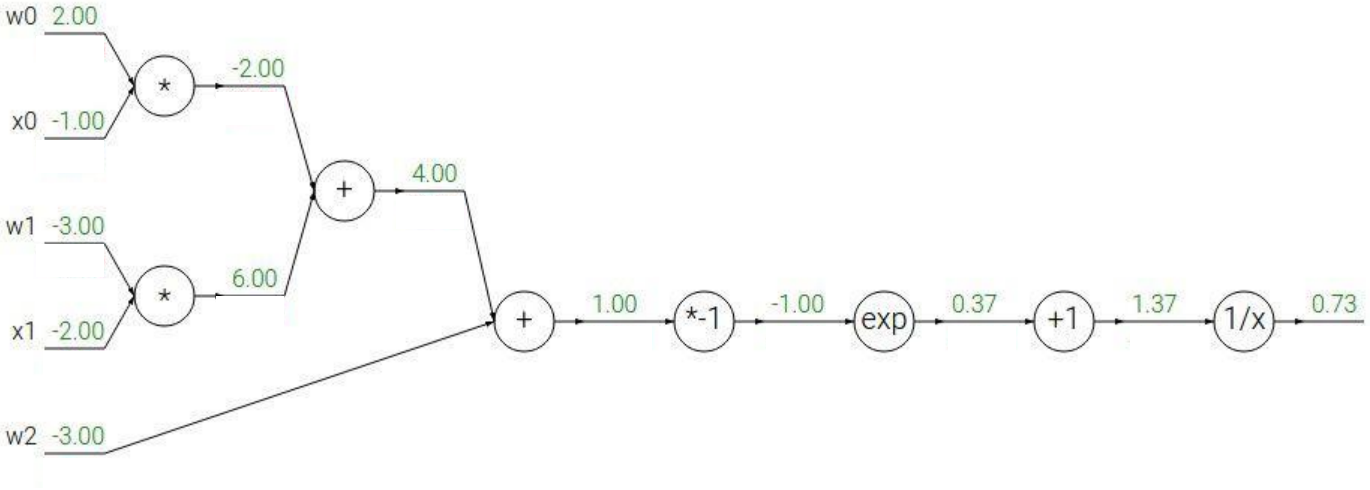



This loop essentially includes a time component into the network as well. This helps in capturing sequential information from the data, which could not be possible in a generic artificial neural network.

This is why the backpropagation in RNN is called Backpropagation through Time, as in backpropagation at each time step.

# Another example:

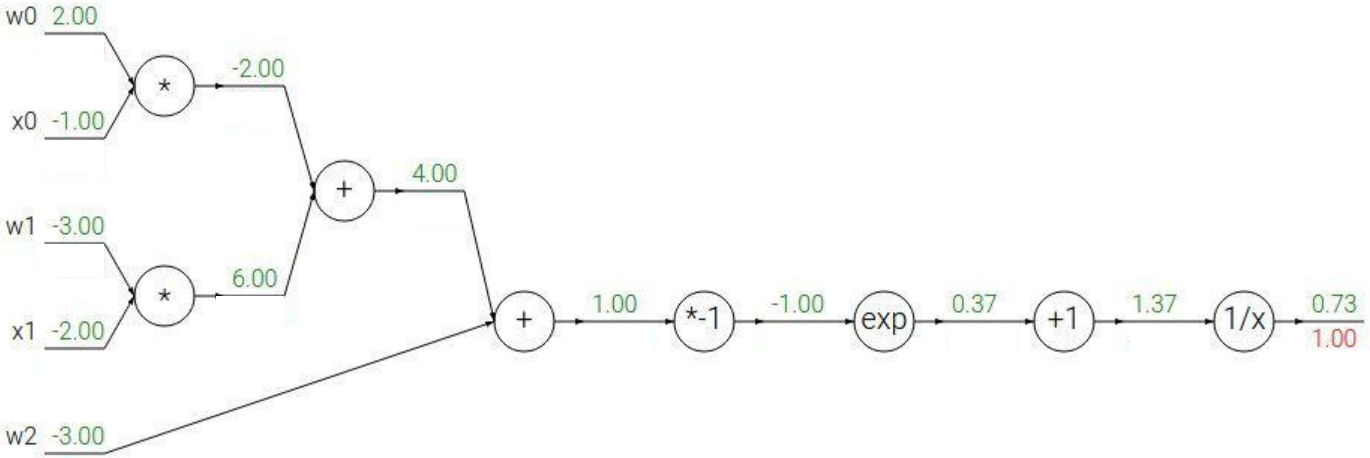
$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



# Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

*Computing a 2D Sigmoid Neuron!*




$$f(x) = e^x \rightarrow \frac{df}{dx} = e^x$$

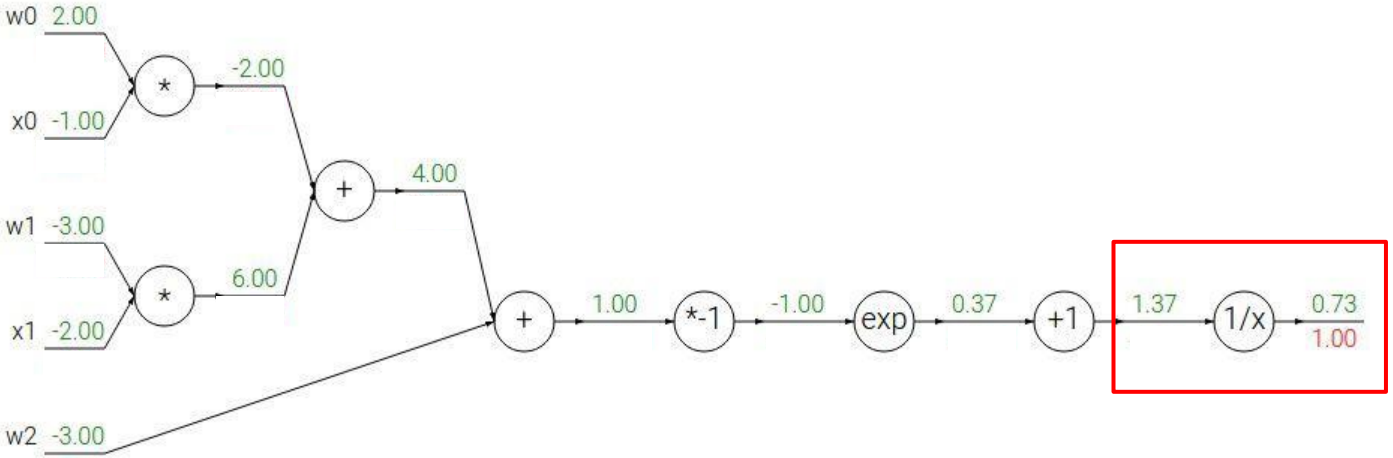
$$f_a(x) = ax \rightarrow \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \rightarrow \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \rightarrow \frac{df}{dx} = 1$$

# Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$




$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

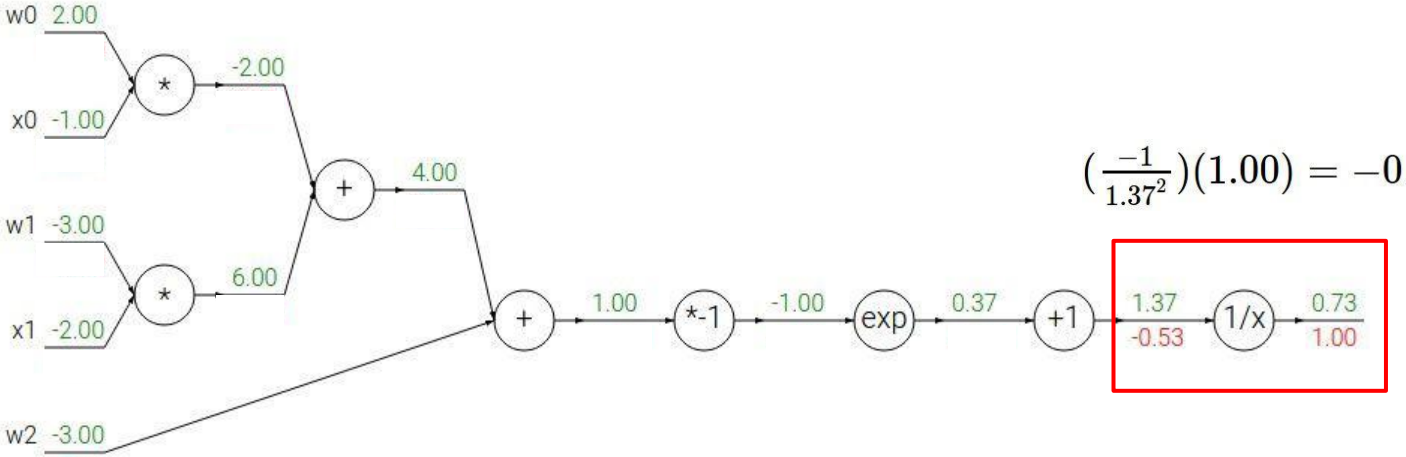
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

# Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$




$$(\frac{-1}{1.37^2})(1.00) = -0.53$$

$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

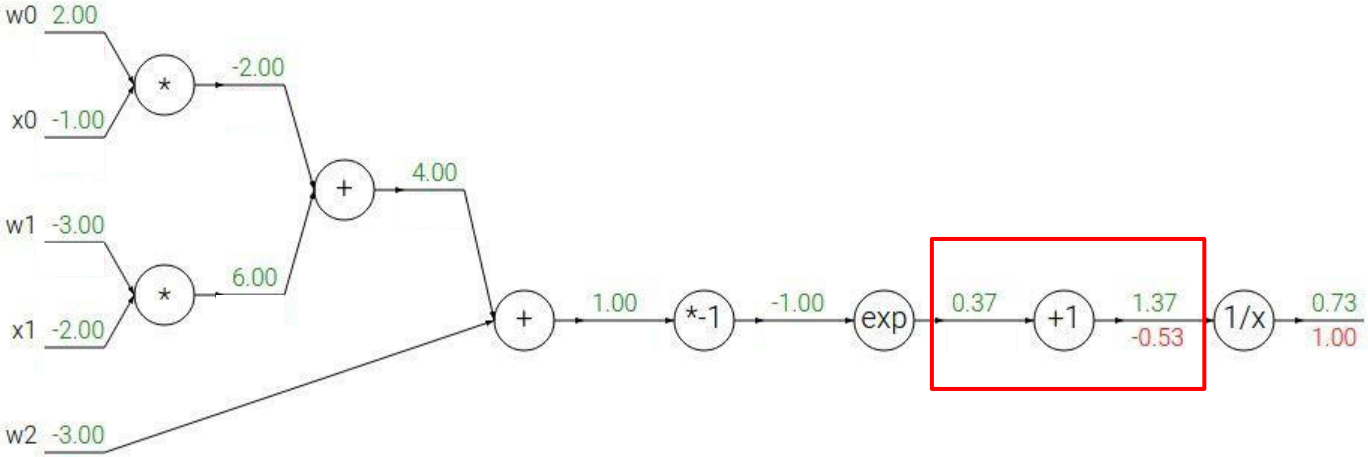
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

# Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$




$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

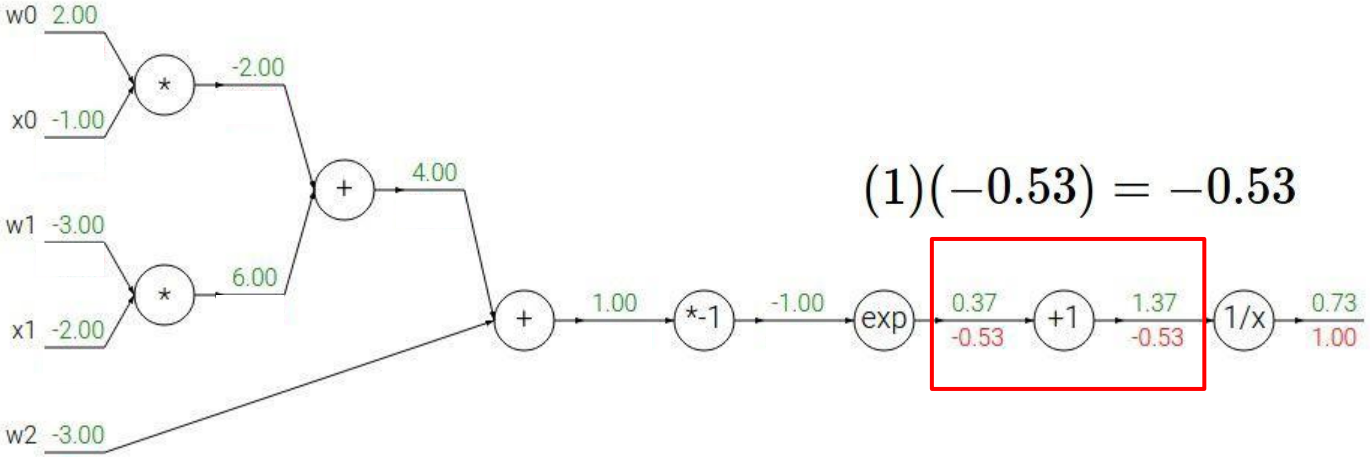
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

# Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$




$$(1)(-0.53) = -0.53$$

$f(x) = e^x \rightarrow \frac{df}{dx} = e^x$

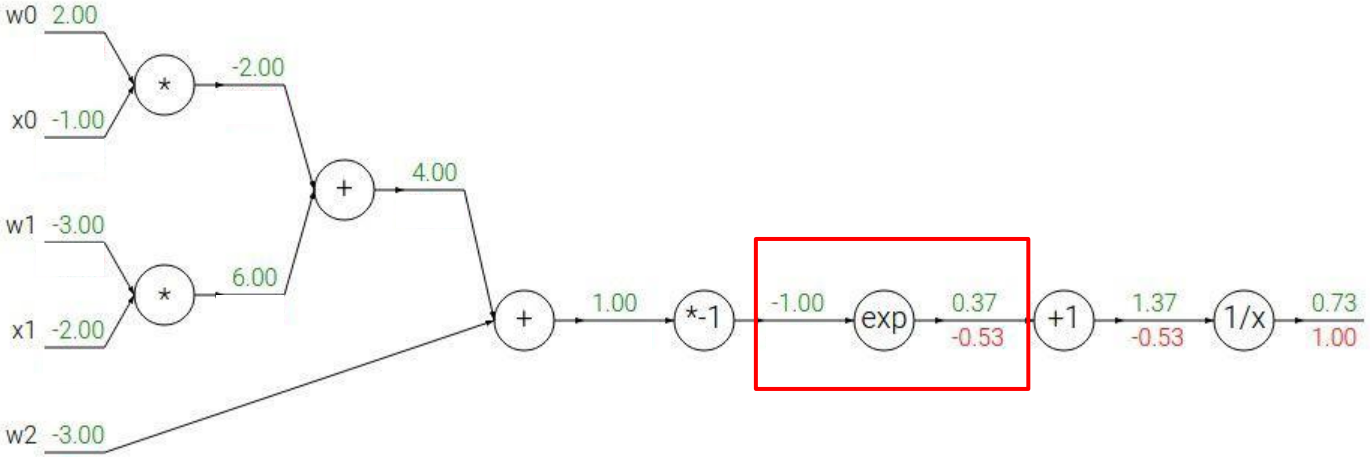
$f_a(x) = ax \rightarrow \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \rightarrow \frac{df}{dx} = -1/x^2$

$f_c(x) = c + x \rightarrow \frac{df}{dx} = 1$

# Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$




$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

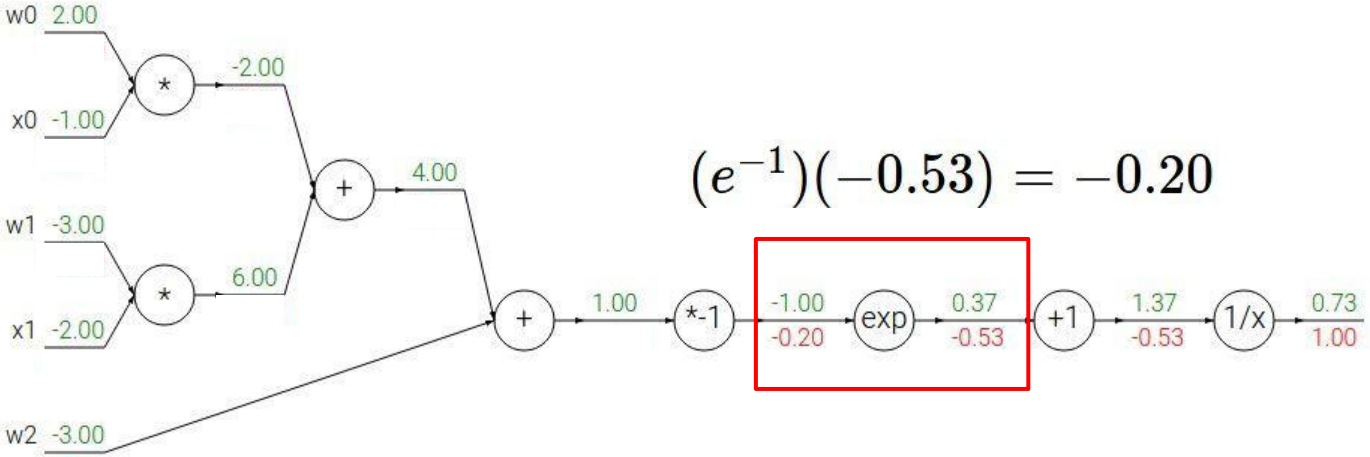
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

# Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$




$$(e^{-1})(-0.53) = -0.20$$

$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

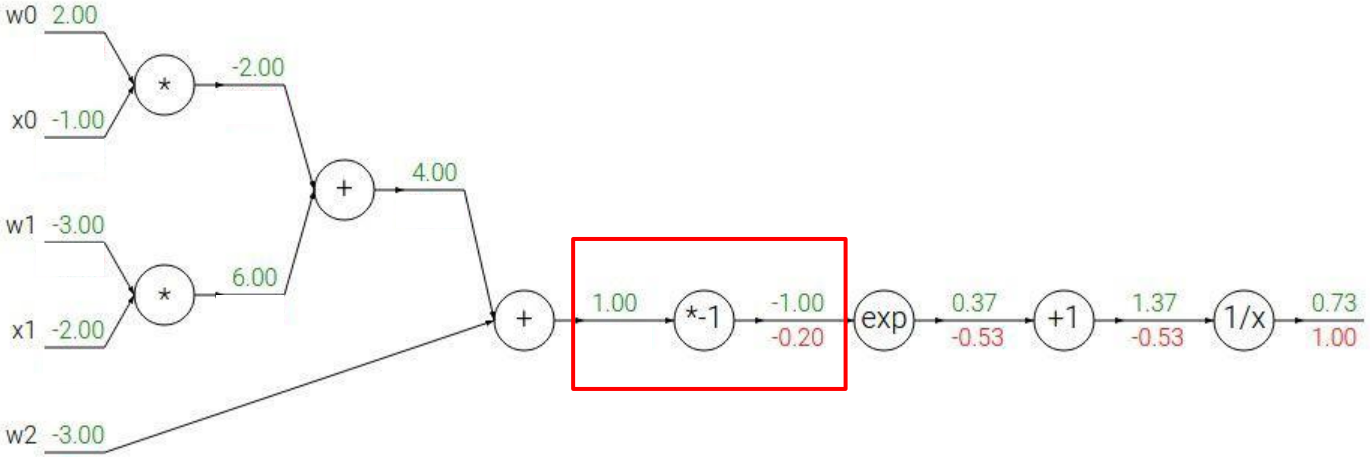
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

# Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$




$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

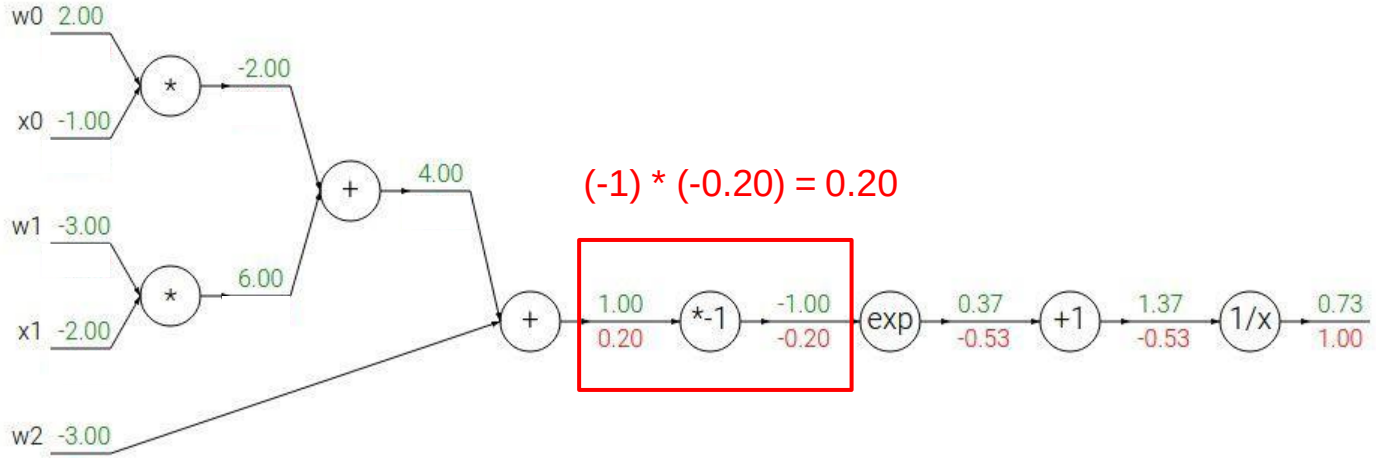
$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

# Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$




(-1) * (-0.20) = 0.20

$$f(x) = e^x \rightarrow \frac{df}{dx} = e^x$$

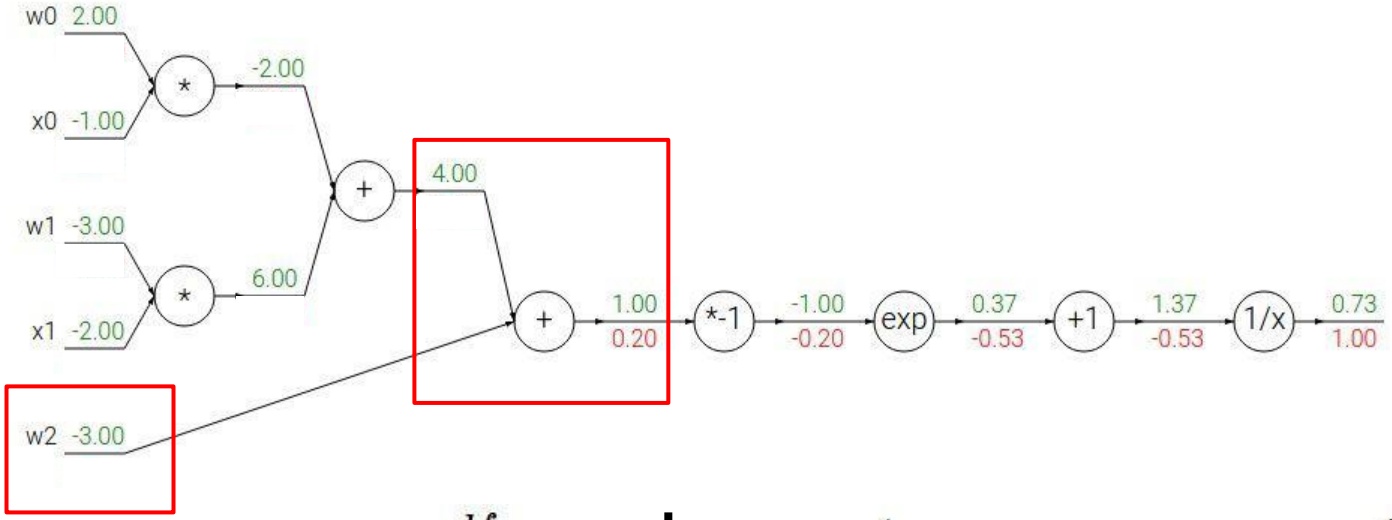
$$f_a(x) = ax \rightarrow \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \rightarrow \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \rightarrow \frac{df}{dx} = 1$$

# Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$




$$f(x) = e^x \rightarrow \frac{df}{dx} = e^x$$

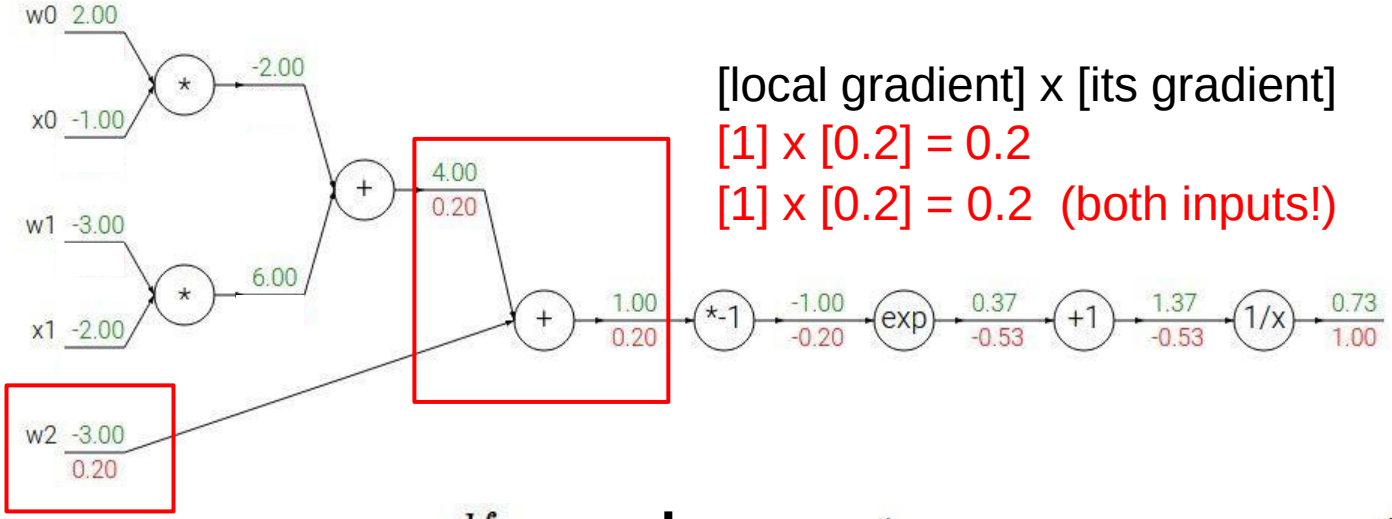
$$f_a(x) = ax \rightarrow \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \rightarrow \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \rightarrow \frac{df}{dx} = 1$$

# Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$




[local gradient] x [its gradient]

[1] x [0.2] = 0.2

[1] x [0.2] = 0.2 (both inputs!)

$f(x) = e^x \rightarrow \frac{df}{dx} = e^x$

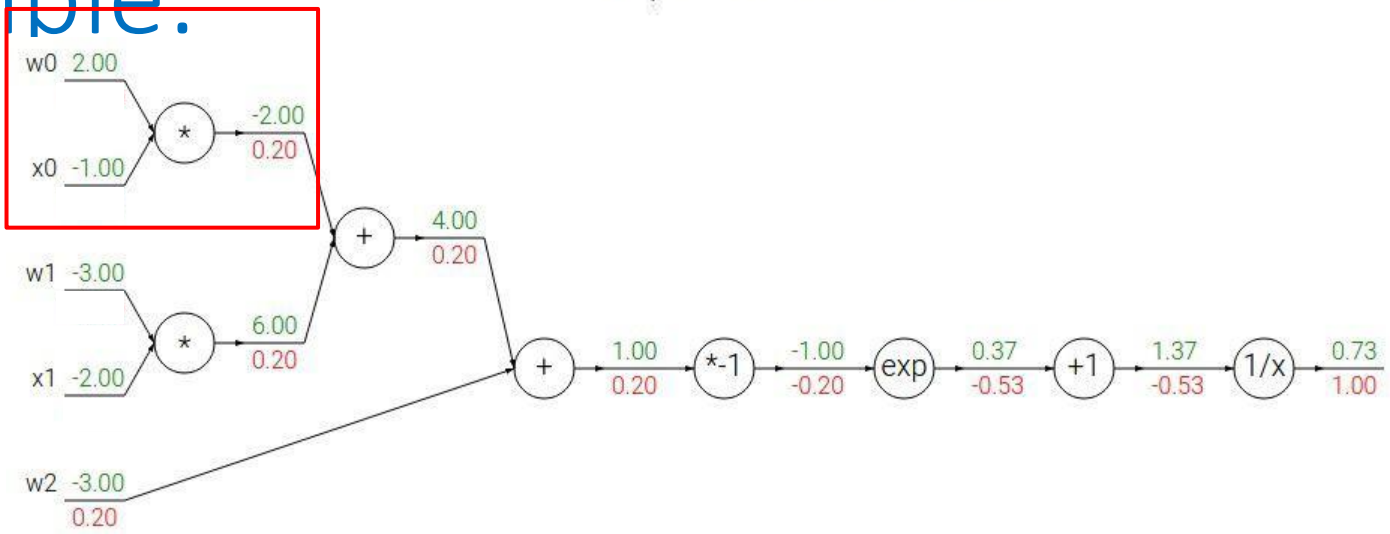
$f_a(x) = ax \rightarrow \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \rightarrow \frac{df}{dx} = -1/x^2$

$f_c(x) = c + x \rightarrow \frac{df}{dx} = 1$

# Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$




$$f(x) = e^x \rightarrow \frac{df}{dx} = e^x$$

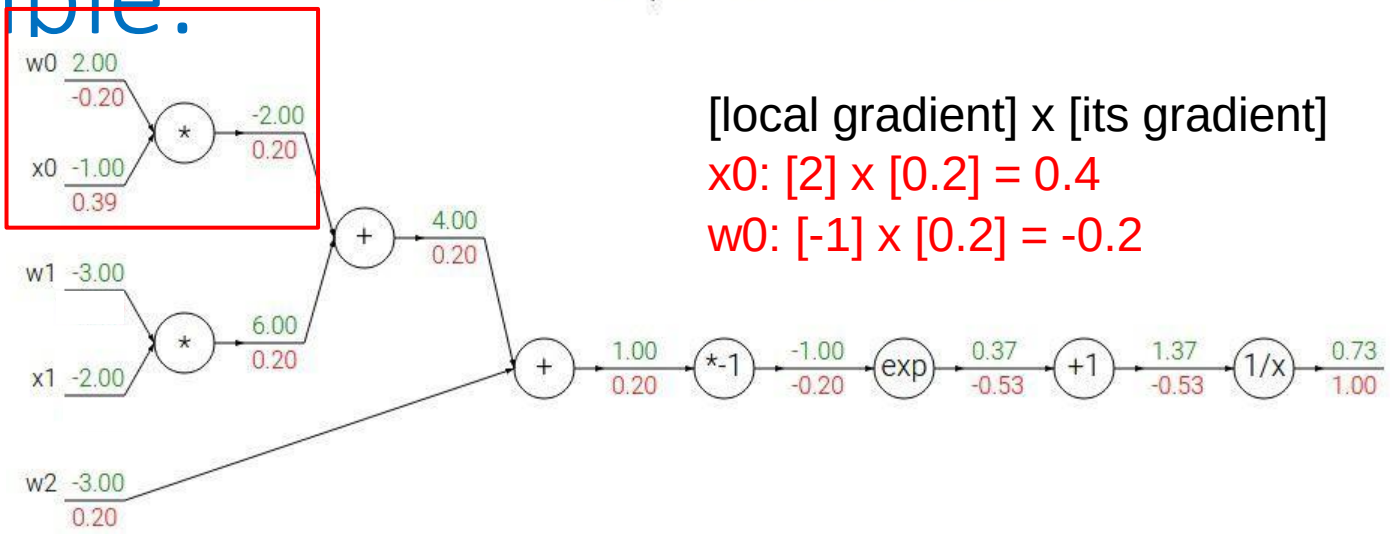
$$f_a(x) = ax \rightarrow \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \rightarrow \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \rightarrow \frac{df}{dx} = 1$$

# Another example:

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$




[local gradient] x [its gradient]

x0: [2] x [0.2] = 0.4

w0: [-1] x [0.2] = -0.2

$$f(x) = e^x \rightarrow \frac{df}{dx} = e^x$$

$$f_a(x) = ax \rightarrow \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \rightarrow \frac{df}{dx} = -1/x^2$$

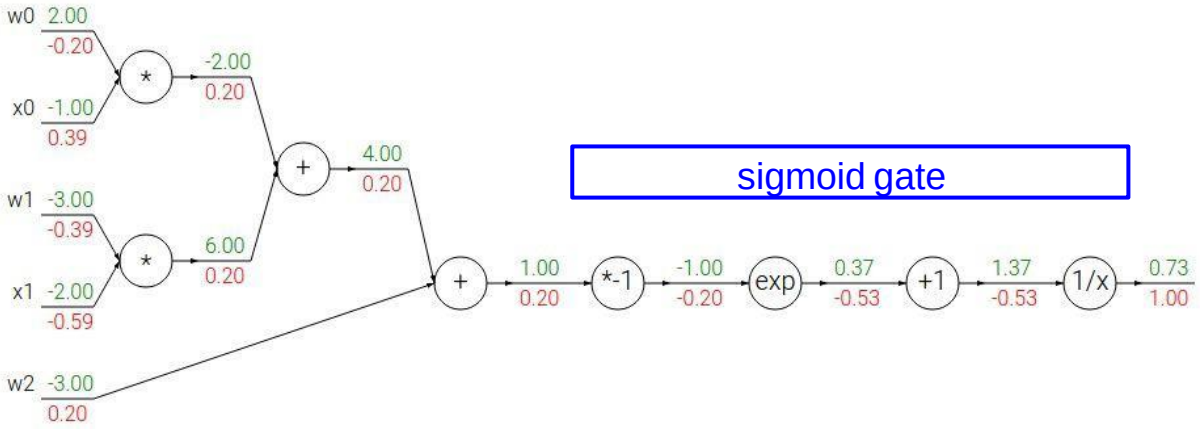
$$f_c(x) = c + x \rightarrow \frac{df}{dx} = 1$$

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

sigmoid function

$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = \left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = (1-\sigma(x))\,\sigma(x)$$

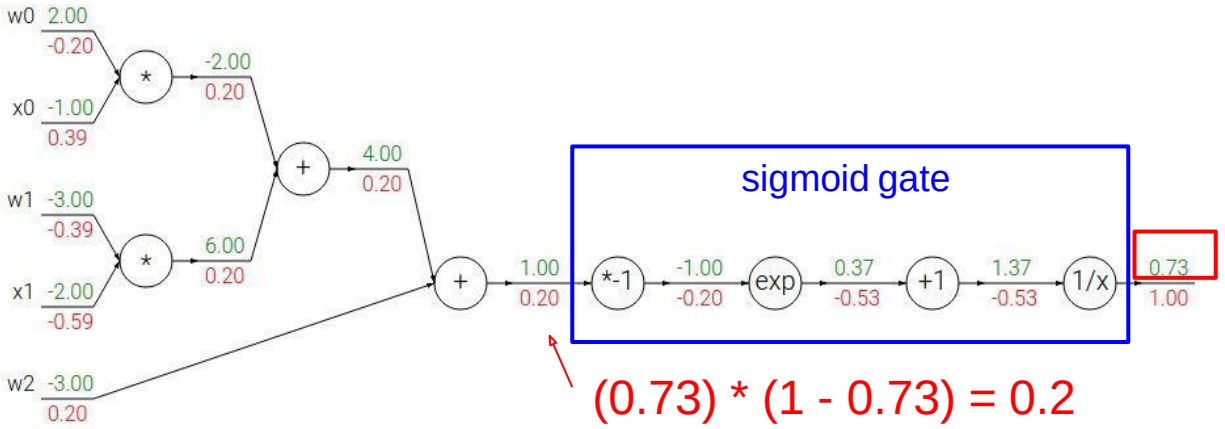


$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

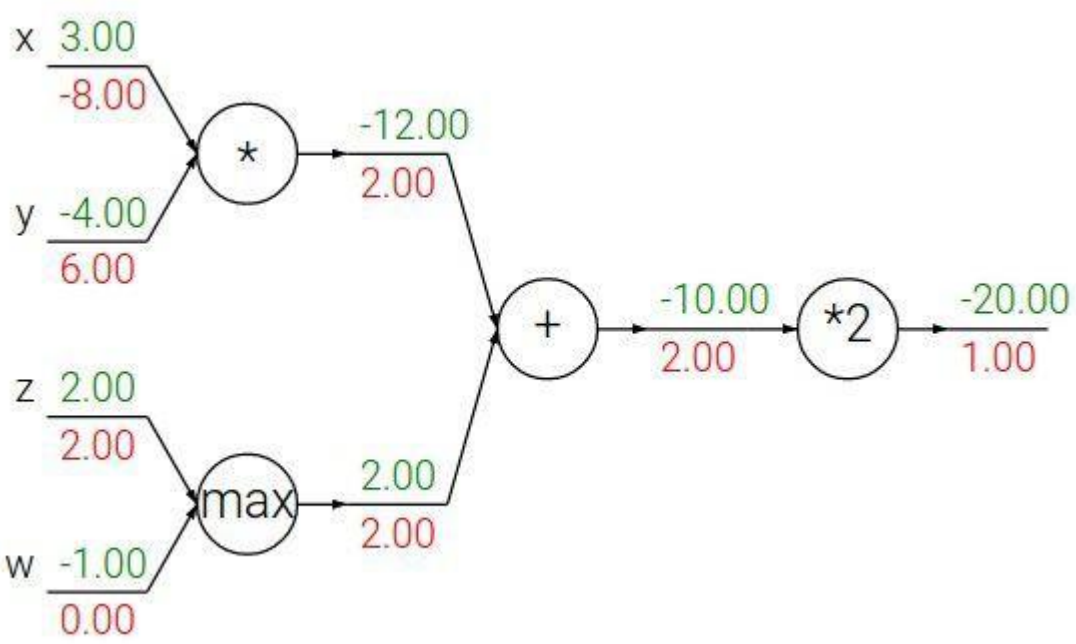
sigmoid function

$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = \left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = (1-\sigma(x))\,\sigma(x)$$



# Patterns in backward flow

**add** gate: gradient distributor
**max** gate: gradient router
**mul** gate: gradient… “switcher”?



# CS 523 – Deep Learning

# Exam

# Summer 2021

***Instructions:***

1- Log onto the lecture zoom link
2- Share your video and audio (make sure your hands are fully in the camera's field of view)
3- Set an alarm for 2:45pm. You have 1 hour and 40 minutes to solve the exam.
4- Solve the exam using paper and pen
5- Print your name and BU ID clearly on the top right of the first page (the page that will have the solution to Problem 1)
6- Start a new page for each question
7- Stop solving at 2:45pm
8- Take photos of your solutions using your phone
9- Sort the photos based on question number
10- Convert the photos into a single pdf
11- Submit your pdf file on GradeScope
12- You will receive a confirmation message from us that we received your submission. **It is your responsibility to make sure the file contains solutions to all the problems you solved.**