"Machine learning is the last invention that humanity will ever need to make"

Deep Learning - 2

Abhinav Gupta

May 25, 2021

Robotics Research Center
IIIT Hyderabad

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Robotics Research Center

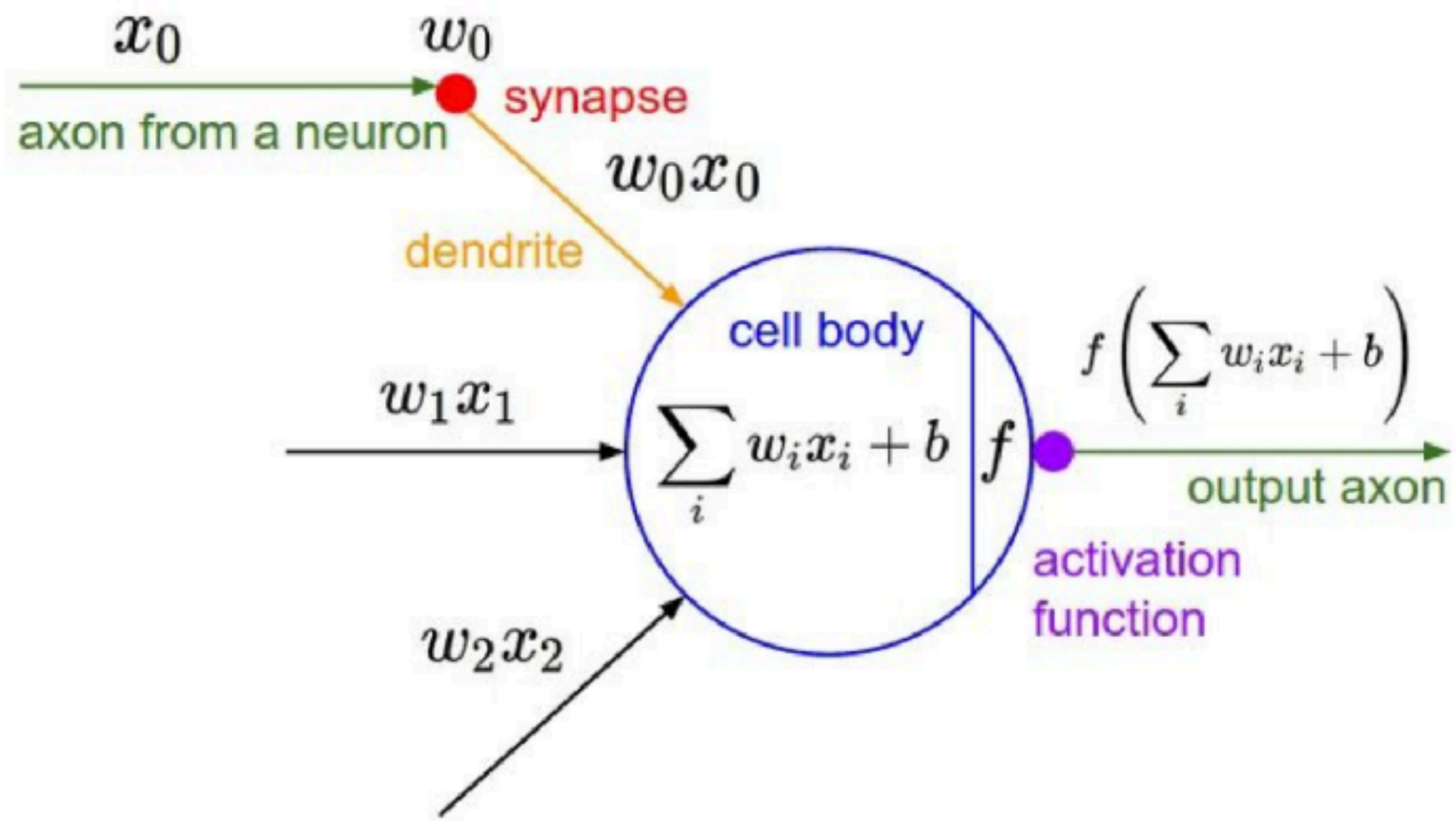
Adapted from Stanford's CS231 course, all images belong to them.

Today's Lecture

- Backpropagation Math
- Optimisers
- Learning Rates
- Beyond Training Error - Regularization

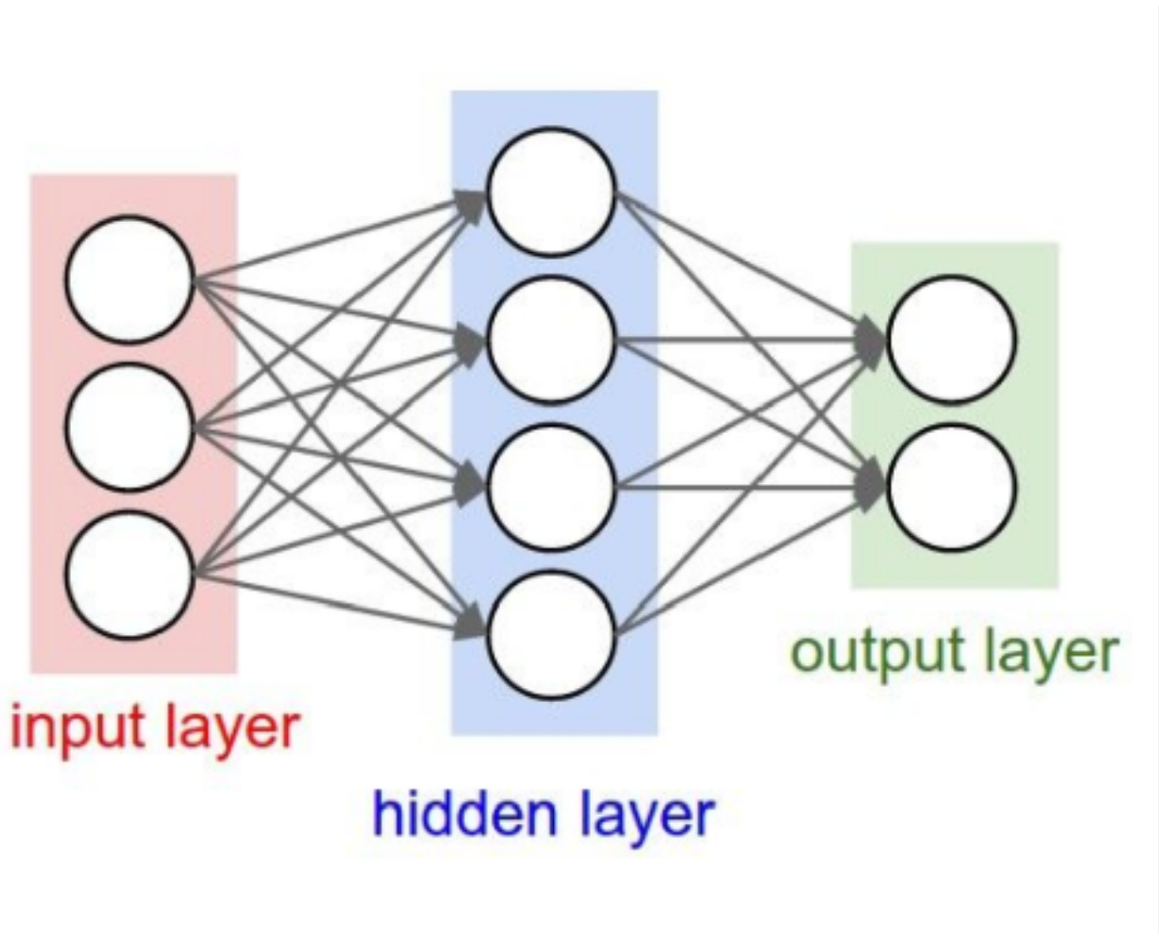
Robotics Research Center

Just a quick recap…



Robotics Research Center

A neural network is a computational graph, with many linear layers and non linearities in between.



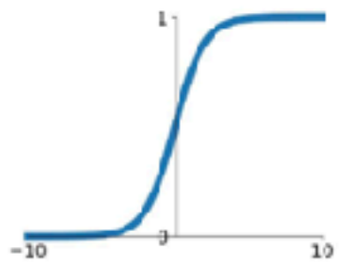
We wanna learn the values of these weights, and we do so through optimisation.

Robotics Research Center

Activation Functions

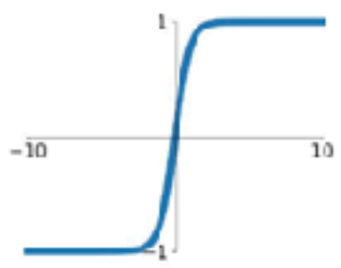
Sigmoid

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



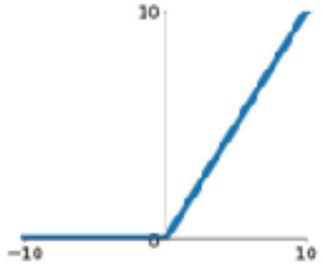
tanh

$$\tanh(x)$$



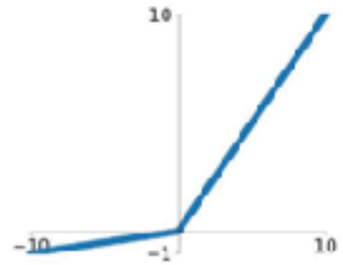
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

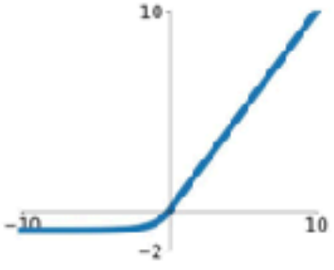


Maxout

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

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



- Use ReLU. Be careful with your learning rates
- Try out Leaky ReLU / Maxout / ELU / SELU
  - To squeeze out some marginal gains
- Don't use sigmoid or tanh

The Basic Pipeline

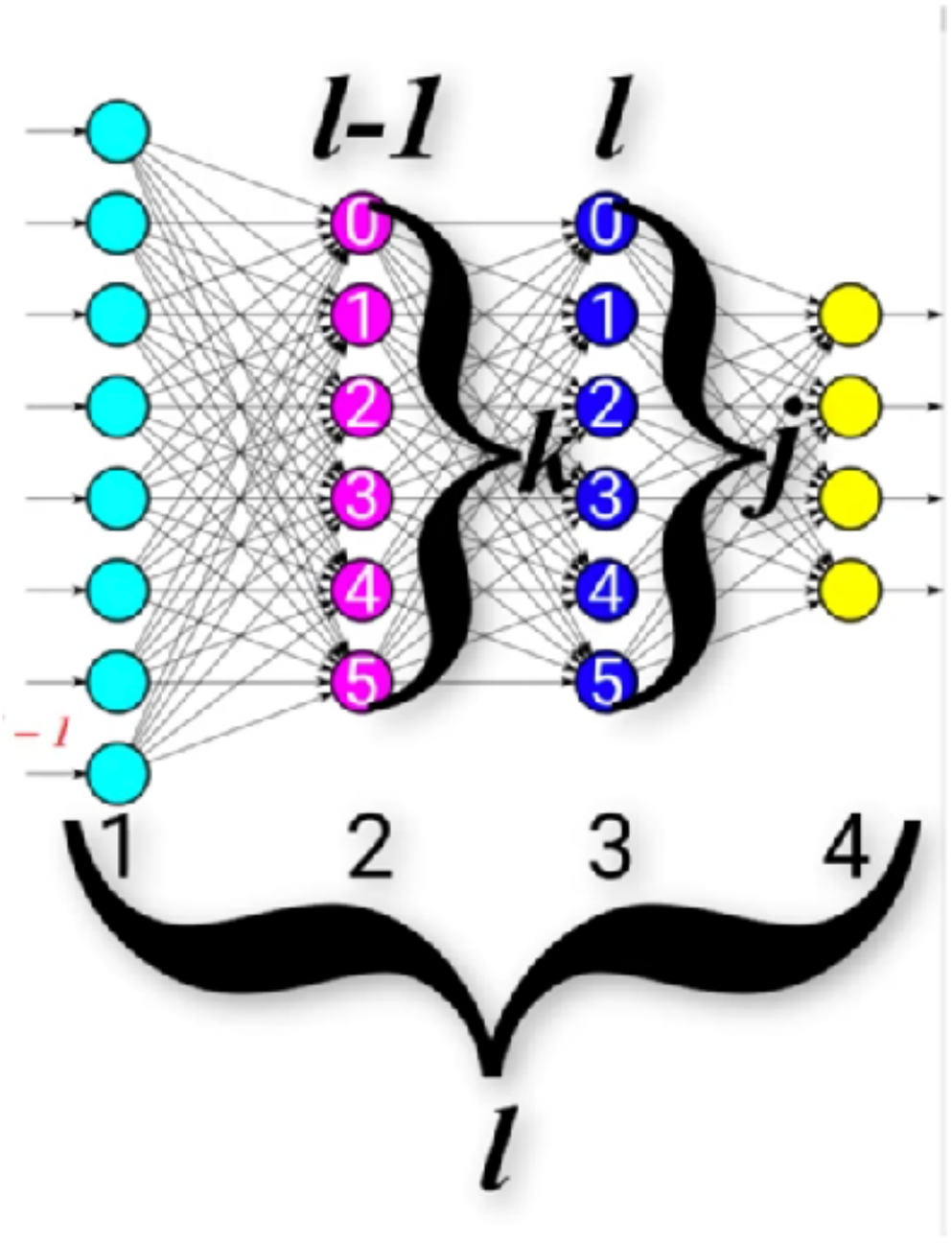
- We sample out a batch of the data.
- We forward propagate it.
- We compute the loss at the end, by comparing prediction and ground truth.
- We back-propagate it and calculate the gradients on the way back.
- We update the parameters/weights using these gradients, by taking steps in the direction of the negative gradient.

Backpropagation

- L = number of layers in the network, from 1->L

- Nodes in a given layer l are indexed as j=0,1,2….n-1

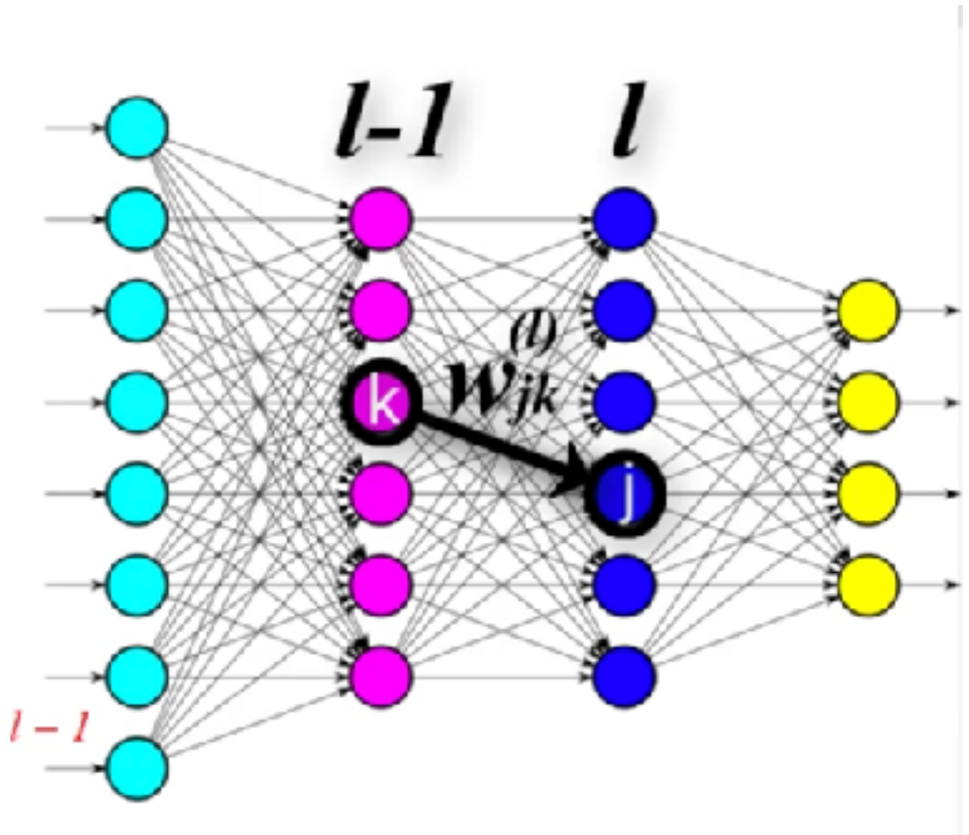
- Nodes in the previous layer are k=0,1,2…n-1



Robotics Research Center

- y_j = the desired value of node j in the output layer L for a single training example. We know the output we desire (since we have GT)

- C_0 = loss function of the network for a single training sample.



$w_{jk}^{(l)}$ = the weight of the connection that connects node k in layer $l-1$ to node j in layer l

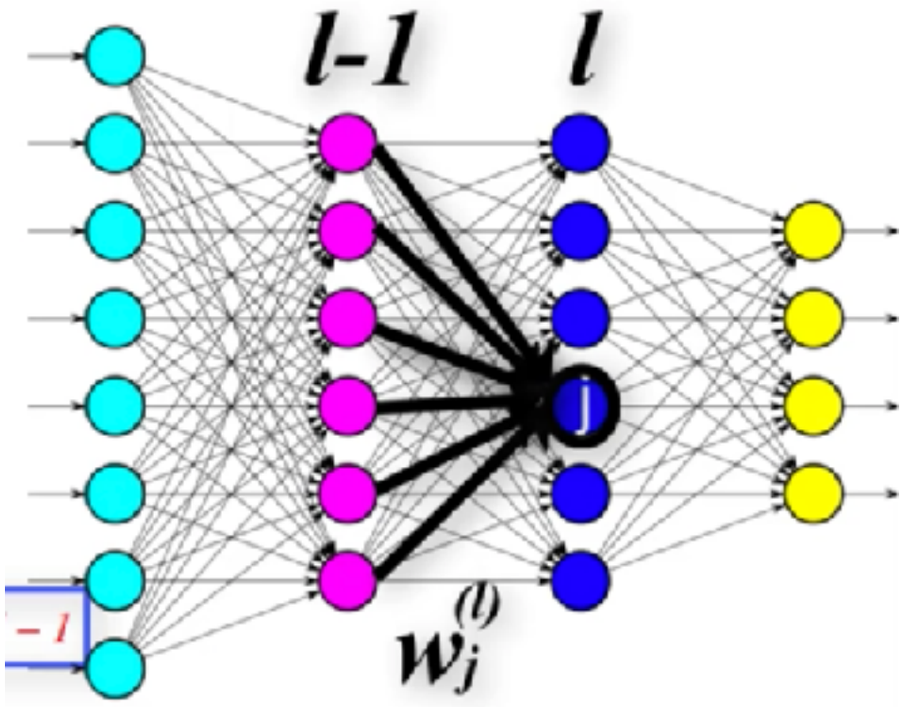
$\vec{w}_j^{(l)}$ = the vector that contains all weights connected to node j in layer l by each node in layer $l-1$

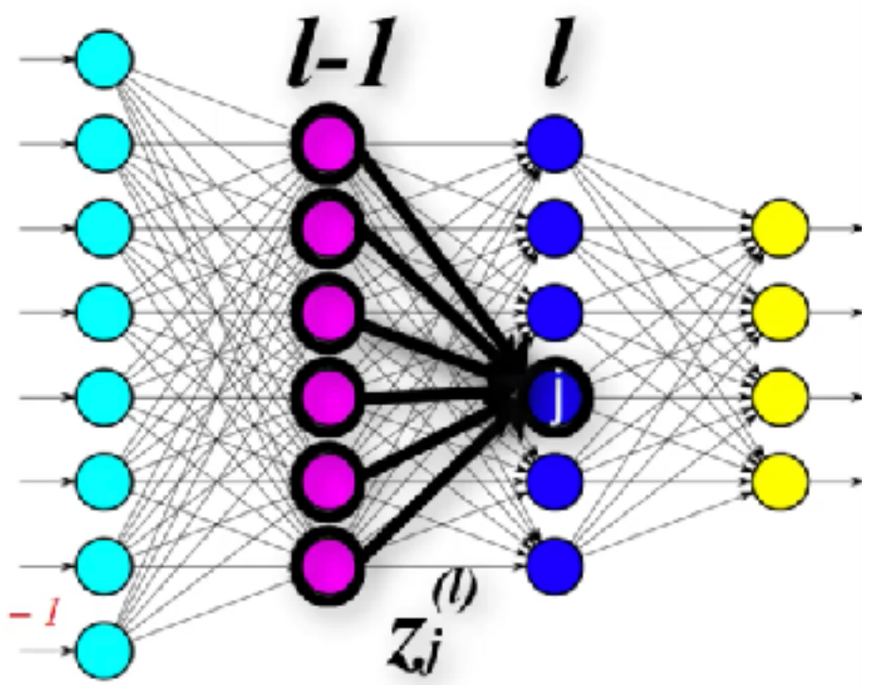
$z_j^{(l)}$ = the input for node j in layer l

$g^{(l)}$ = the activation function used for layer l

$a_j^{(l)}$ = the activation output of node j in layer l

Robotics Research Center





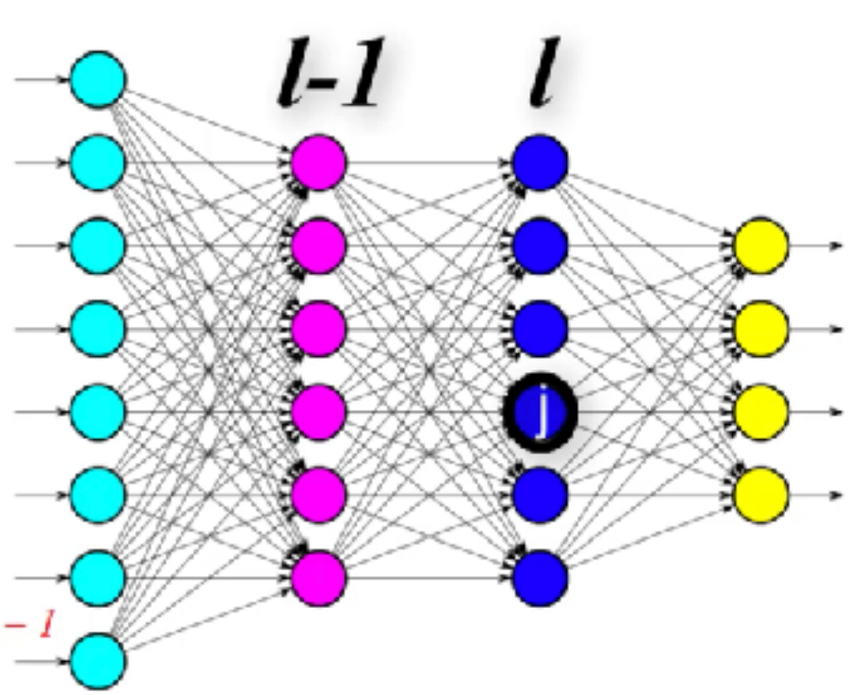
$w_{jk}^{(l)}$ = the weight of the connection that connects node k in layer $l-1$ to node j in layer l

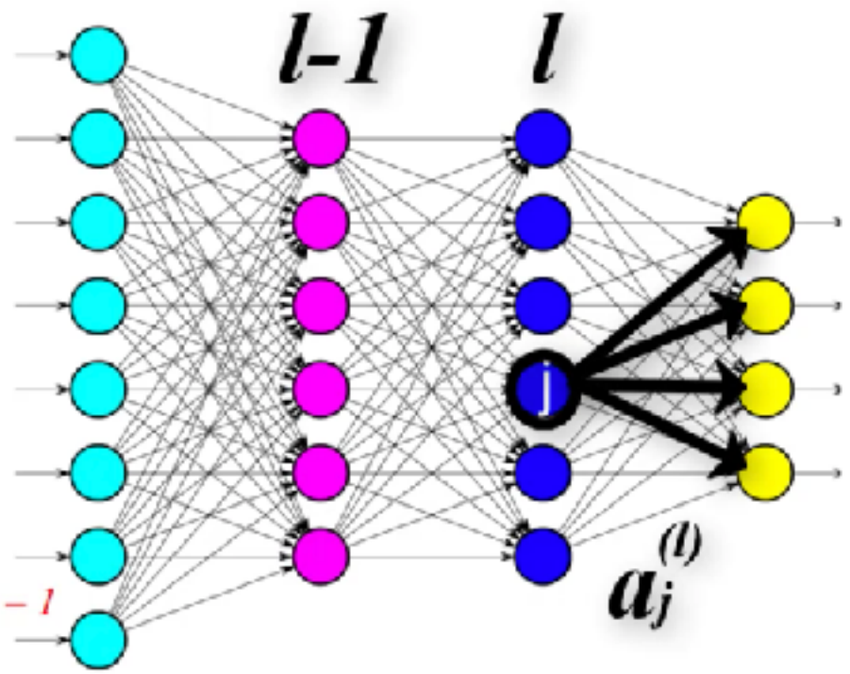
$\vec{w}_j^{(l)}$ = the vector that contains all weights connected to node j in layer l by each node in layer $l-1$

$z_j^{(l)}$ = the input for node j in layer l

$g^{(l)}$ = the activation function used for layer l

$a_j^{(l)}$ = the activation output of node j in layer l





$w_{jk}^{(l)}$ = the weight of the connection that connects node k in layer $l-1$ to node j in layer l

$\vec{w}_j^{(l)}$ = the vector that contains all weights connected to node j in layer l by each node in layer $l-1$

$z_j^{(l)}$ = the input for node j in layer l

$g^{(l)}$ = the activation function used for layer l

$a_j^{(l)}$ = the activation output of node j in layer l

The input for node j is a function of all the weights connected to node j.

So, we can express $z_j^{(L)}$ as a function of $w_j^{(L)}$ as

$$z_j^{(L)}\left(w_j^{(L)}\right).$$

Therefore,

$$C_{0_j} = C_{0_j}\left(a_j^{(L)}\left(z_j^{(L)}\left(w_j^{(L)}\right)\right)\right).$$

So, we can see that C_0 is a composition of functions.

We know that

$$C_0 = \sum_{j=0}^{n-1} C_{0_j},$$

This is useful in order to understand how to differentiate C_0.

To differentiate a composition of functions, we use the chain rule.

Loss function

Loss C_0

Observe that the expression

$$\left(a_j^{(L)} - y_j\right)^2$$

is the squared difference of the activation output and the desired output for node j in the output layer L.

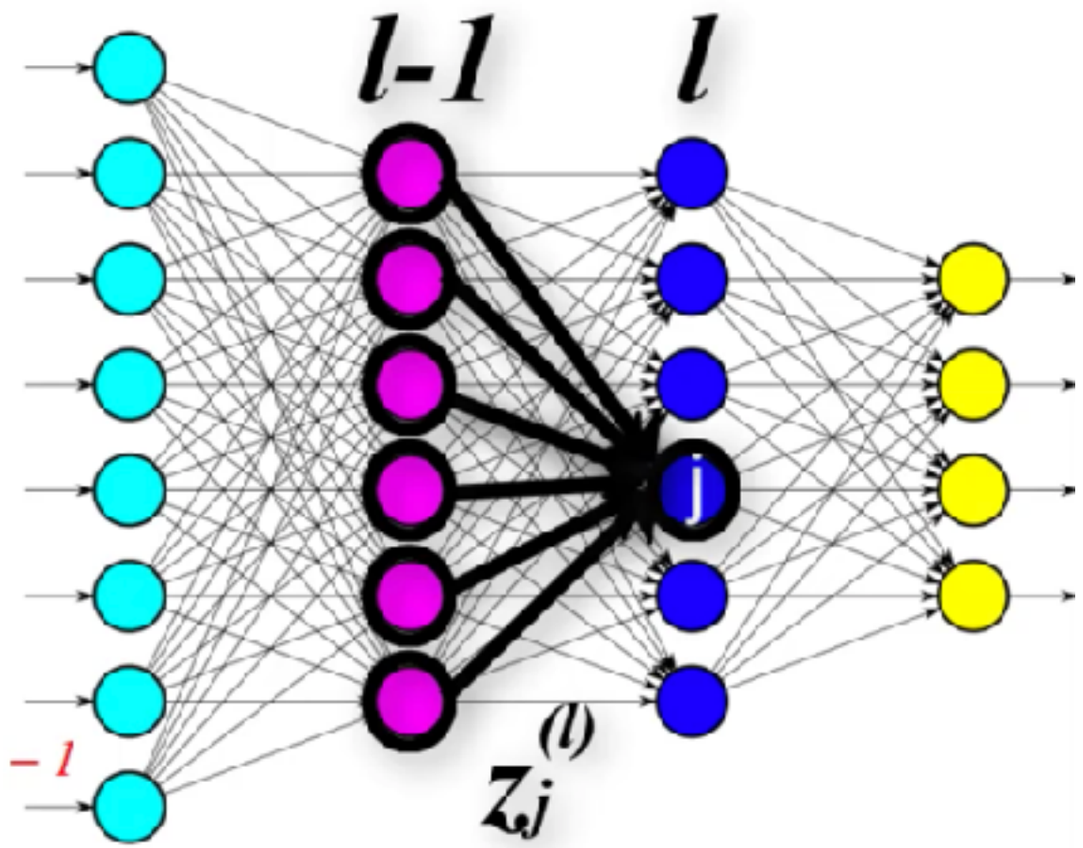
This can be interpreted as the loss for node j in layer L.

Therefore, to calculate the total loss, we should sum this squared difference for each node j in the output layer L.

This is expressed as

$$C_0 = \sum_{j=0}^{n-1} \left(a_j^{(L)} - y_j\right)^2.$$

Robotics Research Center



Input $z_j^{(l)}$

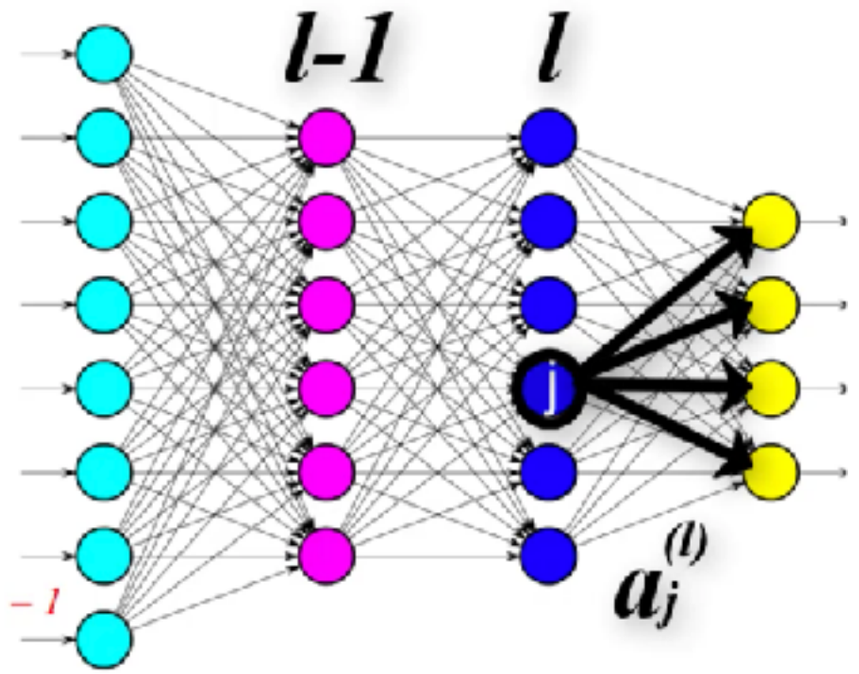
We know that the input for node j in layer l is the weighted sum of the activation outputs from the previous layer $l-1$.

An individual term from the sum looks like this:

$$w_{jk}^{(l)} a_k^{(l-1)}$$

So, the input for a given node j in layer l is expressed as

$$z_j^{(l)} = \sum_{k=0}^{n-1} w_{jk}^{(l)} a_k^{(l-1)}.$$



Activation Output $a_j^{(l)}$

We know that the activation output of a given node j in layer l is the result of passing the input, $z_j^{(l)}$, to whatever activation function we choose to use $g^{(l)}$.

Therefore, the activation output of node j in layer l is expressed as

$$a_j^{(l)} = g^{(l)}\left(z_j^{(l)}\right).$$

Recall the definition of C_0,

$$C_0 = \sum_{j=0}^{n-1} \left(a_j^{(L)} - y_j\right)^2.$$

So the loss of a single node j in the output layer L can be expressed as

$$C_{0_j} = \left(a_j^{(L)} - y_j\right)^2.$$

We see that C_{0_j} is a function of the activation output of node j in layer L.

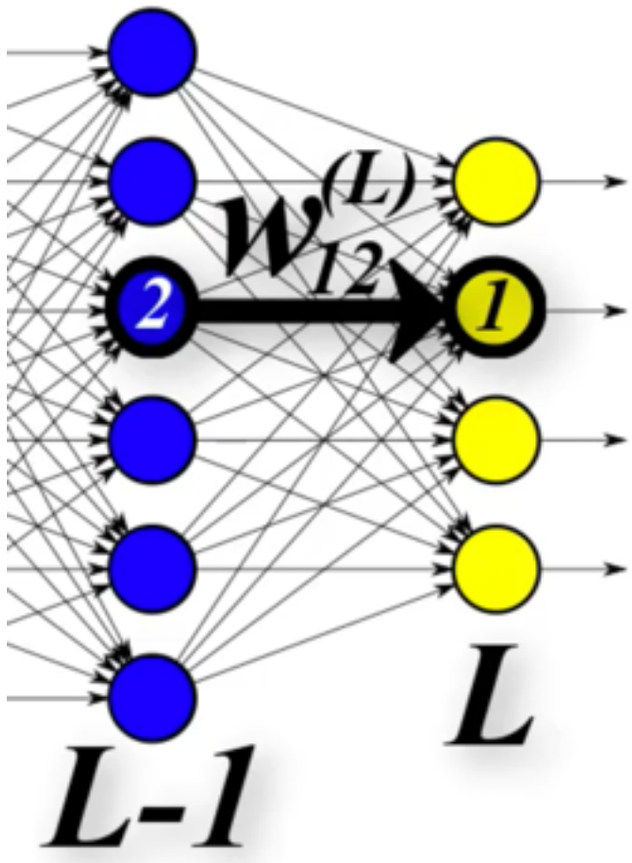
So, we can express C_{0_j} as a function of $a_j^{(L)}$ as

$$C_{0_j}\left(a_j^{(L)}\right).$$

The activation output of node j in the output layer L is a function of the input for node j.

From an earlier observation, we know we can express this as

$$a_j^{(L)} = g^{(L)}\left(z_j^{(L)}\right).$$

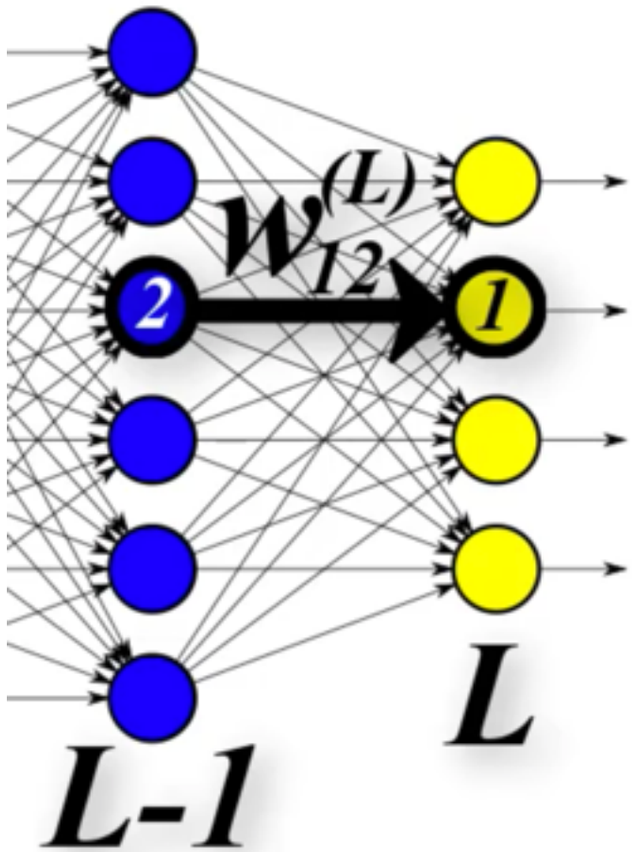


Let's look at a single weight that connects node 2 in layer $L-1$ to node 1 in layer L. This weight is denoted as

$$w_{12}^{(L)}.$$

The derivative of the loss C_0 with respect to this particular weight $w_{12}^{(L)}$ is denoted as

$$\frac{\partial C_0}{\partial w_{12}^{(L)}}.$$



Since C_0 depends on $a_1^{(L)}$, and $a_1^{(L)}$ depends on $z_1^{(L)}$, and $z_1^{(L)}$ depends on $w_{12}^{(L)}$:

Term 1 Term 2 Term 3

$$\frac{\partial C_0}{\partial w_{12}^{(L)}} = \left(\frac{\partial C_0}{\partial a_1^{(L)}}\right)\left(\frac{\partial a_1^{(L)}}{\partial z_1^{(L)}}\right)\left(\frac{\partial z_1^{(L)}}{\partial w_{12}^{(L)}}\right).$$

Activation output of node 1 Input of node 1 The weight

Robotics Research Center

$$C_0 = \sum_{j=0}^{n-1} \left(a_j^{(L)} - y_j \right)^2 .$$

$$\frac{\partial C_0}{\partial a_1^{(L)}} = \frac{\partial}{\partial a_1^{(L)}} \left(\sum_{j=0}^{n-1} \left(a_j^{(L)} - y_j \right)^2 \right).$$

$$\begin{aligned}
\frac{\partial}{\partial a_1^{(L)}} \left(\sum_{j=0}^{n-1} \left(a_j^{(L)} - y_j \right)^2 \right) &= \frac{\partial}{\partial a_1^{(L)}} \left(\left(a_0^{(L)} - y_0 \right)^2 + \left(a_1^{(L)} - y_1 \right)^2 + \left(a_2^{(L)} - y_2 \right)^2 + \left(a_3^{(L)} - y_3 \right)^2 \right) \\
&= \frac{\partial}{\partial a_1^{(L)}} \left(\left(a_0^{(L)} - y_0 \right)^2 \right) + \frac{\partial}{\partial a_1^{(L)}} \left(\left(a_1^{(L)} - y_1 \right)^2 \right) + \frac{\partial}{\partial a_1^{(L)}} \left(\left(a_2^{(L)} - y_2 \right)^2 \right) + \frac{\partial}{\partial a_1^{(L)}} \left(\left(a_3^{(L)} - y_3 \right)^2 \right) \\
&= 2 \left(a_1^{(L)} - y_1 \right).
\end{aligned}$$

The loss from the network for a single input image will respond to a small change in the activation output from the node in layer L by an amount equal to two times the difference of the activation output of that node with the ground truth (desired output).

Robotics Research Center

The Second Term

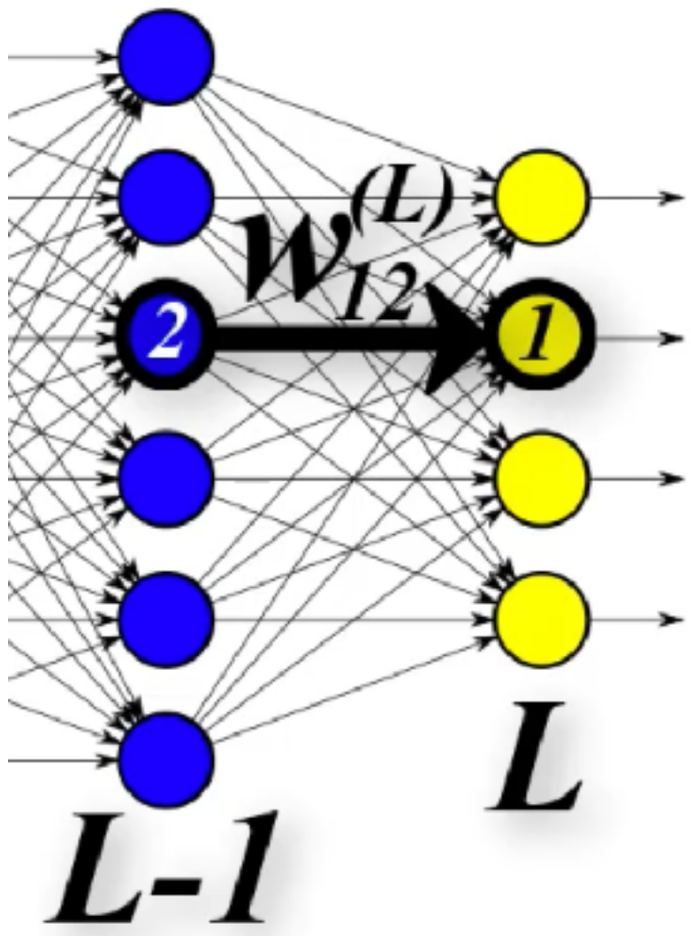
$$a_j^{(L)} = g^{(L)}\left(z_j^{(L)}\right),$$

$$a_1^{(L)} = g^{(L)}\left(z_1^{(L)}\right).$$

Activation output of node 1

$$\frac{\partial a_1^{(L)}}{\partial z_1^{(L)}} = \frac{\partial}{\partial z_1^{(L)}}\left(g^{(L)}\left(z_1^{(L)}\right)\right)$$

$$= g'^{(L)}\left(z_1^{(L)}\right).$$



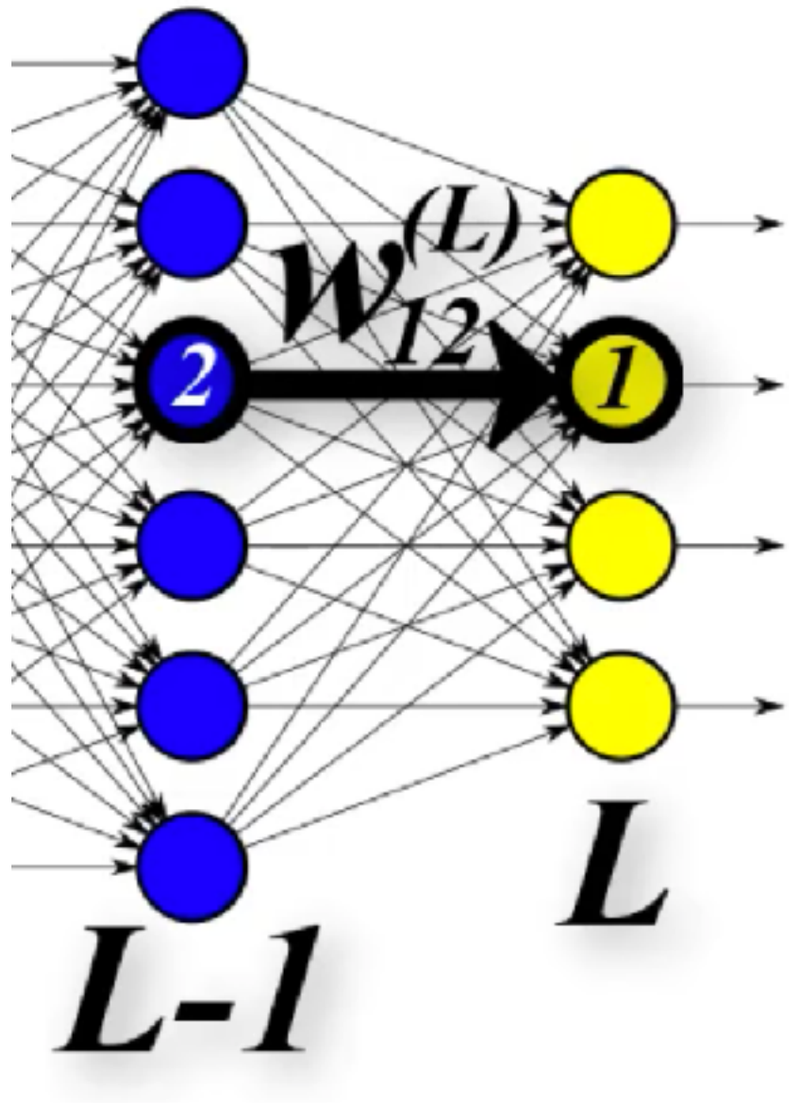
Robotics Research Center

The Third Term

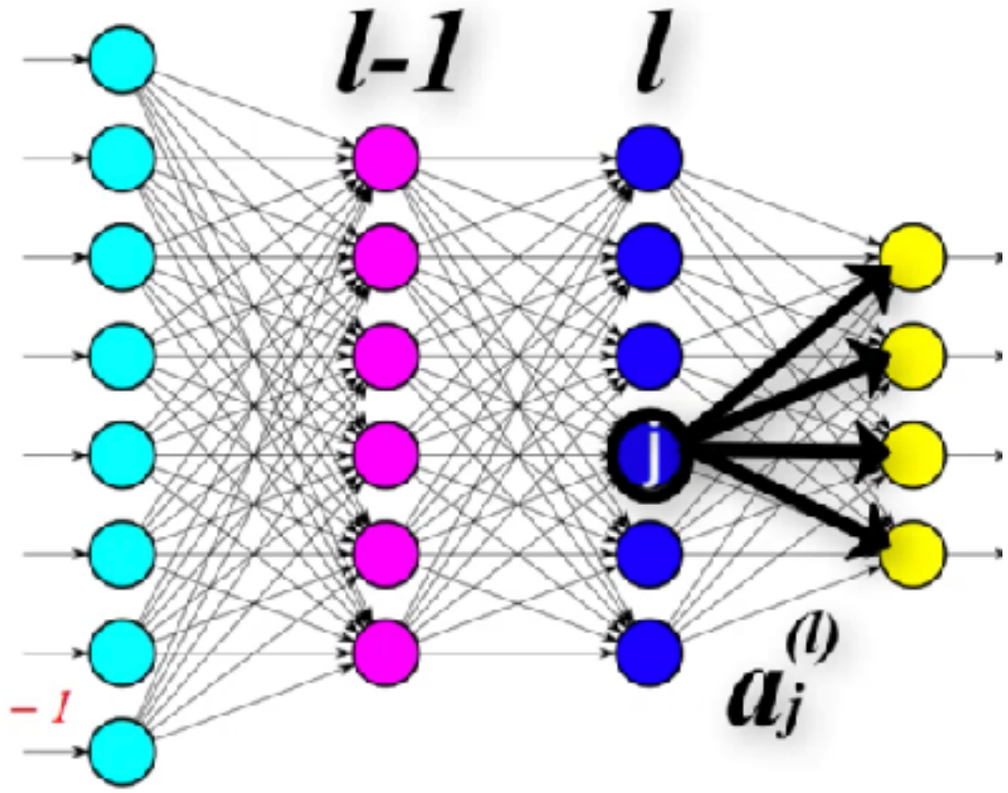
$$z_j^{(L)} = \sum_{k=0}^{n-1} w_{jk}^{(L)} a_k^{(L-1)}.$$

$$z_1^{(L)} = \sum_{k=0}^{n-1} w_{1k}^{(L)} a_k^{(L-1)}.$$

$$\frac{\partial z_1^{(L)}}{\partial w_{12}^{(L)}} = \frac{\partial}{\partial w_{12}^{(L)}} \left(\sum_{k=0}^{n-1} w_{1k}^{(L)} a_k^{(L-1)} \right).$$



Robotics Research Center



$$\frac{\partial}{\partial w_{12}^{(L)}}\left(\sum_{k=0}^{n-1} w_{1k}^{(L)} a_k^{(L-1)}\right) = \frac{\partial}{\partial w_{12}^{(L)}}\left(w_{10}^{(L)} a_0^{(L-1)} + w_{11}^{(L)} a_1^{(L-1)} + w_{12}^{(L)} a_2^{(L-1)} + \cdots + w_{15}^{(L)} a_5^{(L-1)}\right)$$

$$= \frac{\partial}{\partial w_{12}^{(L)}} w_{10}^{(L)} a_0^{(L-1)} + \frac{\partial}{\partial w_{12}^{(L)}} w_{11}^{(L)} a_1^{(L-1)} + \boxed{\frac{\partial}{\partial w_{12}^{(L)}} w_{12}^{(L)} a_2^{(L-1)}} + \cdots + \frac{\partial}{\partial w_{12}^{(L)}} w_{15}^{(L)} a_5^{(L-1)}$$

$$= a_2^{(L-1)}$$

The input for the node in layer L will respond to a change in the weight w_{12} by an amount equal to the activation output of node 2 in the previous layer.

Robotics Research Center

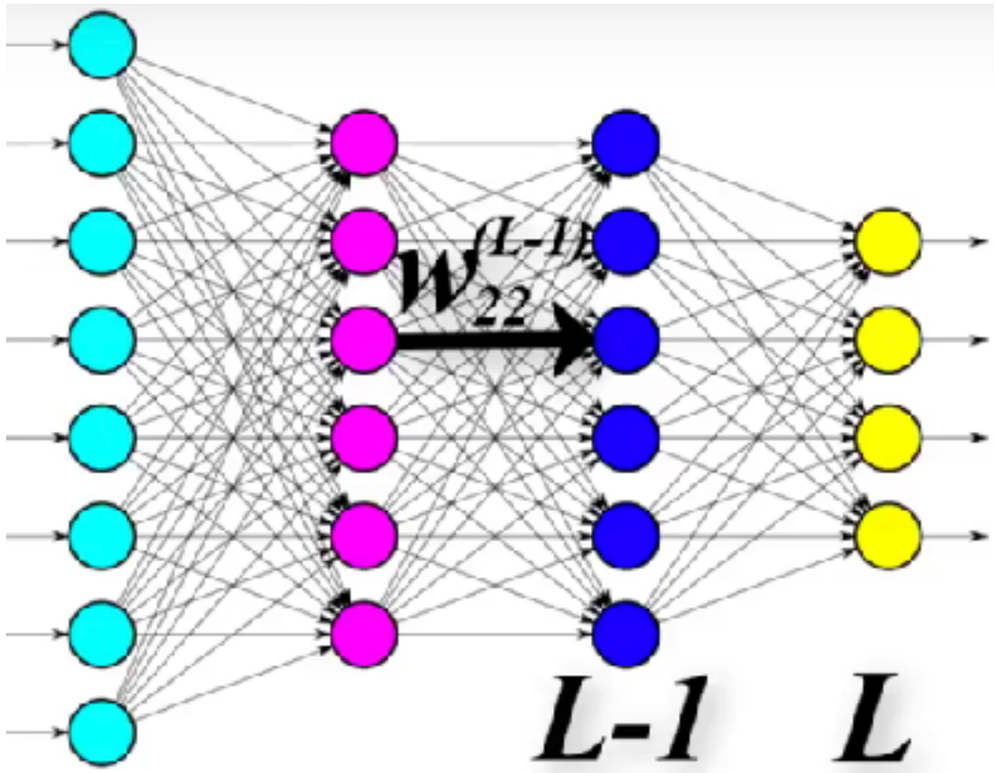
$$\frac{\partial C_0}{\partial w_{12}^{(L)}} = \left(\frac{\partial C_0}{\partial a_1^{(L)}}\right)\left(\frac{\partial a_1^{(L)}}{\partial z_1^{(L)}}\right)\left(\frac{\partial z_1^{(L)}}{\partial w_{12}^{(L)}}\right)$$
$$= 2\left(a_1^{(L)} - y_1\right)\left(g'^{(L)}\left(z_1^{(L)}\right)\right)\left(a_2^{(L-1)}\right)$$

Average derivative of the loss function over n training samples

$$\frac{\partial C}{\partial w_{12}^{(L)}} = \frac{1}{n}\sum_{i=0}^{n-1}\frac{\partial C_i}{\partial w_{12}^{(L)}}.$$

We do this process for each weight in the network to compute the derivative of the loss function with respect to each weight!

Now suppose we wanna compute the gradient with respect to earlier weights in the network?



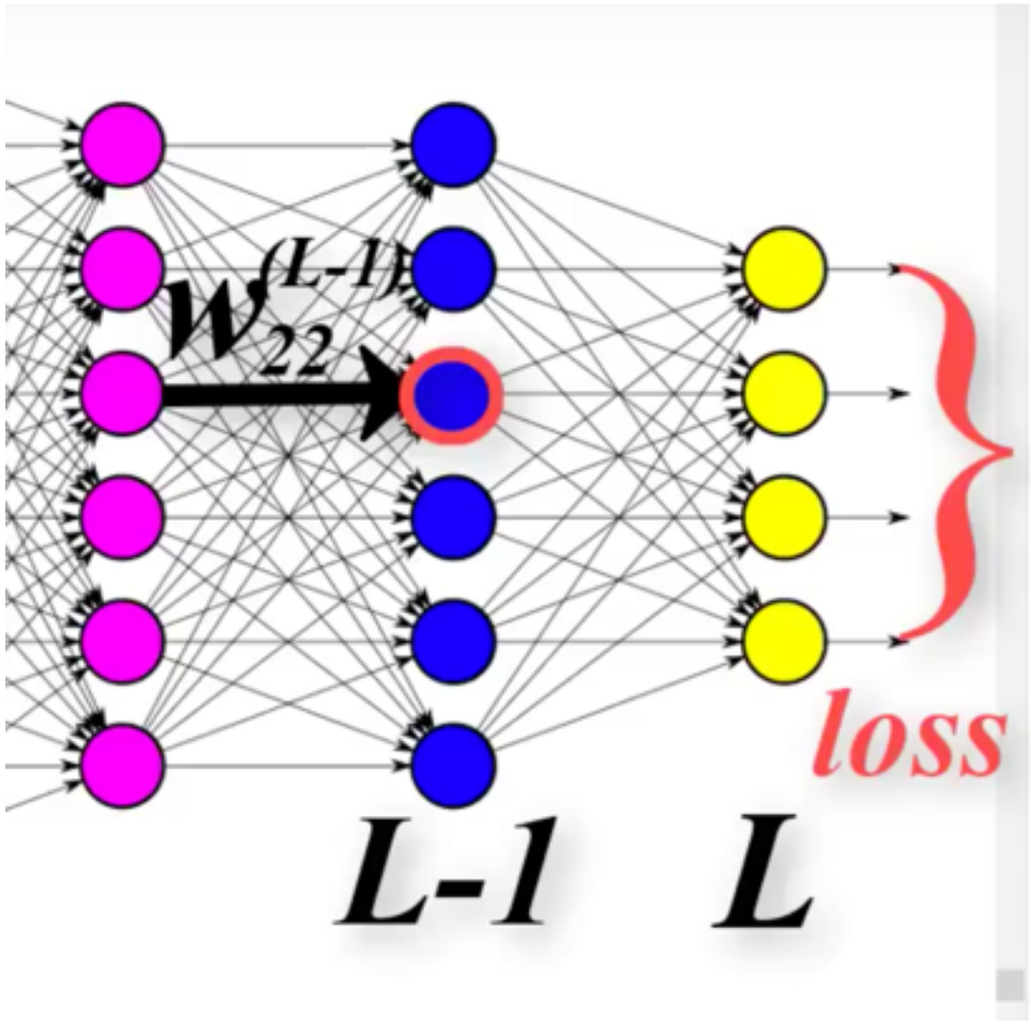
The equation previously

$$\frac{\partial C_0}{\partial w_{12}^{(L)}} = \left(\frac{\partial C_0}{\partial a_1^{(L)}}\right)\left(\frac{\partial a_1^{(L)}}{\partial z_1^{(L)}}\right)\left(\frac{\partial z_1^{(L)}}{\partial w_{12}^{(L)}}\right)$$

$$\frac{\partial C_0}{\partial w_{22}^{(L-1)}} = \left(\frac{\partial C_0}{\partial a_2^{(L-1)}}\right)\left(\frac{\partial a_2^{(L-1)}}{\partial z_2^{(L-1)}}\right)\left(\frac{\partial z_2^{(L-1)}}{\partial w_{22}^{(L-1)}}\right)$$

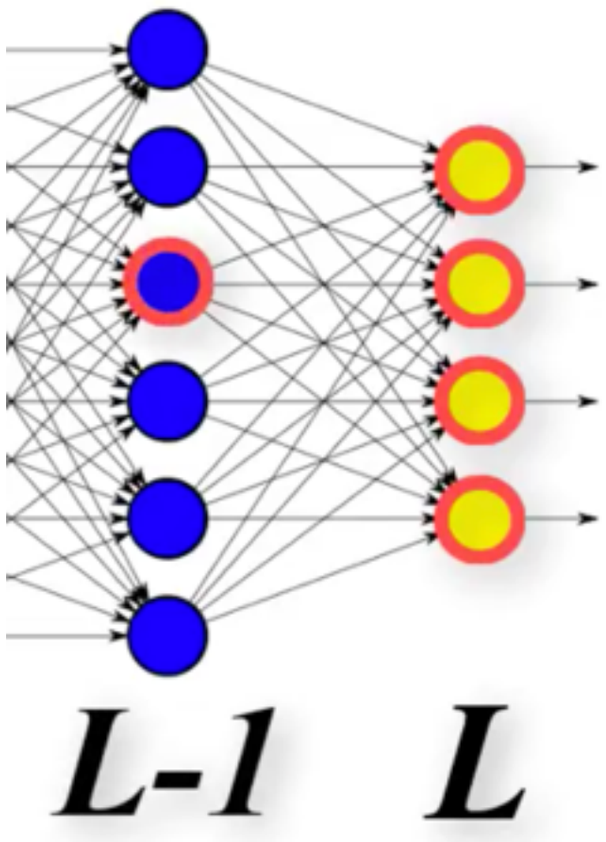
Derivative of loss wrt activation output of node 2 in layer L-1

Robotics Research Center



$$\frac{\partial C_0}{\partial w_{22}^{(L-1)}} = \left(\frac{\partial C_0}{\partial a_2^{(L-1)}}\right)\left(\frac{\partial a_2^{(L-1)}}{\partial z_2^{(L-1)}}\right)\left(\frac{\partial z_2^{(L-1)}}{\partial w_{22}^{(L-1)}}\right)$$

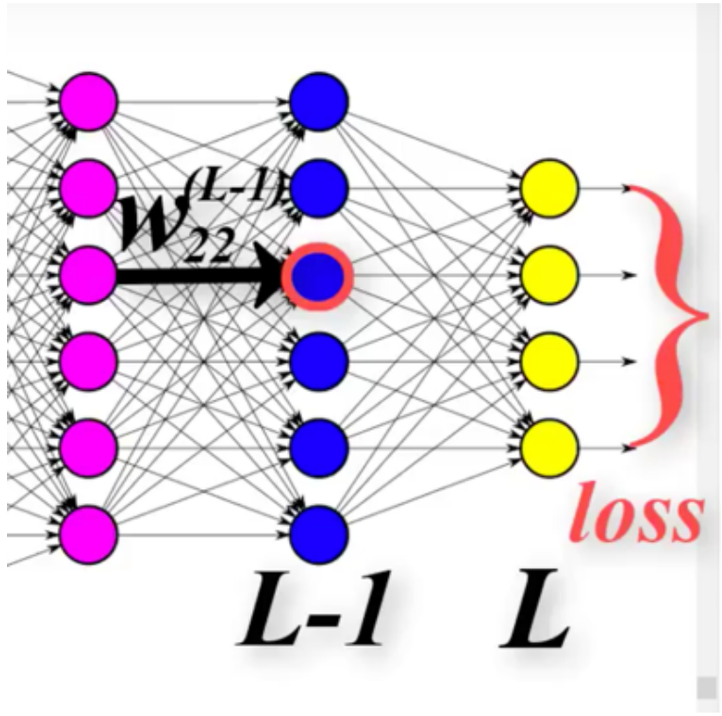
$$\frac{\partial C_0}{\partial a_2^{(L-1)}} = \sum_{j=0}^{n-1} \left(\left(\frac{\partial C_0}{\partial a_j^{(L)}} \right) \left(\frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \right) \left(\frac{\partial z_j^{(L)}}{\partial a_2^{(L-1)}} \right) \right).$$



$$\frac{\partial}{\partial a_2^{(L-1)}} \sum_{k=0}^{n-1} w_{jk}^{(L)} a_k^{(L-1)} = \frac{\partial}{\partial a_2^{(L-1)}} \left(w_{j0}^{(L)} a_0^{(L-1)} + w_{j1}^{(L)} a_1^{(L-1)} + w_{j2}^{(L)} a_2^{(L-1)} \cdots + w_{j5}^{(L)} a_5^{(L-1)} \right)$$

$$= \frac{\partial}{\partial a_2^{(L-1)}} w_{j0}^{(L)} a_0^{(L-1)} + \frac{\partial}{\partial a_2^{(L-1)}} w_{j1}^{(L)} a_1^{(L-1)} + \boxed{\frac{\partial}{\partial a_2^{(L-1)}} w_{j2}^{(L)} a_2^{(L-1)}} \cdots + \frac{\partial}{\partial a_2^{(L-1)}} w_{j5}^{(L)} a_5^{(L-1)}$$

$$= w_{j2}^{(L)}.$$

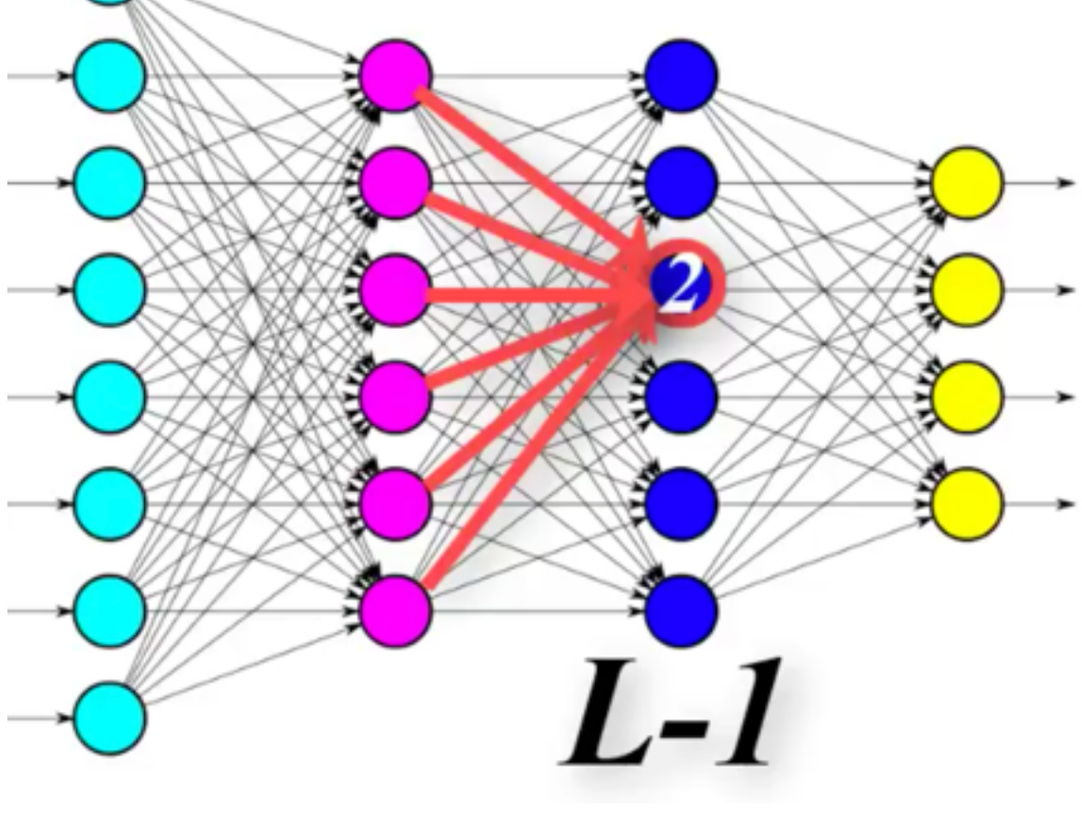


The input for any node in the output layer L will respond to a change in the activation output of the previous layer by an amount equal to the weight connecting node 2 in the previous layer to a node j in the output layer L.

$$\frac{\partial C_0}{\partial a_2^{(L-1)}} = \sum_{j=0}^{n-1}\left(\left(\frac{\partial C_0}{\partial a_j^{(L)}}\right)\left(\frac{\partial a_j^{(L)}}{\partial z_j^{(L)}}\right)\left(\frac{\partial z_j^{(L)}}{\partial a_2^{(L-1)}}\right)\right)$$

$$= \sum_{j=0}^{n-1}\left(2\left(a_j^{(L)} - y_j\right)\left(g'^{(L)}\left(z_j^{(L)}\right)\right)\left(w_{j2}^{(L)}\right)\right).$$

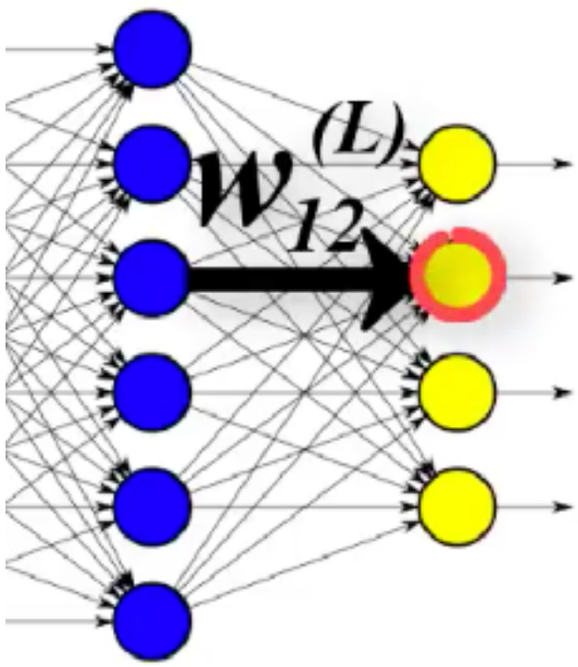
We did all this so we can compete the gradient of the loss with respect to any weight connected to node 2 in layer L-1.



Robotics Research Center

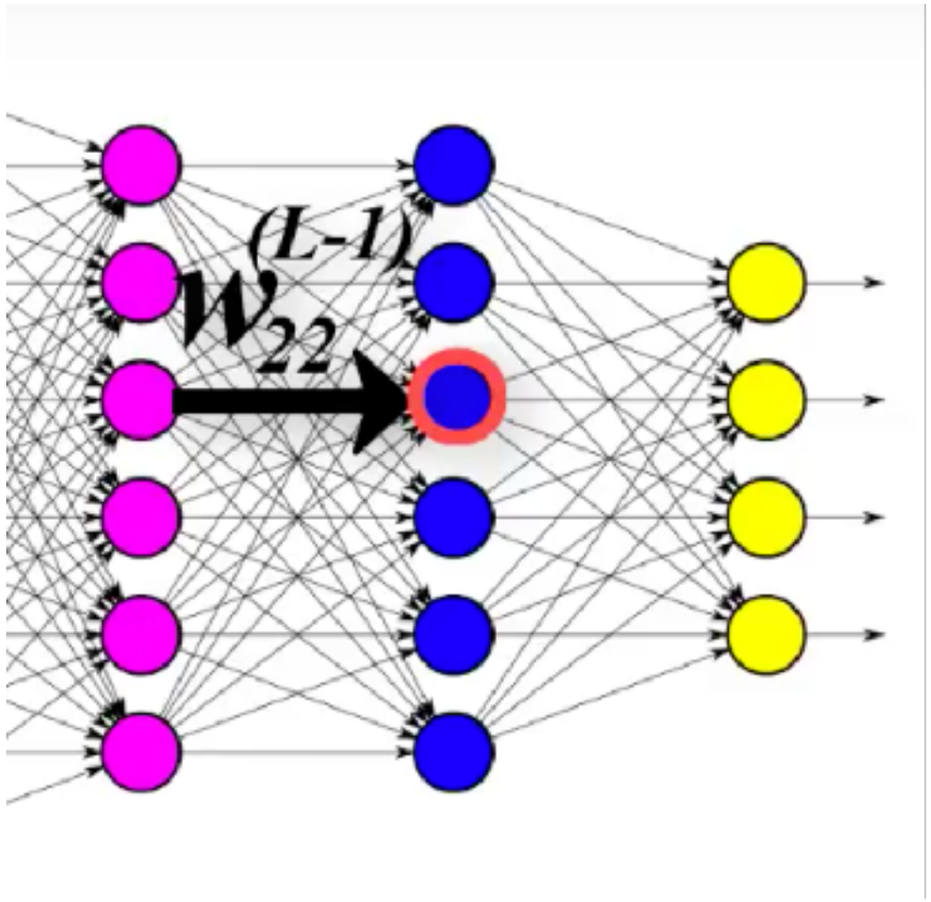
$$\frac{\partial C_0}{\partial w_{22}^{(L-1)}} = \left(\frac{\partial C_0}{\partial a_2^{(L-1)}} \right) \left(\frac{\partial a_2^{(L-1)}}{\partial z_2^{(L-1)}} \right) \left(\frac{\partial z_2^{(L-1)}}{\partial w_{22}^{(L-1)}} \right)$$

We used the chain rule twice! Once for the entire equation, and one for the first term - the derivative of the loss wrt to the activation output of previous layer.



To compute the gradient of the loss with respect to w_{12} , we needed the derivatives that depended on the activation output and the input for the node in **the output layer.**

Robotics Research Center



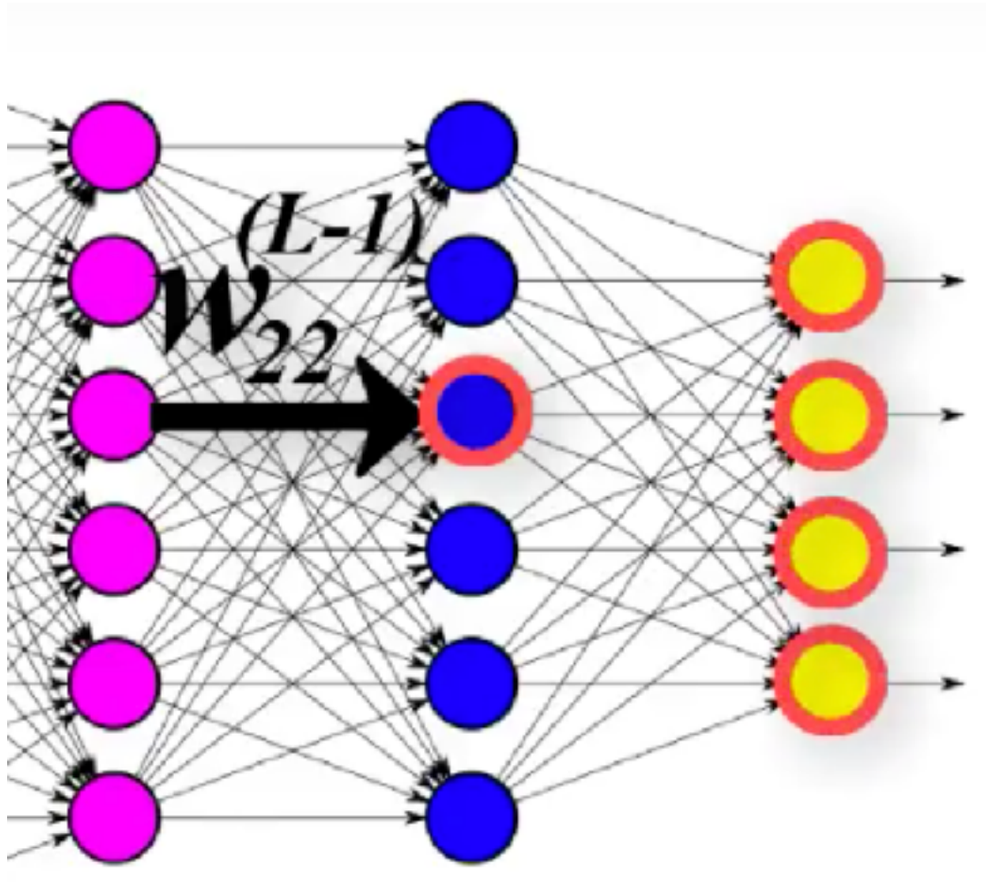
To compute the gradient of the loss with respect to w_{22} , we needed the derivatives that depended on the activation output and the input for the particular node in the **second last layer!**

$$\frac{\partial C_0}{\partial w_{22}^{(L-1)}} = \left(\frac{\partial C_0}{\partial a_2^{(L-1)}}\right)\left(\frac{\partial a_2^{(L-1)}}{\partial z_2^{(L-1)}}\right)\left(\frac{\partial z_2^{(L-1)}}{\partial w_{22}^{(L-1)}}\right)$$

$$\frac{\partial C_0}{\partial a_2^{(L-1)}} = \sum_{j=0}^{n-1}\left(\left(\frac{\partial C_0}{\partial a_j^{(L)}}\right)\left(\frac{\partial a_j^{(L)}}{\partial z_j^{(L)}}\right)\left(\frac{\partial z_j^{(L)}}{\partial a_2^{(L-1)}}\right)\right)$$

$$= \sum_{j=0}^{n-1}\left(2\left(a_j^{(L)} - y_j\right)\left(g'^{(L)}\left(z_j^{(L)}\right)\right)\left(w_{j2}^{(L)}\right)\right).$$

Robotics Research Center



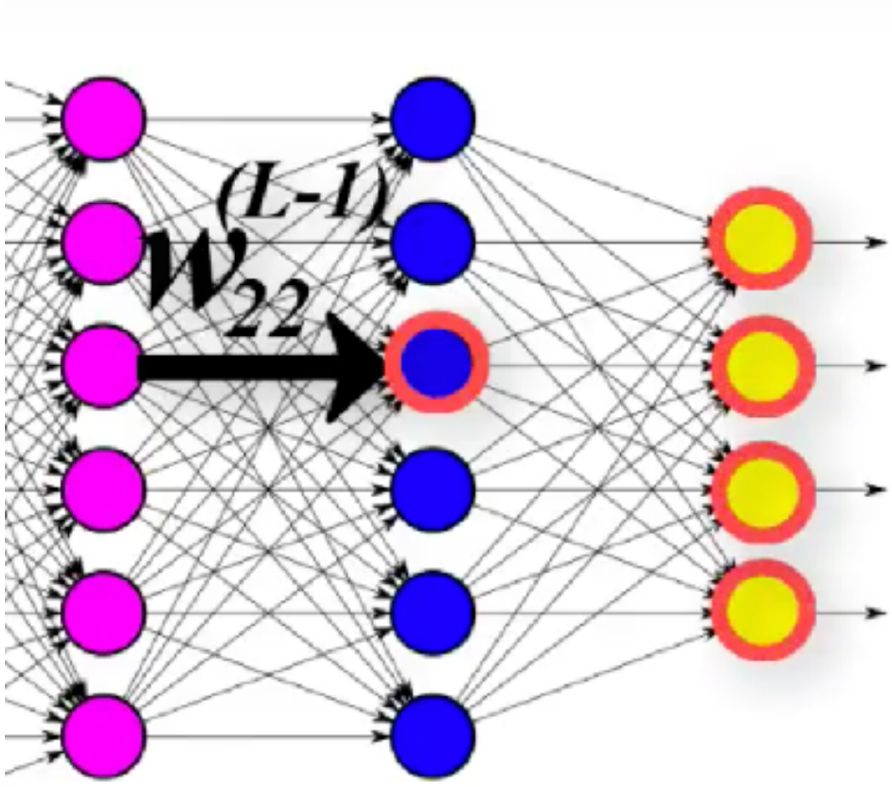
And the derivative that depended on this activation output (the blue box) needs the derivatives that depend on ALL the activation outputs and all the inputs in the final layer!

$$\frac{\partial C_0}{\partial w_{22}^{(L-1)}} = \left(\frac{\partial C_0}{\partial a_2^{(L-1)}}\right)\left(\frac{\partial a_2^{(L-1)}}{\partial z_2^{(L-1)}}\right)\left(\frac{\partial z_2^{(L-1)}}{\partial w_{22}^{(L-1)}}\right)$$

$$\frac{\partial C_0}{\partial a_2^{(L-1)}} = \sum_{j=0}^{n-1}\left(\left(\frac{\partial C_0}{\partial a_j^{(L)}}\right)\left(\frac{\partial a_j^{(L)}}{\partial z_j^{(L)}}\right)\left(\frac{\partial z_j^{(L)}}{\partial a_2^{(L-1)}}\right)\right)$$

$$= \sum_{j=0}^{n-1}\left(2\left(a_j^{(L)} - y_j\right)\left(g'^{(L)}\left(z_j^{(L)}\right)\right)\left(w_{j2}^{(L)}\right)\right).$$

Robotics Research Center



Apply chain rule repeatedly in the backwards fashion!

Hence, we need to compute derivatives that depend components later in the network FIRST, and then use these for computations of gradients with respect to weights that come EARLIER in the network!

$$\frac{\partial C_0}{\partial w_{22}^{(L-1)}} = \left(\frac{\partial C_0}{\partial a_2^{(L-1)}}\right)\left(\frac{\partial a_2^{(L-1)}}{\partial z_2^{(L-1)}}\right)\left(\frac{\partial z_2^{(L-1)}}{\partial w_{22}^{(L-1)}}\right)$$

$$\frac{\partial C_0}{\partial a_2^{(L-1)}} = \sum_{j=0}^{n-1}\left(\left(\frac{\partial C_0}{\partial a_j^{(L)}}\right)\left(\frac{\partial a_j^{(L)}}{\partial z_j^{(L)}}\right)\left(\frac{\partial z_j^{(L)}}{\partial a_2^{(L-1)}}\right)\right)$$

$$= \sum_{j=0}^{n-1}\left(2\left(a_j^{(L)} - y_j\right)\left(g'^{(L)}\left(z_j^{(L)}\right)\right)\left(w_{j2}^{(L)}\right)\right).$$

Robotics Research Center

Vanishing gradients

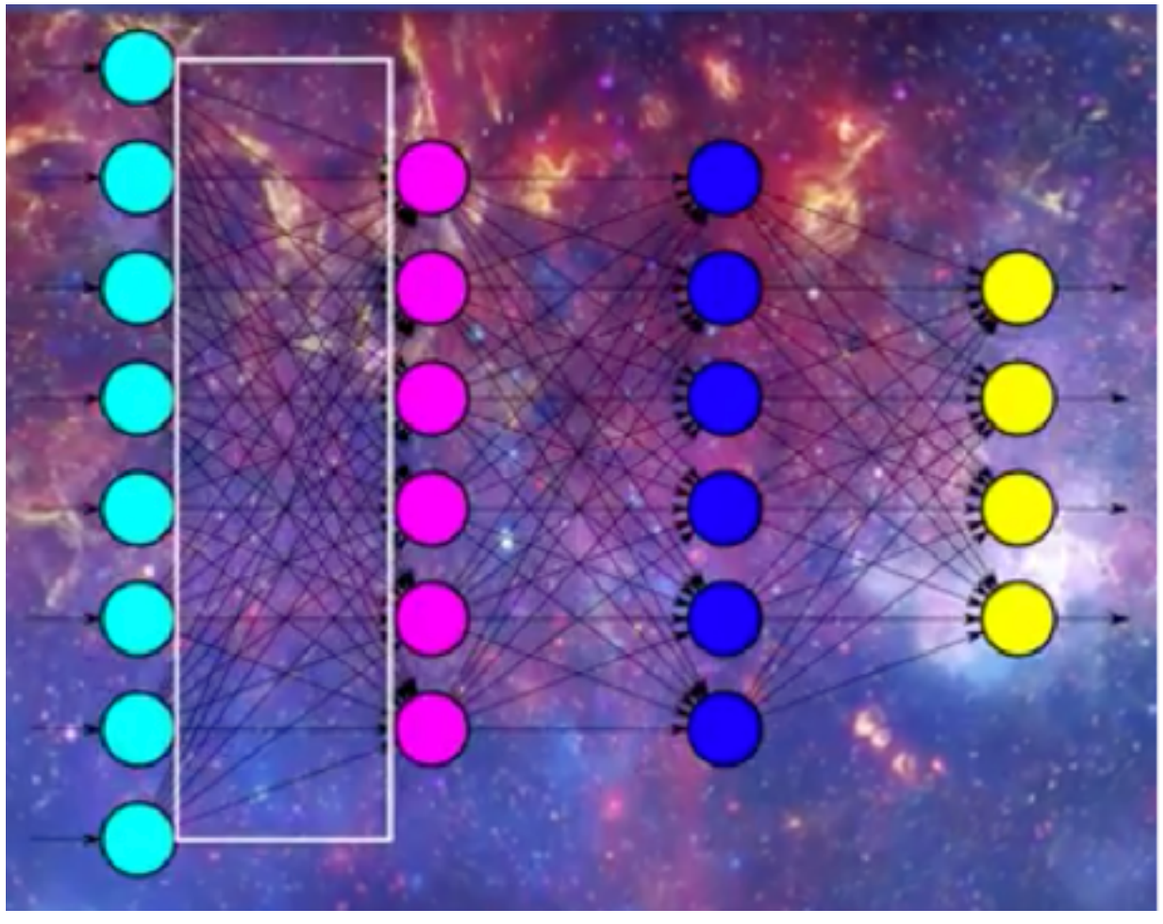
- One problem that we frequently come across while training neural networks is the problem of unstable gradients.
- This is known popularly as the **vanishing gradient** problem.
- When we say gradient - we mean the gradient of the loss function with respect to the weights in the neural network.
- We use these gradients to update the weights of our network, during backprop.

What is vanishing gradient?

- During training, SGD works to calculate the gradient of the loss wrt the weights.

- However during training, the gradient of the earlier weights in the network become REALLY SMALL. They **vanish**.

- Earlier weights as in, the weights of the earlier layers of the artificial neural net.



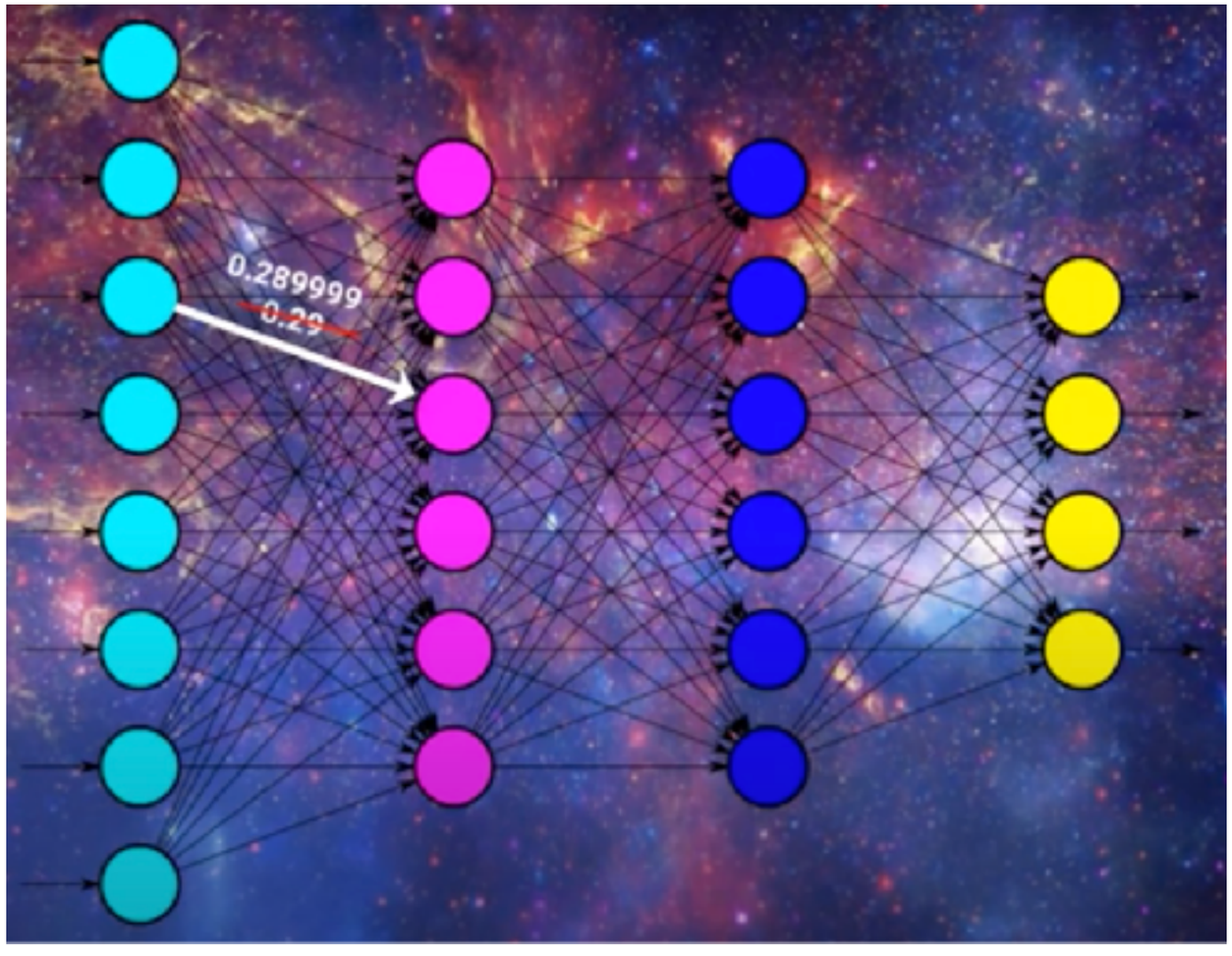
Robotics Research Center

Vanishing Gradient

- If the gradient is vanishingly small, then there'll barely be any update to the weight!

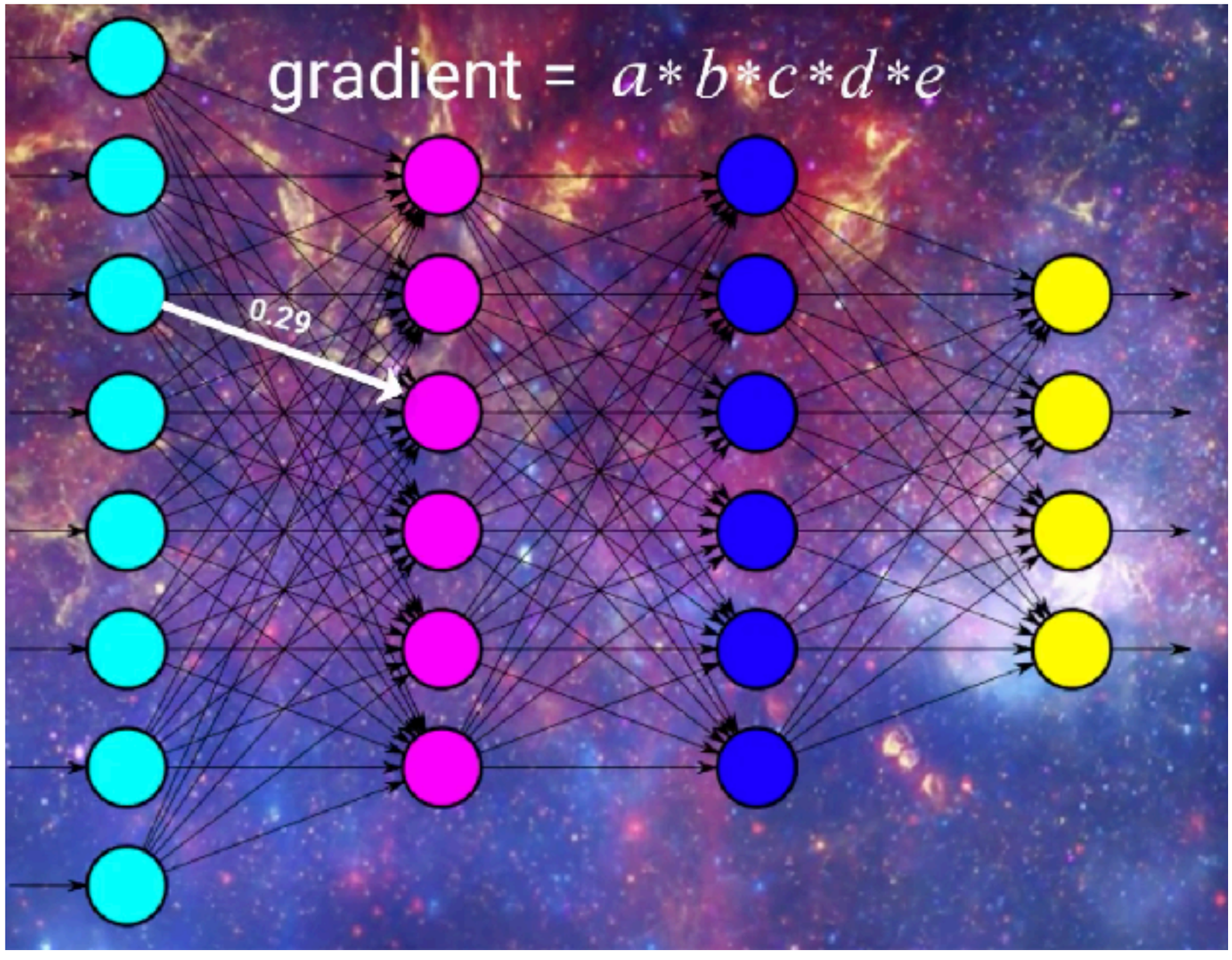
- The weights will barely change! It won't carry through the network very well to reduce the loss, as it hasn't changed much from the previous iteration.

- And we know that the the input of subsequent layers depend on the output of the initial layers. If the initial layers themselves don't change much, it impairs the ability of the neural net to learn.



Robotics Research Center

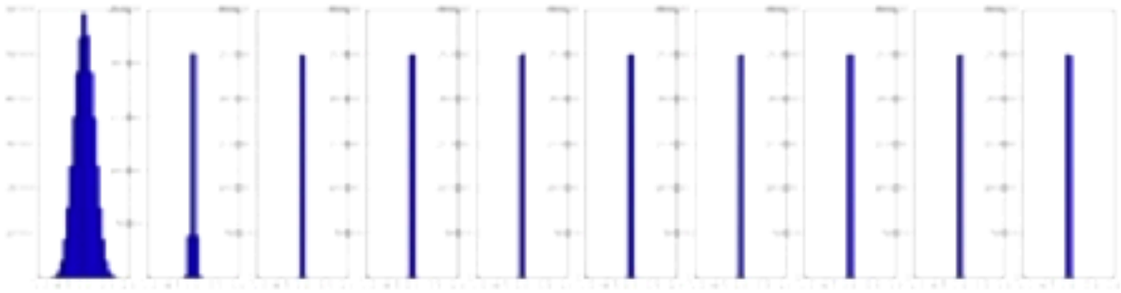
The product for earlier weights in the network are going to be really small and when we subtract this number from the weight, it's barely going to make the difference. It won't learn!



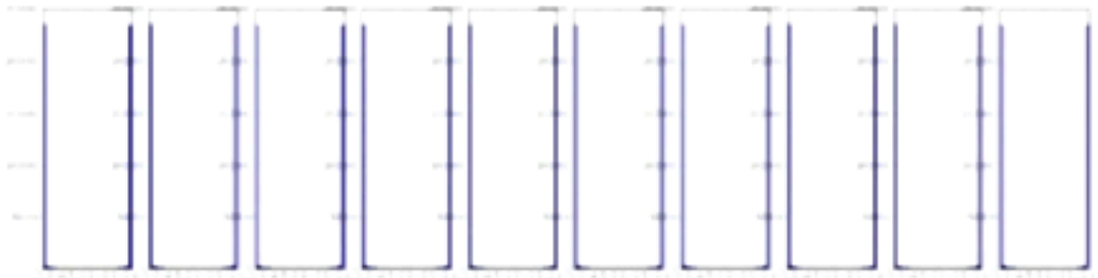
The earlier the weight in the network resides, the more terms will be included in the gradient to be multiplied!

Robotics Research Center

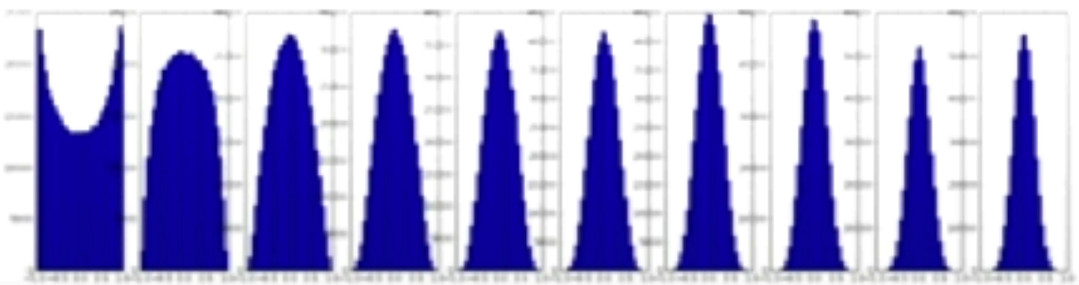
Weight Initialisation



Initialization too small:
Activations go to zero, gradients also zero,
No learning



Initialization too big:
Activations saturate (for tanh),
Gradients zero, no learning



Initialization just right:
Nice distribution of activations at all layers,
Learning proceeds nicely

10 minute break!

Robotics Research Center

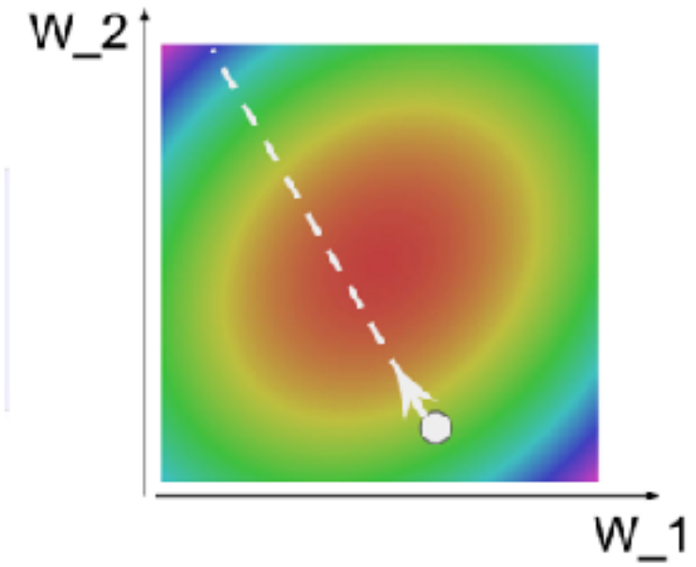
Optimisation Algorithms

- The core strategy in training Neural Networks in solving an optimisation problem! We have some sort of a loss function, that tells us how good our weights are. We need to optimise over the weights and get to the most minimum loss possible.

```
# Vanilla Gradient Descent

while True:
  weights_grad = evaluate_gradient(loss_fun, data, weights)
  weights += - step_size * weights_grad # perform parameter update
```

Stochastic Gradient Descent
A simple optimisation algorithm

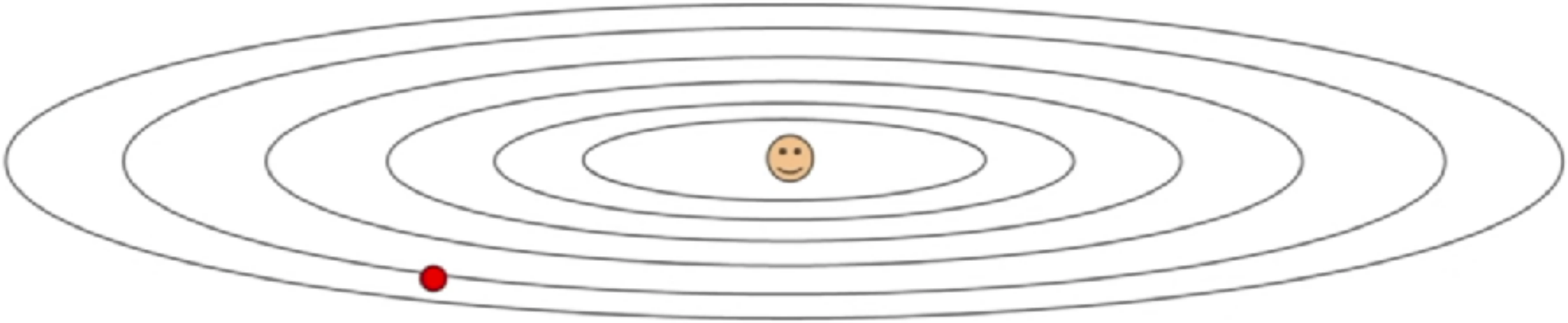


Move in the direction of the greatest decrease in the loss function

Robotics Research Center

Are there any problems with Vanilla Gradient Descent?

- What if our objective function looks like this?



Do you think SGD will work in such a case?

- Loss has a bad condition number: ratio of largest to smallest singular value of the Hessian matrix is large

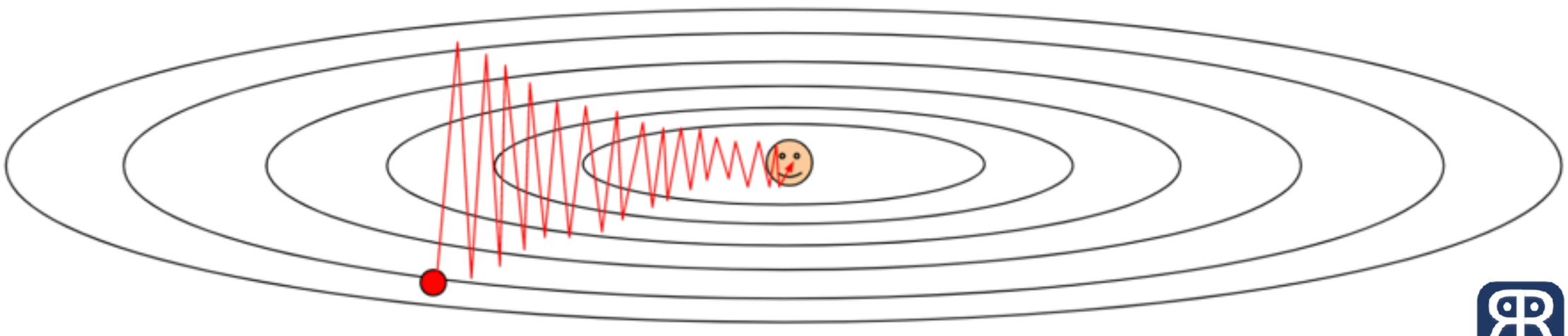
Robotics Research Center

Problems with SGD

- What if the loss is sensitive in one direction and not in the other? That means, it changes quickly in w1 but slowly in w2.

- Very slow progress along horizontal dimension (the slow one). When we calculate the gradients and update, we kinda zig zag back and forth. Nasty progress along the fast changing one.

- This gets worse when we have more dimensions - which means more directions along which it can move.



Robotics Research Center

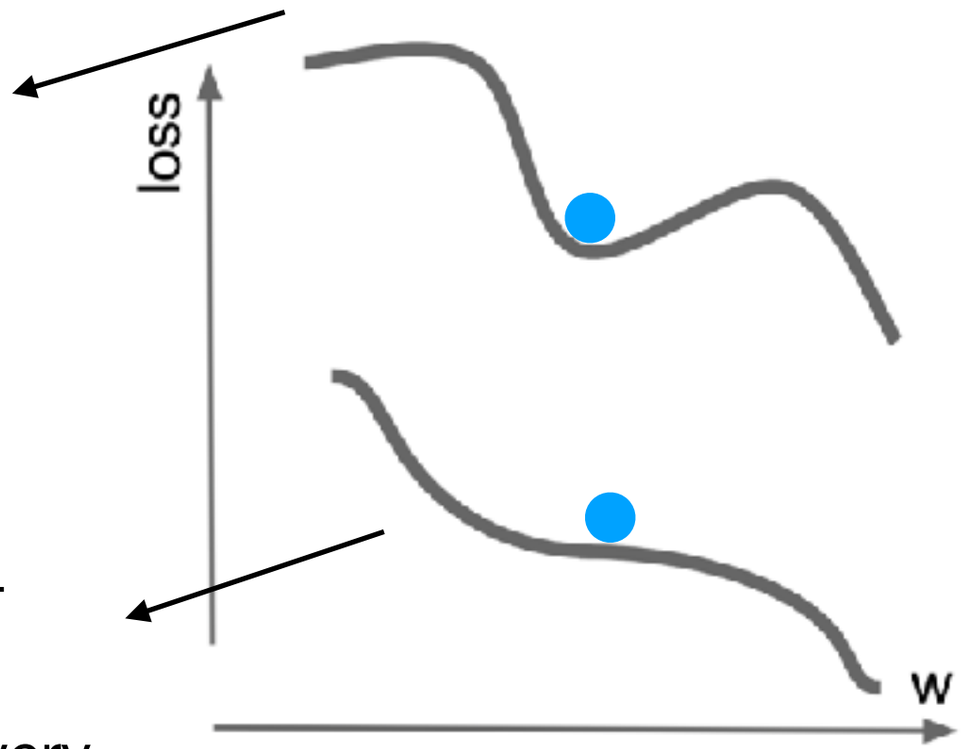
Problems with SGD

- If there a hundreds of different directions to move in, and the ratio of the largest and smallest one is quite large, then SGD wouldn’t perform well.

- Another problem is that of local minima or saddle point.

In this situation, SGD will get stuck!
Because at the local minima, the gradient is zero as it’s locally flat.
In SGD, we compute the gradient and we move in the negative direction. But since we are at a local minima, we’ll get stuck at this point.

Another Problem: In one direction we go up,
And in another we go down - SADDLE POINT
Then also it may get stuck.
The gradient is exactly zero, but the slope is very small. So if the gradient is very small,
we make REALLY SLOW progress

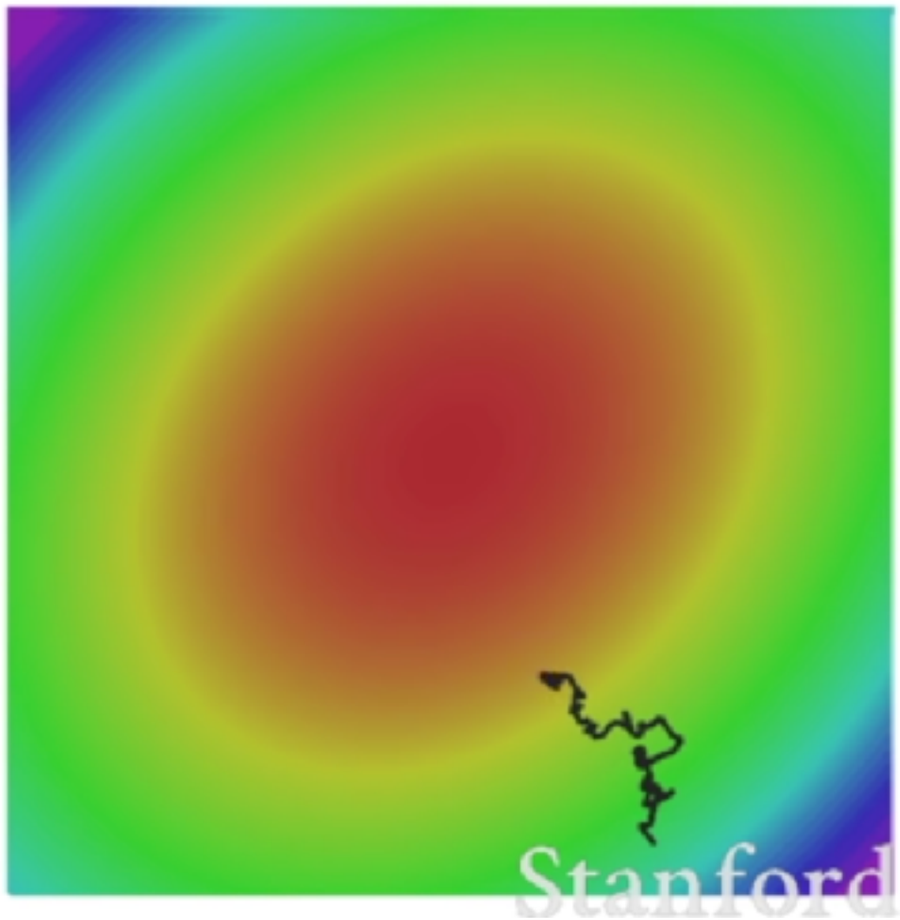


One dimensional weight

Robotics Research Center

‘Stochastic’ Gradient Descent

- Loss function is computed by finding the loss for many different examples. But there could be too many, so we estimate the loss and gradient using a mini batch of training examples.



- So if there’s noise in the gradient estimates, then vanilla SGD might just wander/meander around a bit before actually going towards the minima.

- What if we use just normal GD, that is, full-batch? Doesn’t really solve these problems, so we need a fancier optimisation problem.

Robotics Research Center

SGD + Momentum

- We maintain a ‘velocity’ over time. We add our gradient estimates to the velocity. **And we step in the direction of the velocity rather than the gradient.**

- And a new ‘friction’ hyper-parameter which is used to decay the velocity at every time step.

SGD

$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

```
while True:
  dx = compute_gradient(x)
  x -= learning_rate * dx
```

SGD+Momentum

$$v_{t+1} = \rho v_t + \nabla f(x_t)$$

$$x_{t+1} = x_t - \alpha v_{t+1}$$

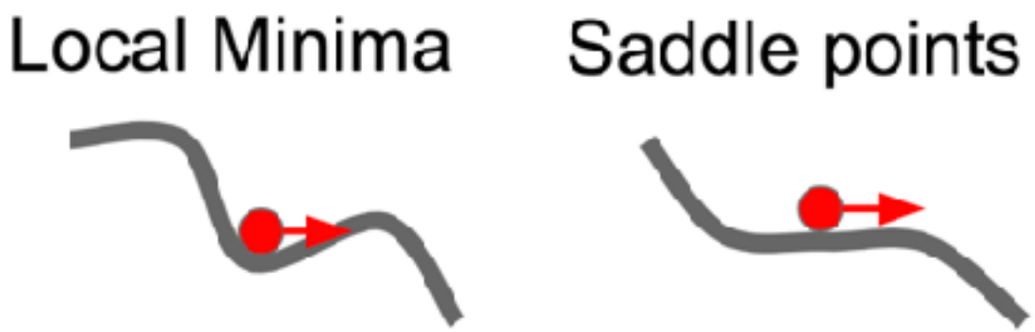
```
vx = 0
while True:
  dx = compute_gradient(x)
  vx = rho * vx + dx
  x -= learning_rate * vx
```

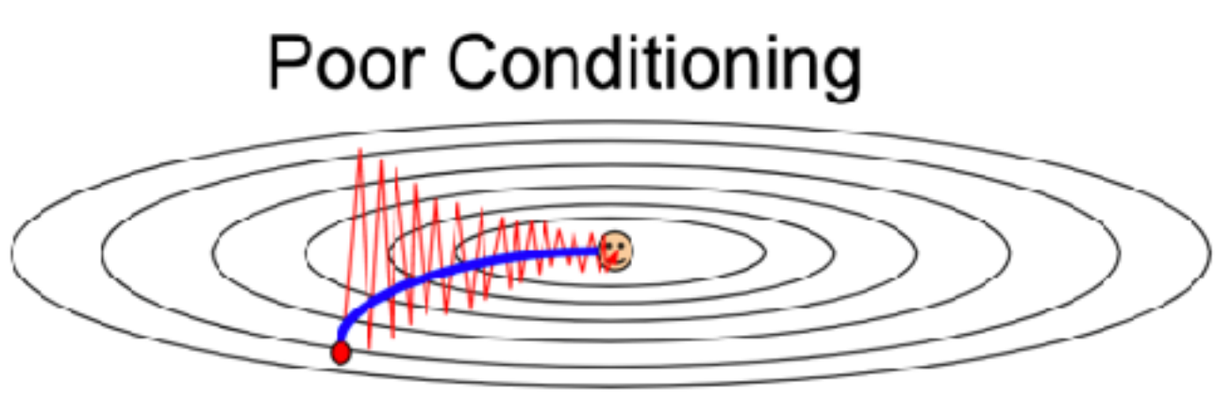
Robotics Research Center

This solves all the problems

At local minima, the point may not have any gradient, but i**t'll still have a 'velocity'.**

So we can get over this local minima and still continue to let the boll roll.





At saddle points, even though the gradient is really less, we have built up a decent amount of velocity as we rolled downhill.

For the 'zig-zag' problem, the zig zags would cancel each other out while using momentum, reducing the amount by which we step in the sensitive direction, and the velocity will keep building up in the horizontal direction and accelerate across the less sensitive direction.

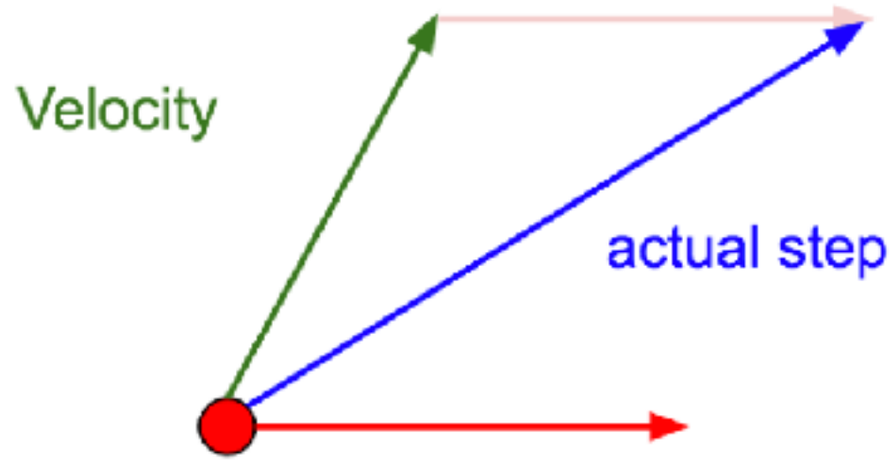
Robotics Research Center

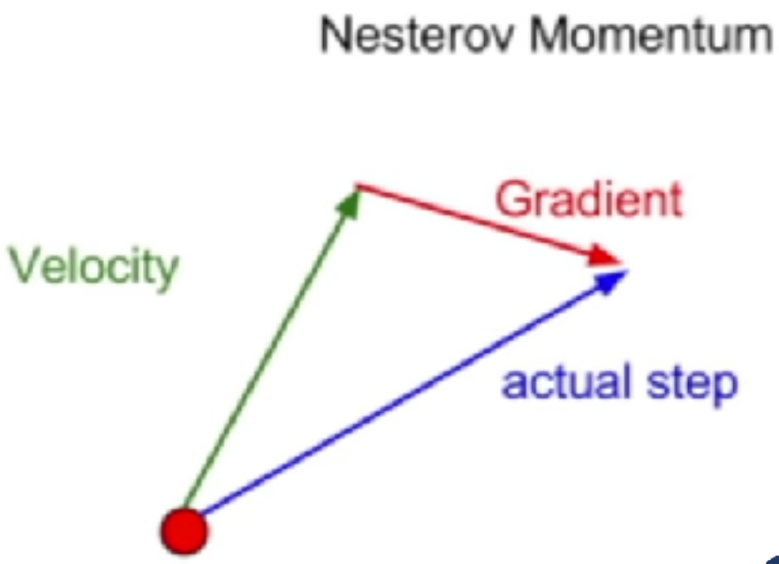
Momentum Updates

- We compute the weighted average of the gradient estimate and the velocity vector and we step across this ‘actual step’.

- So we’re taking a mix of the gradient and the velocity to overcome the noise and all the difficulties discussed.

- That’s why, SGD + Momentum





Robotics Research Center

AdaGrad: Another optimisation strategy

- AdaGrad keeps a running sum estimate of the squared gradients. So instead of velocity, we now have a ‘grad_squared’ term.

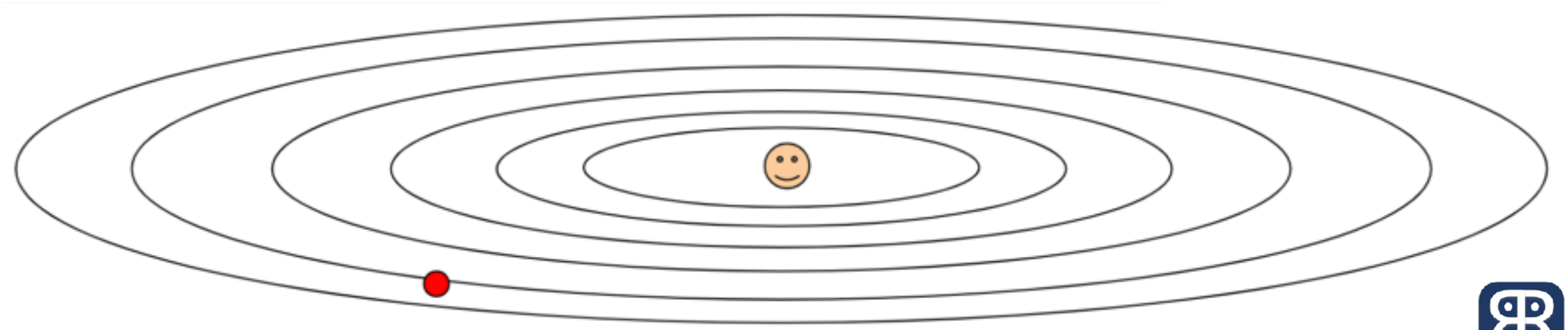
```
grad_squared = 0
while True:
  dx = compute_gradient(x)
  grad_squared += dx * dx
  x -= learning_rate * dx / (np.sqrt(grad_squared) + 1e-7)
```

(So that we don’t divide by zero…)

Robotics Research Center

How it helps!

- The idea is that, if there's a high gradient along one direction and low gradient along the other, then if add the sum of squares of the small gradient, we divide by a smaller number and hence, we'll accelerate movement along the slower dimension.
- And along the other dimension where the gradients tend to be very large, we divide by a larger number, thereby slowing down the progress along that wiggling dimension.



Robotics Research Center

What happens over the course of training?

- In AdaGrad, progress along ‘steep’ directions is damped and progress along ‘flat’ directions is accelerated.

- But over time, the step size gets smaller and smaller. It’s good in the convex case because as you approach the minima, you want to slow down and converge.

- But in non convex case, if you come across a saddle point, you get stuck and can’t make any progress.

The grad_squared term keeps building up

RMSProp: Leaky AdaGrad

AdaGrad

```
grad_squared = 0
while True:
  dx = compute_gradient(x)
  grad_squared += dx * dx
  x -= learning_rate * dx / (np.sqrt(grad_squared) + 1e-7)
```

RMSProp

```
grad_squared = 0
while True:
  dx = compute_gradient(x)
  grad_squared = decay_rate * grad_squared + (1 - decay_rate) * dx * dx
  x -= learning_rate * dx / (np.sqrt(grad_squared) + 1e-7)
```

We still keep the estimates of the squared gradients. But instead of just letting it accumulate over training, we let it decay over time - just like the momentum update we had seen earlier.
So it's like a momentum update over squared gradients!
And we don't slow down where don't want to, as the estimates are leaky.

Adam

- The core idea behind momentum was to move in the direction of the velocity to not get stuck in local minima.

- The core idea behind AdaGrad and RMSProp was to compute a square of the gradients to slow down movement along the sensitive direction and prevent the zig zags.

- The core idea behind Adam: Why not combine these two?!

```
first_moment = 0
second_moment = 0
while True:
  dx = compute_gradient(x)
  first_moment = beta1 * first_moment  + (1 - beta1) * dx
  second_moment = beta2 * second_moment + (1 - beta2) * dx * dx
  x -= learning_rate * first_moment / (np.sqrt(second_moment) + 1e-7))
```

Robotics Research Center

```
first_moment = 0
second_moment = 0
while True:
  dx = compute_gradient(x)
  first_moment = beta1 * first_moment  + (1 - beta1) * dx
  second_moment = beta2 * second_moment + (1 - beta2) * dx * dx
  x -= learning_rate * first_moment / (np.sqrt(second_moment) + 1e-7))
```

Momentum

AdaGrad / RMSProp

Sort of like RMSProp with momentum

First moment: the velocity term
Second moment: the square gradients term

But there's a slight problem here…what happens at the first time-step?

Robotics Research Center

- The second moment is initialised to zero. And the decay rate is very very low, like 0.9 or something (beta2). So after one update, second moment is still very close to zero.

- When we make the update step, we divide by a very small number and hence, make a huge step in the beginning, which is because we initialised it to zero.

- Of course, since the first moment is also very small, it may cancel out the second moment, but sometimes, it does result in taking very large steps in the beginning.

```
first_moment = 0
second_moment = 0
while True:
  dx = compute_gradient(x)
  first_moment = beta1 * first_moment  + (1 - beta1) * dx
  second_moment = beta2 * second_moment + (1 - beta2) * dx * dx
  x -= learning_rate * first_moment / (np.sqrt(second_moment) + 1e-7))
```

Robotics Research Center

We add a ‘bias correction’ step

```
first_moment = 0
second_moment = 0
for t in range(1, num_iterations):
  dx = compute_gradient(x)
  first_moment = beta1 * first_moment  + (1 - beta1) * dx
  second_moment = beta2 * second_moment + (1 - beta2) * dx * dx
  first_unbias = first_moment / (1 - beta1 ** t)
  second_unbias = second_moment / (1 - beta2 ** t)
  x -= learning_rate * first_unbias / (np.sqrt(second_unbias) + 1e-7))
```

Momentum

Bias correction

AdaGrad / RMSProp

We create an unbiased estimate of the first and second moments using the current tilmestep ’t’. We make our steps using these unbiased estimates

Adam with beta1 = 0.9,
beta2 = 0.999, and learning_rate = 1e-3 or 5e-4
is a great starting point for many models!

Check out the lecture video at 43:05 - shows how well Adam works: https://www.youtube.com/watch?v=_JB0AO7QxSA

Robotics Research Center

Recap: Optimisation Algorithms

- Stochastic Gradient Descent (The classic one)
- SGD + Momentum
- AdaGrad
- RMSProp
- Adam (The Best One)

Robotics Research Center

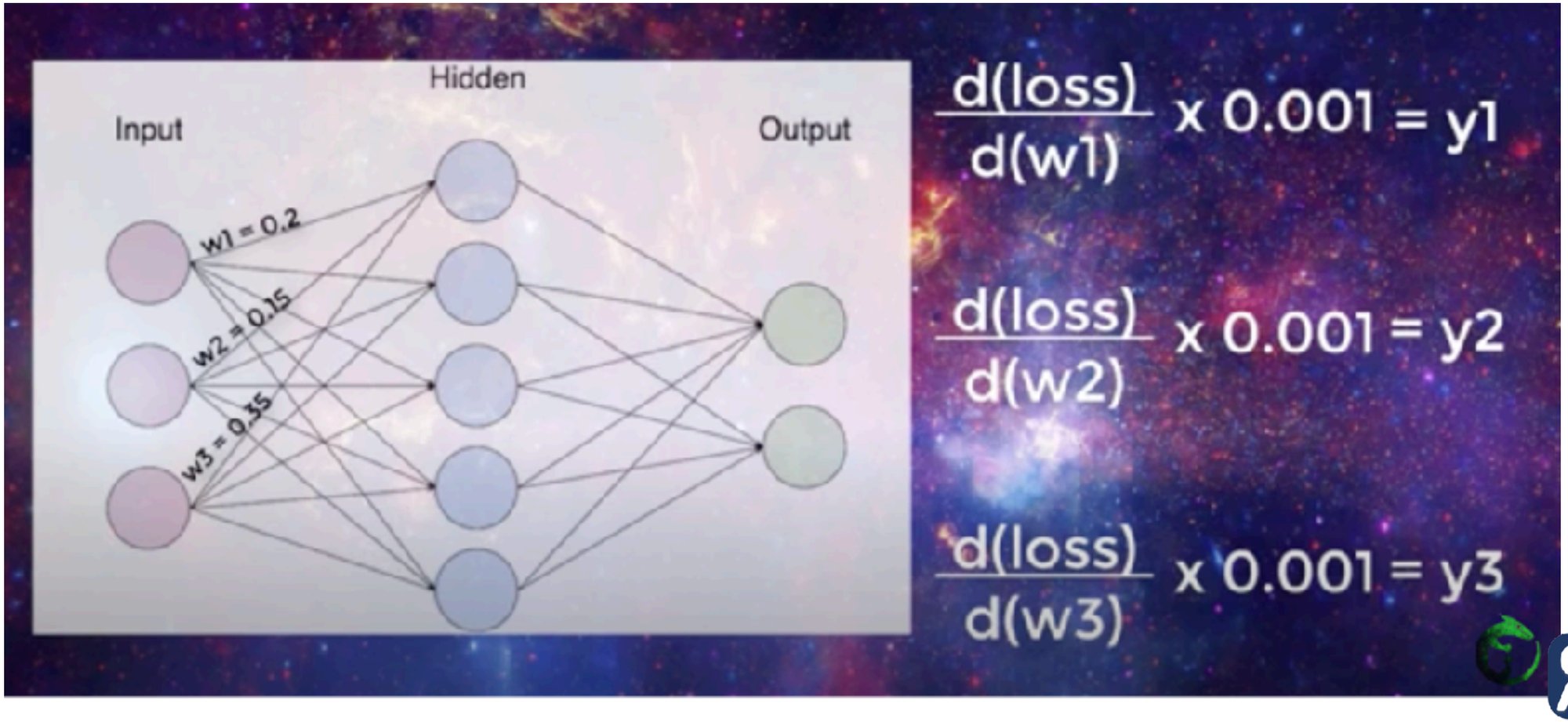
Learning Rate or Step Size

- A number by which we multiply our resulting gradient by!
- The objective during training is for SGD or Adam to minimise the loss between the actual output and predicted output.
- We start with an initialisation of the weights and keep updating them with every iteration to move towards the minimum loss.
- The size of the steps we take to reach this minimised loss is going to depend on the learning rate.

- After calculating loss for the given inputs, the gradient of that loss is calculated with respect to each of the weights in the model.

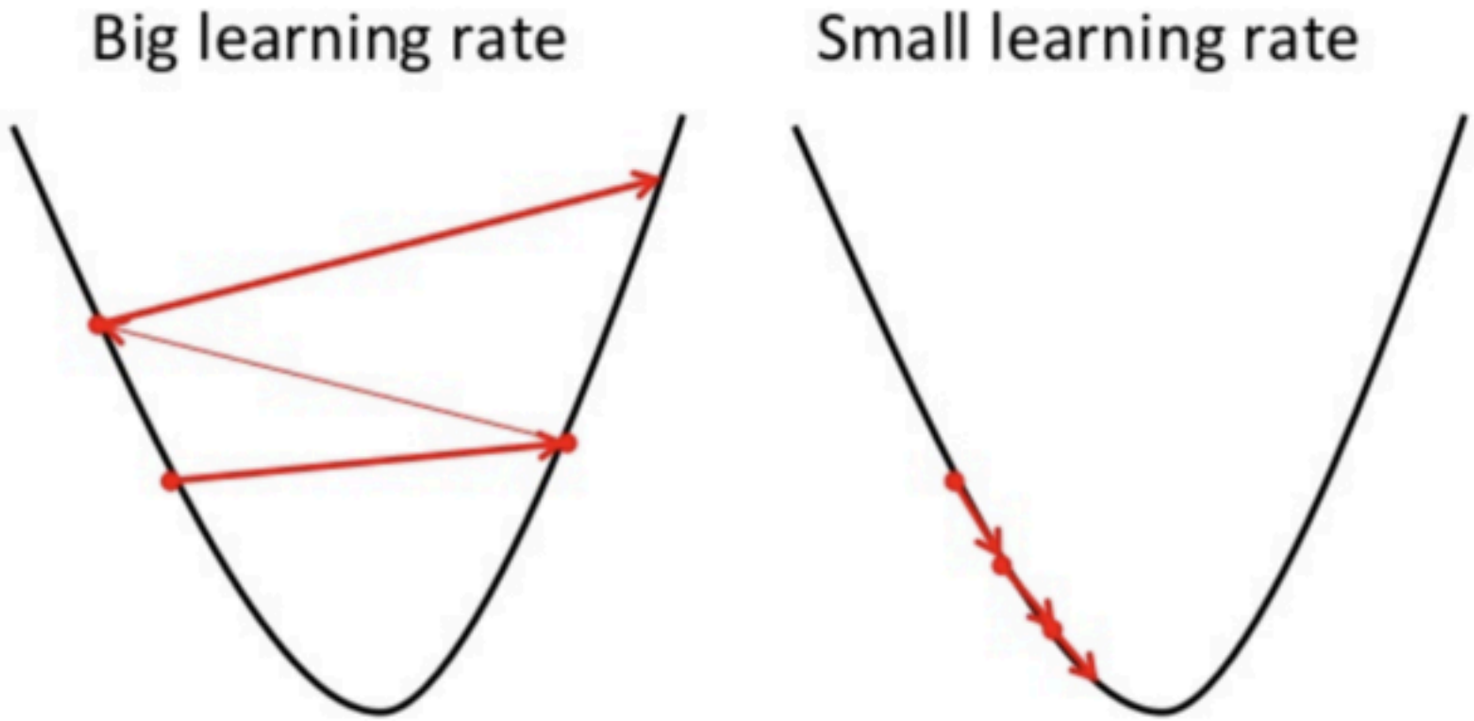
- These gradients will be multiplied by the learning rate (somewhere between 0.1 and 0.001).

- We then update the weights by subtracting these values from them (to move in the negative direction).



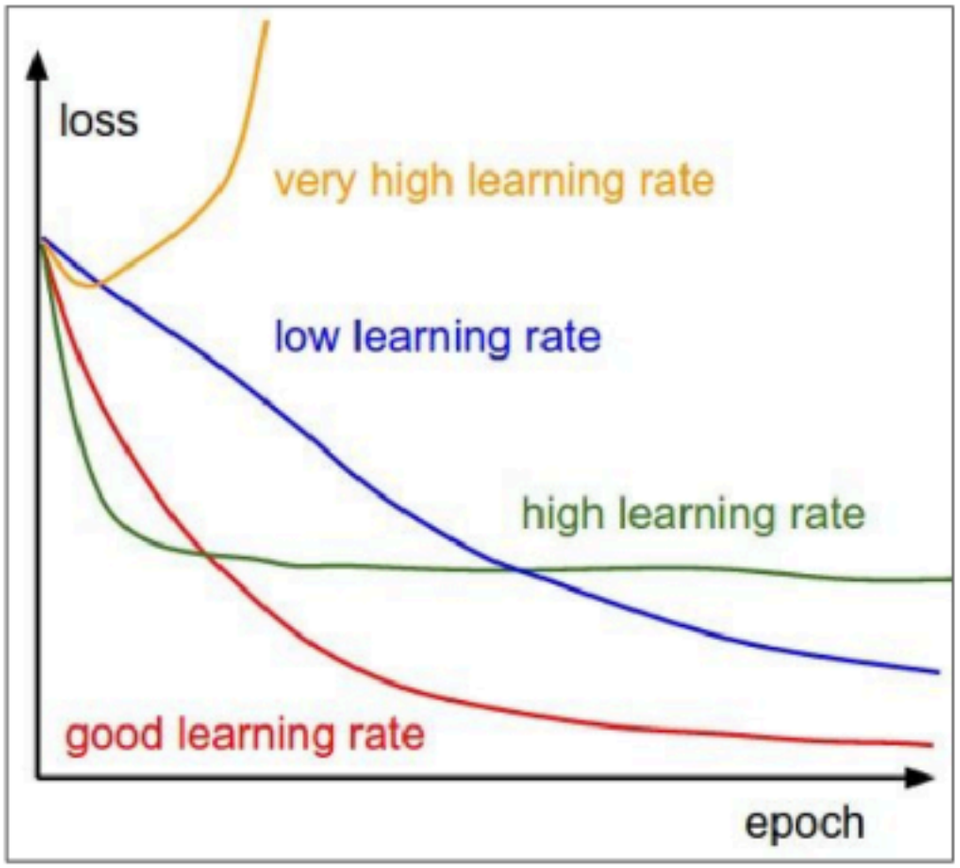
- If learning rate is too high? OVERSHOOTING, that is we take a step that's too large in the direction of the loss. So we shoot past the minimum and miss it.

- If small learning rate, then we take very small baby steps.



Effect of learning rate on gradient descent

SGD, SGD+Momentum, Adagrad, RMSProp, Adam all have **learning rate** as a hyperparameter.



=> Learning rate decay over time!

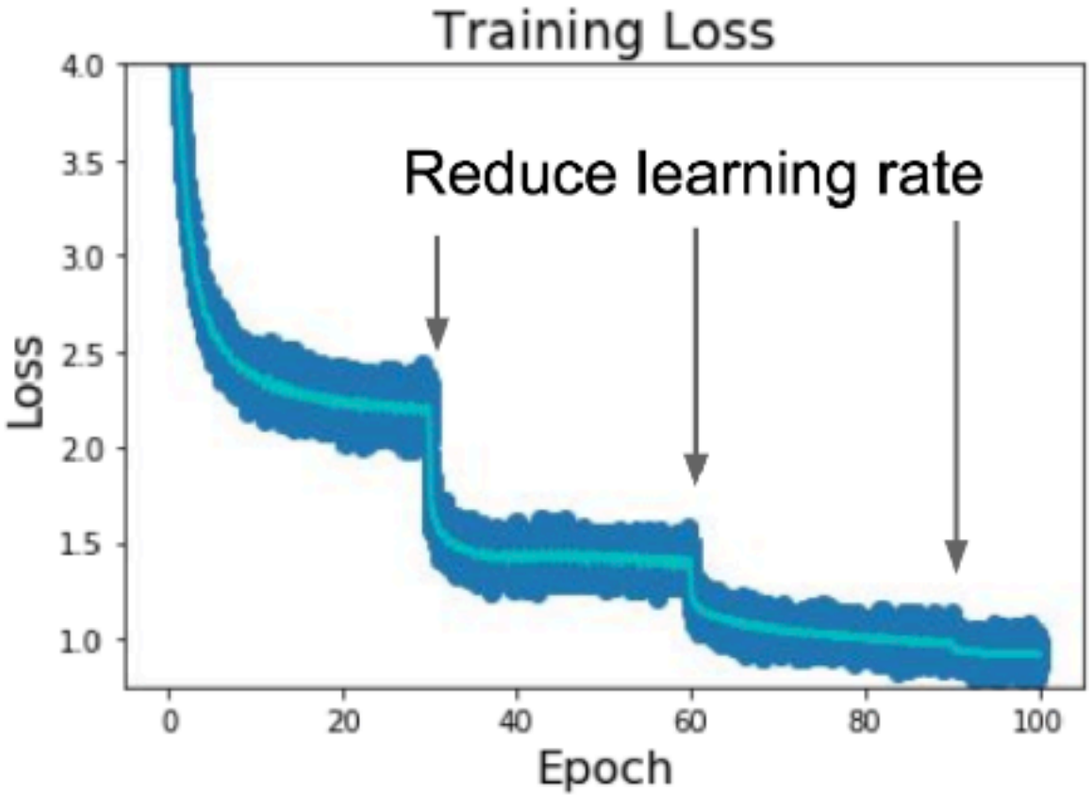
step decay:
e.g. decay learning rate by half every few epochs.

exponential decay:

$$\alpha = \alpha_0 e^{-kt}$$

1/t decay:

$$\alpha = \alpha_0 / (1 + kt)$$

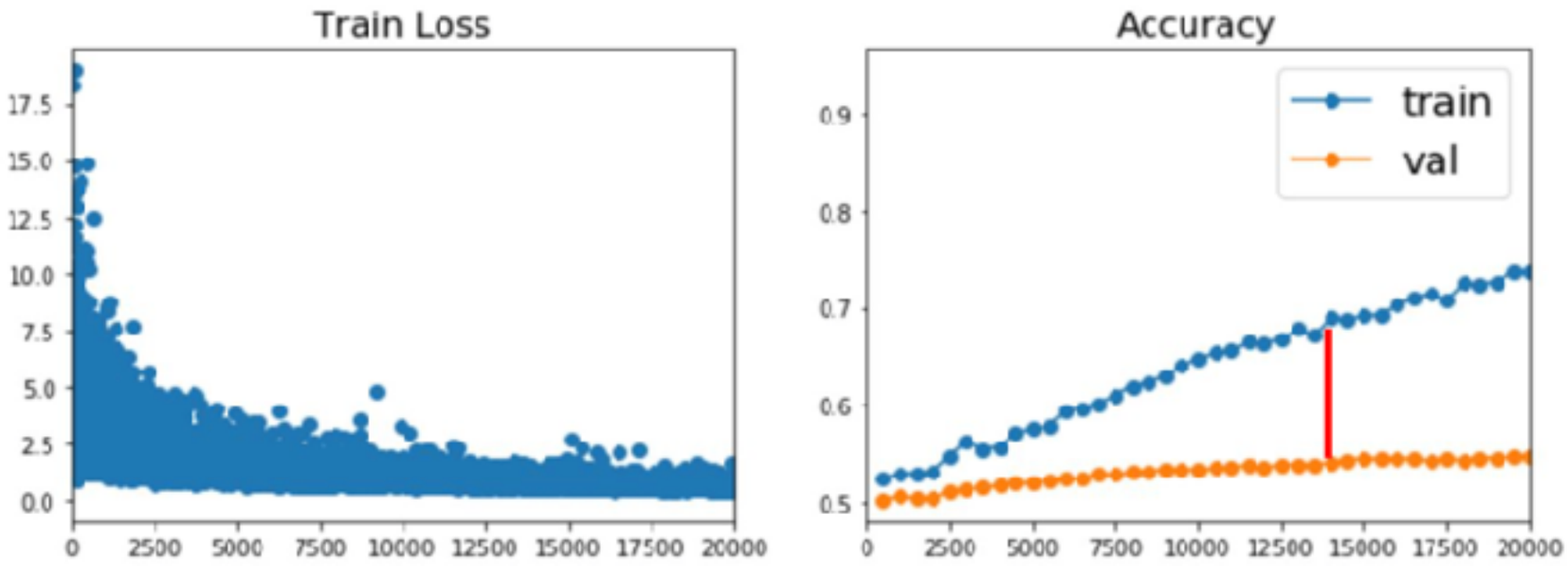


Step: Reduce learning rate at a few fixed points. E.g. for ResNets, multiply LR by 0.1 after epochs 30, 60, and 90.

‘Step Decay’ - we reduce the learning rate at particular steps, so if it would’ve gotten stuck somewhere, we can take smaller baby steps and maybe reach a better region.

Beyond Training Error

- So far, we've been trying to reduce the train error, the one between predicted output and actual output of the network.
- But what we care about is the performance on unseen data!



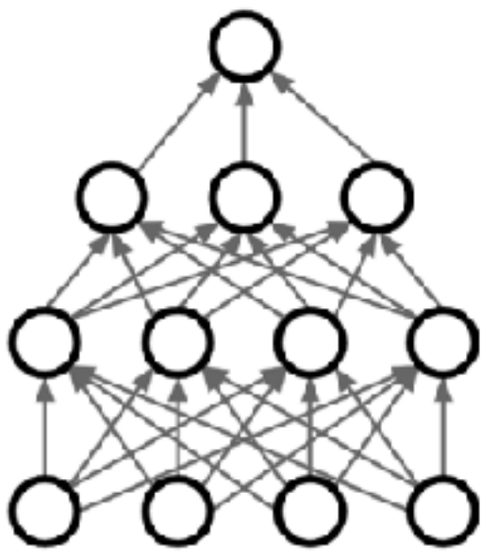
Robotics Research Center

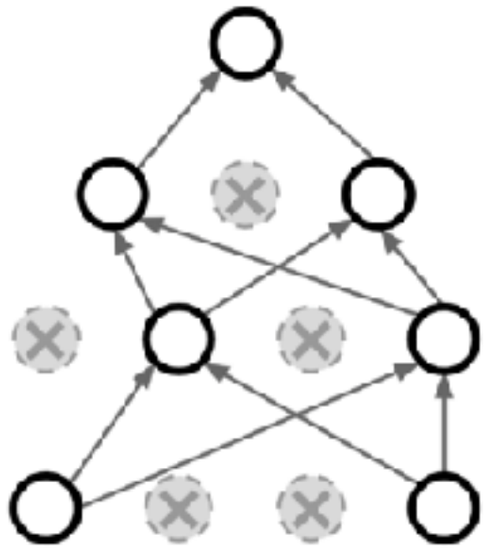
Regularisation: Dropout

- We add something to our model to prevent it from fitting our training model too well, so it could perform better on unseen data!

- In dropout, we randomly set some neurons to zero for each forward pass.

- Dropout is like training a large ensemble of models!

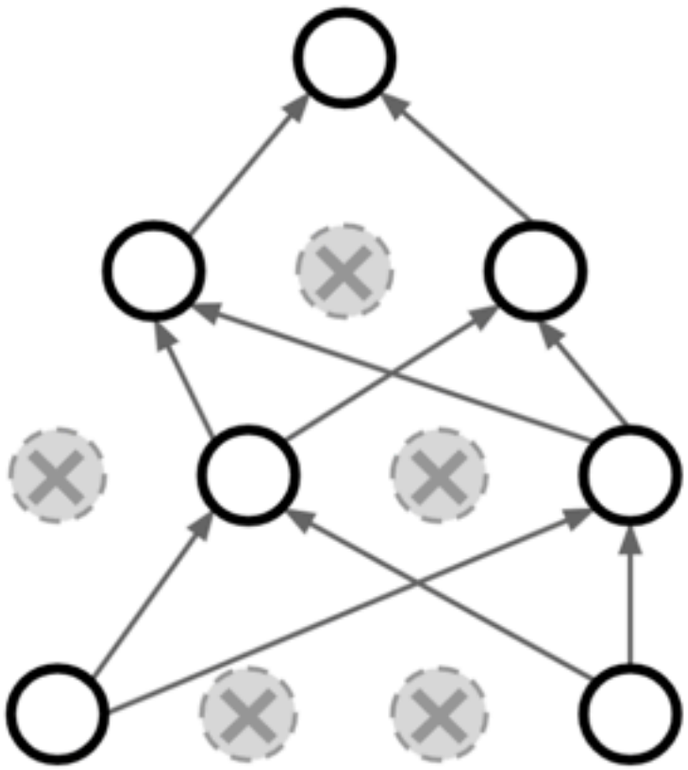




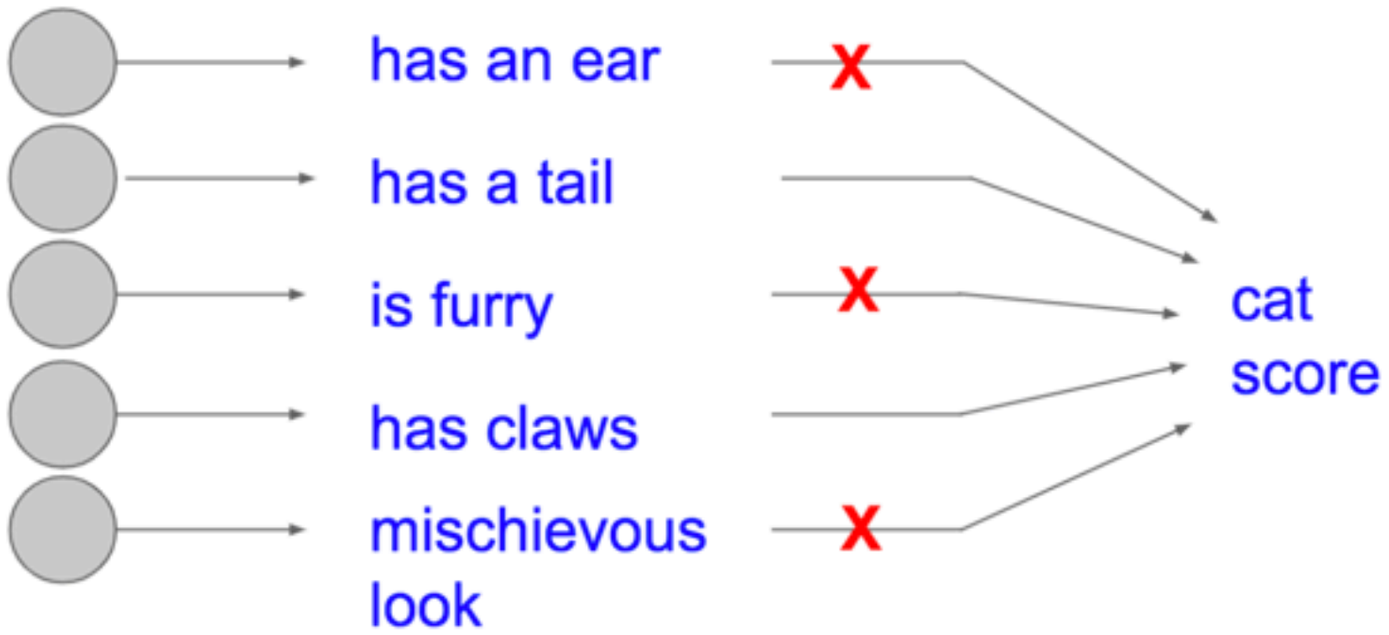
Robotics Research Center

Regularization: Dropout

How can this possibly be a good idea?



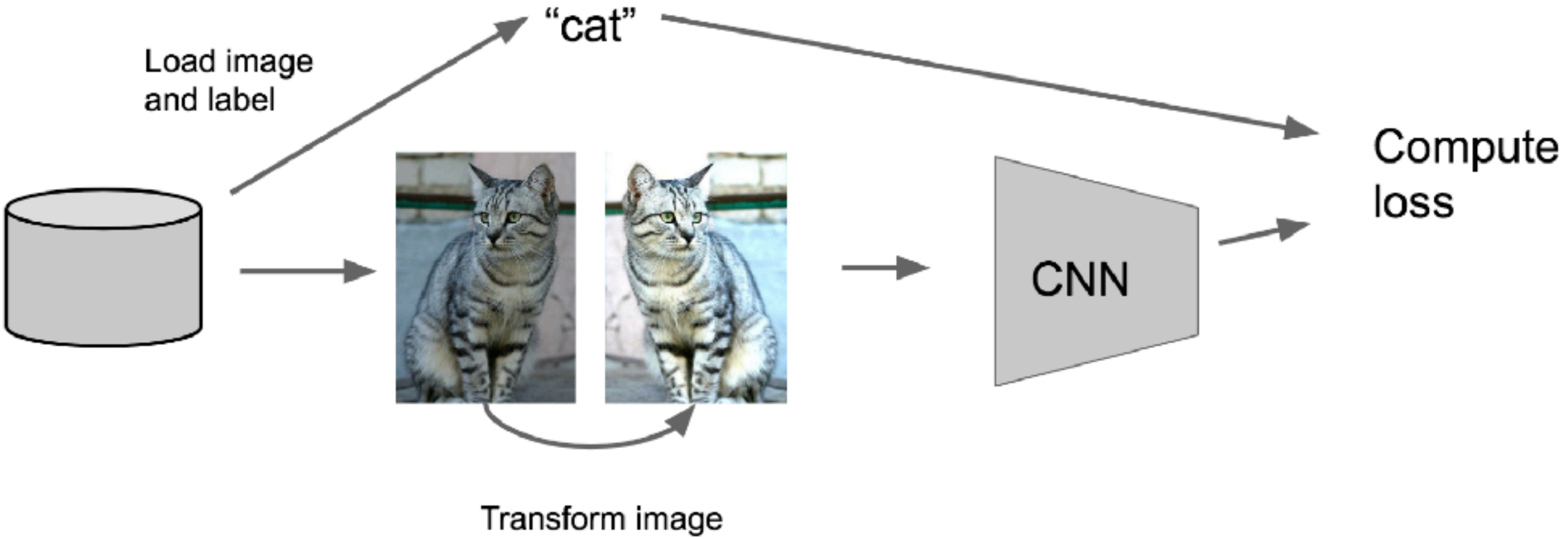
Forces the network to have a redundant representation;
Prevents co-adaptation of features



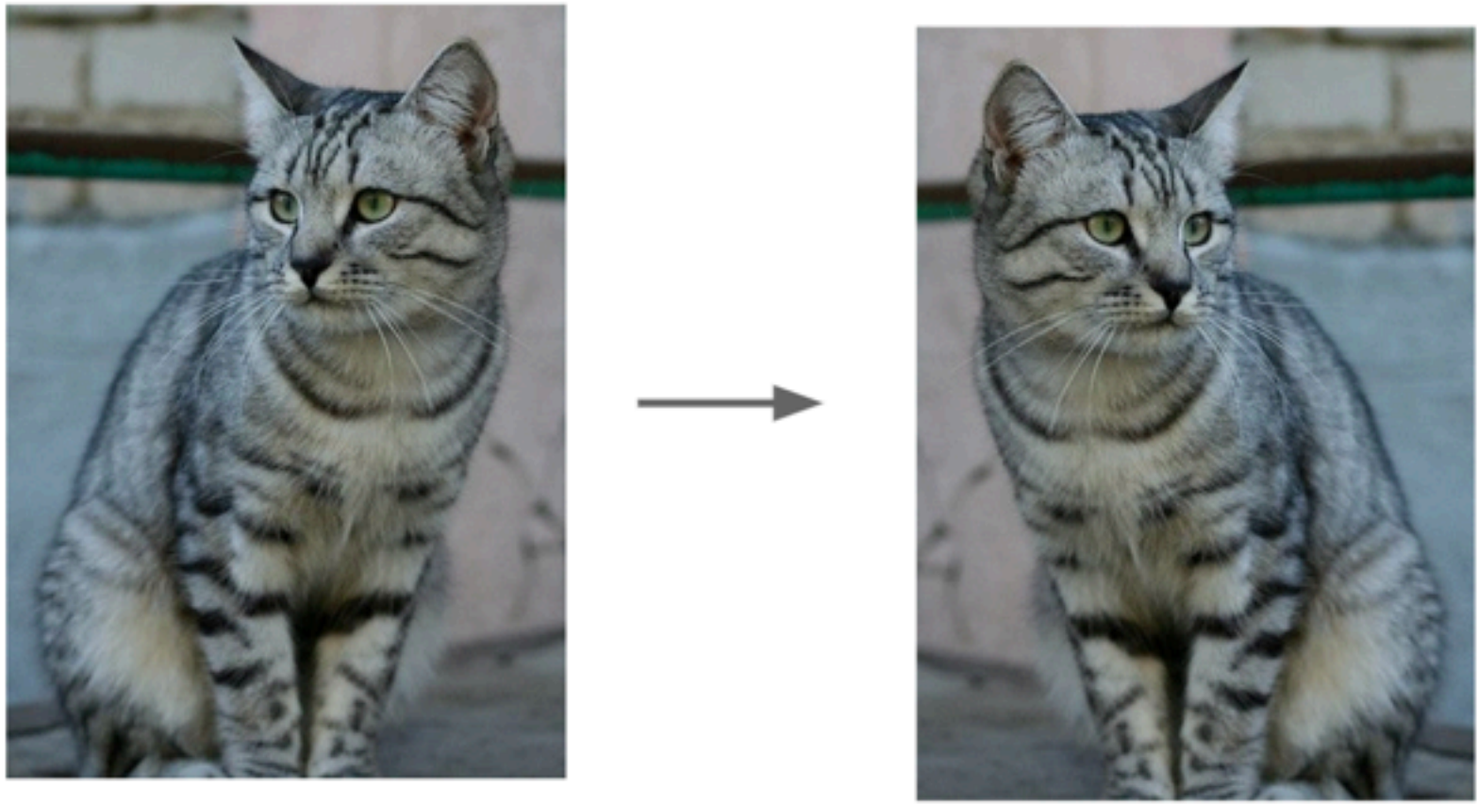
We want to prevent over fitting and co adaptation of all the features. So even if the cat doesn't have a tail or isn't furry, the model should be able to detect the cat-ness. This helps increase performance for unseen data.

Robotics Research Center

Regularisation: Data Augmentation



Horizontal Flip



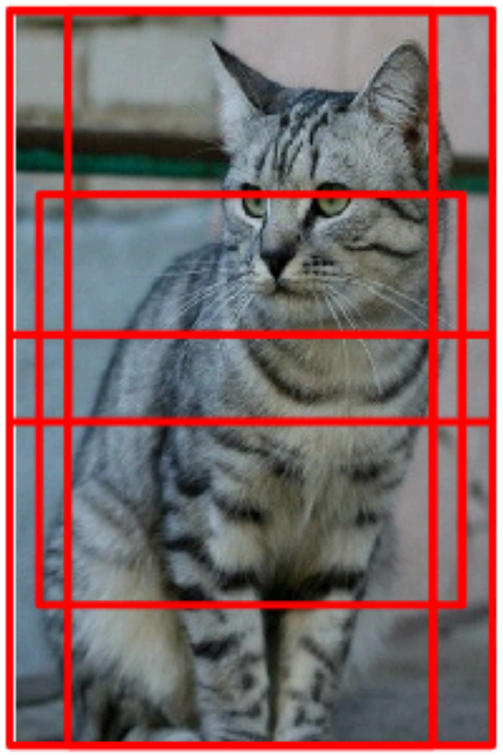
Data Augmentation

Random crops and scales

Training: sample random crops / scales

ResNet:

1. Pick random L in range [256, 480]
2. Resize training image, short side = L
3. Sample random 224 x 224 patch



Testing: average a fixed set of crops

ResNet:

1. Resize image at 5 scales: {224, 256, 384, 480, 640}
2. For each size, use 10 224 x 224 crops: 4 corners + center, + flips

Color Jitter

Simple: Randomize
contrast and brightness

Random mix/combinations of :

- translation
- rotation
- stretching
- shearing,
- lens distortions, … (go crazy)

Regularization: DropConnect

Training: Drop connections between neurons (set weights to 0)

Testing: Use all the connections

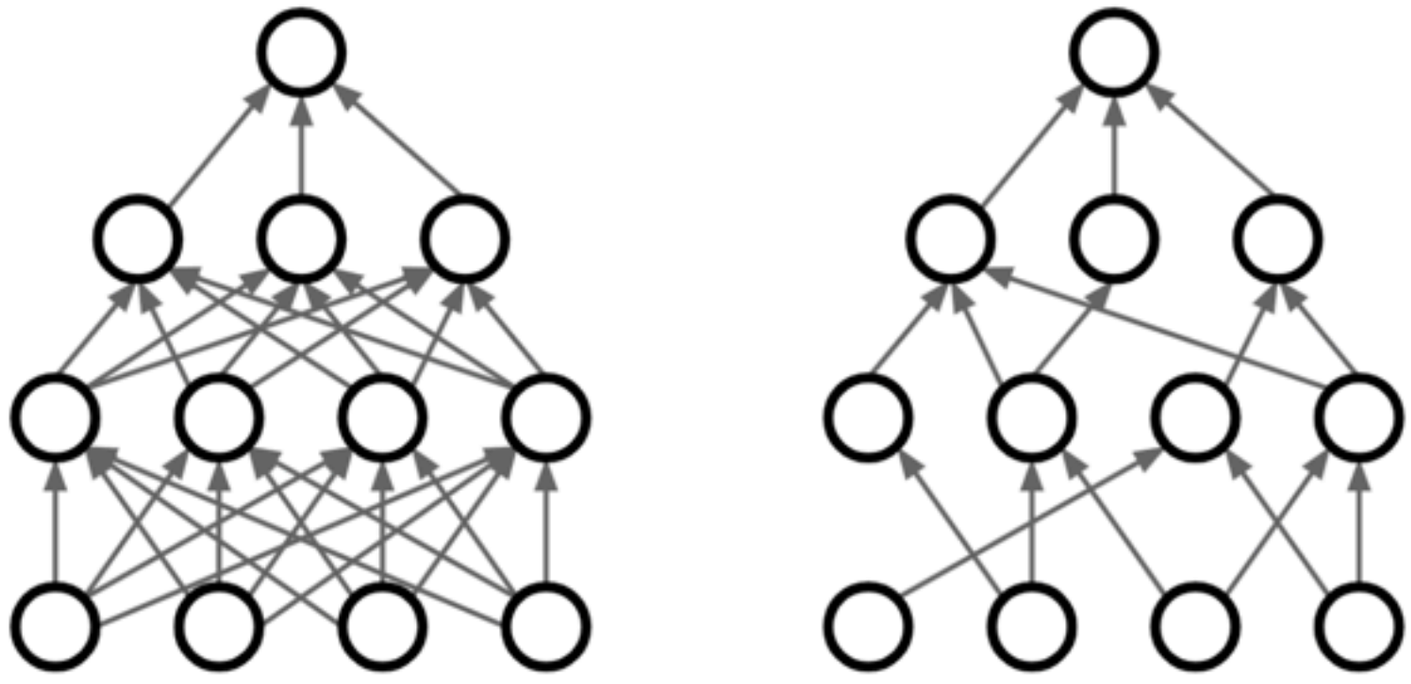
Examples:

Dropout

Batch Normalization

Data Augmentation

DropConnect



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

Regularization: Mixup

Training: Train on random blends of images

Testing: Use original images

Examples:

Dropout
Batch Normalization
Data Augmentation
DropConnect
Fractional Max Pooling
Stochastic Depth
Cutout / Random Crop
Mixup

Target label:
cat: 0.4
dog: 0.6

Randomly blend the pixels of pairs of training images, e.g. 40% cat, 60% dog

Zhang et al, "*mixup*: Beyond Empirical Risk Minimization", ICLR 2018

Transfer Learning

- You need a lot of data if you wanna train CNN's. If you have a really small dataset, then data augmentation helps to increase the amount of data you have.

- But we can also accomplish this without having to do data augmentation, by Transfer Learning.

- As the name suggests, you 'transfer' the weights the network learns and learn some more with the small dataset. That is, you fine tune the network for the new dataset, keeping the weights of the previous one.

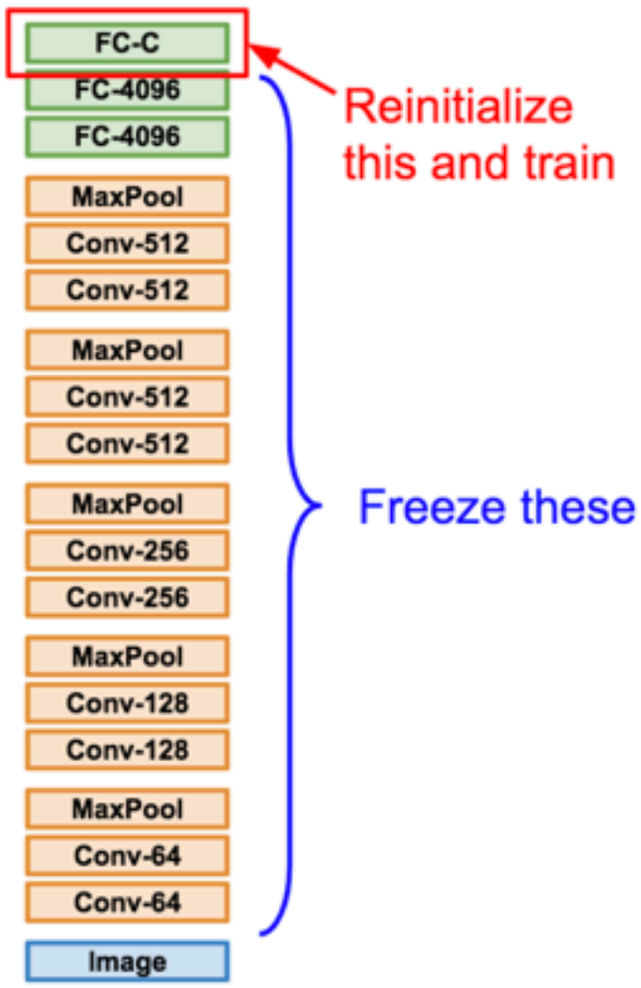
- Very common nowadays, even Fast RCNN uses a CNN that is pertained on ImageNet.

Transfer Learning with CNNs

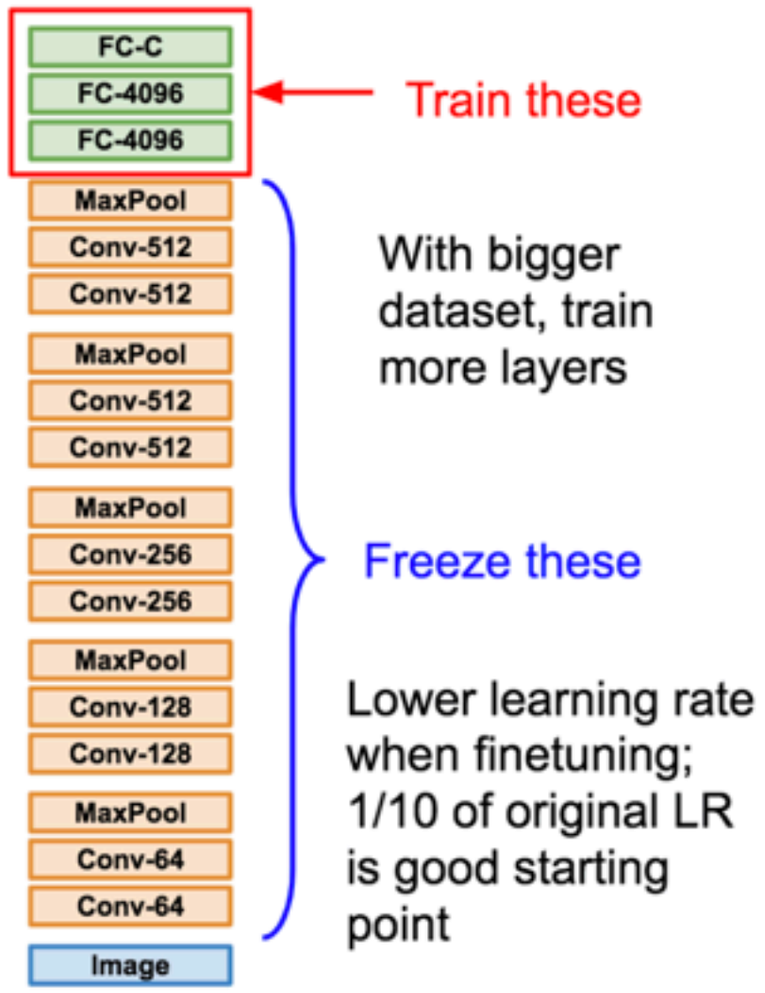
Donahue et al, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", ICML 2014
Razavian et al, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition", CVPR Workshops 2014

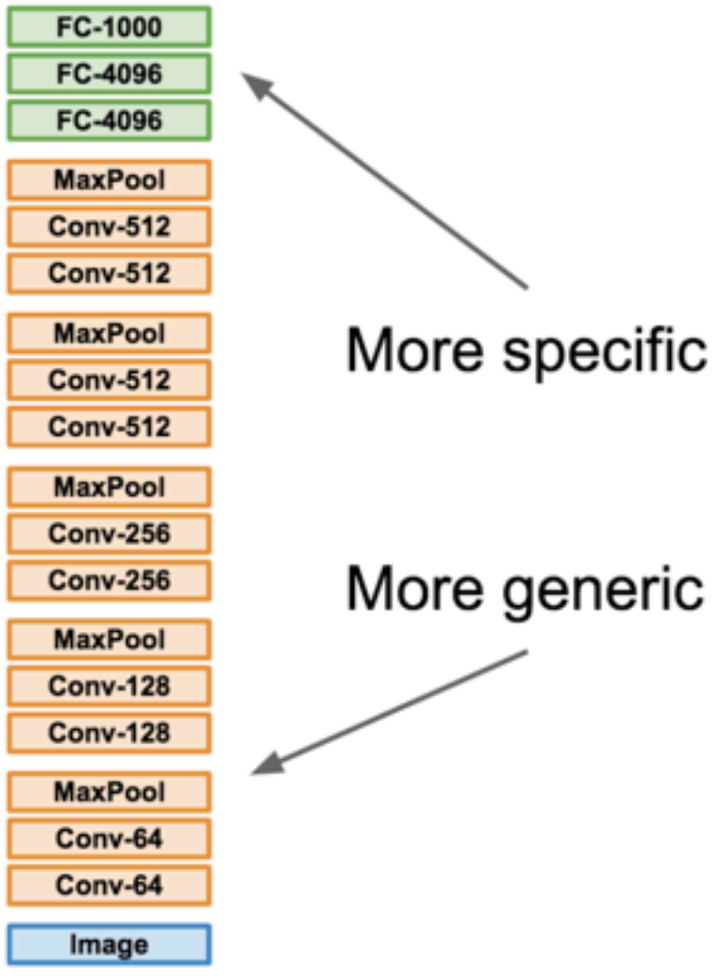
1. Train on Imagenet

2. Small Dataset (C classes)



3. Bigger dataset





	very similar dataset	**very different dataset**
very little data	Use Linear Classifier on top layer	You're in trouble… Try linear classifier from different stages
quite a lot of data	Finetune a few layers	Finetune a larger number of layers