

Lecture 3

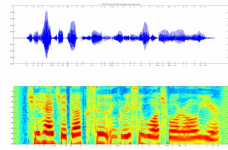
Introduction to Deep Learning

Vladislav Goncharenko

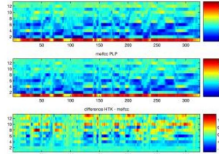
Moscow, 2021

1. Neural Networks in different areas, historical overview
2. Backpropagation
3. More on backpropagation
4. Activation functions
5. Playground

Audio Features



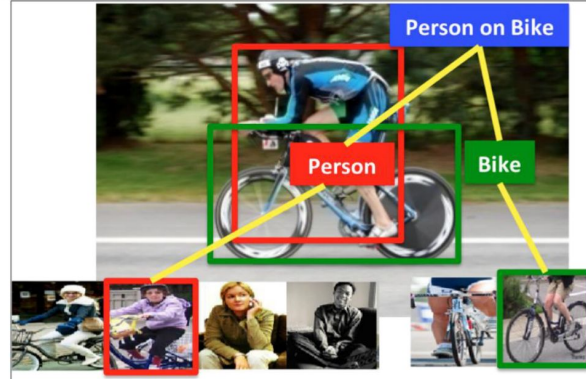
Spectrogram



MFCC



- Object detection
- Action classification
- Image captioning
- ...

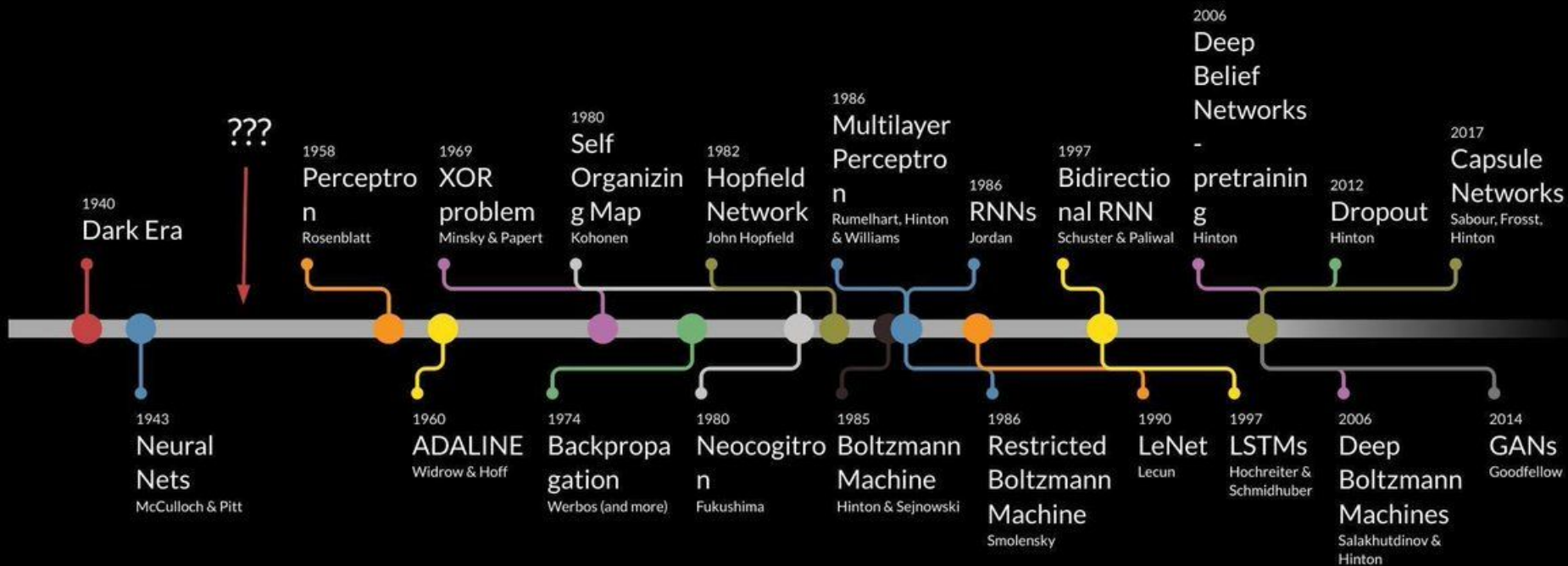


Real world applications

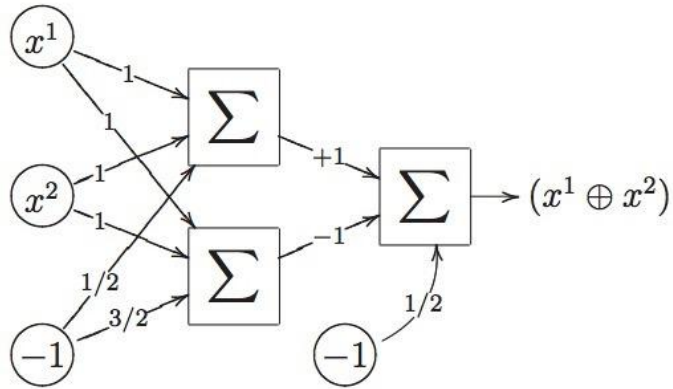


man in black shirt is playing guitar.

Deep Learning Timeline

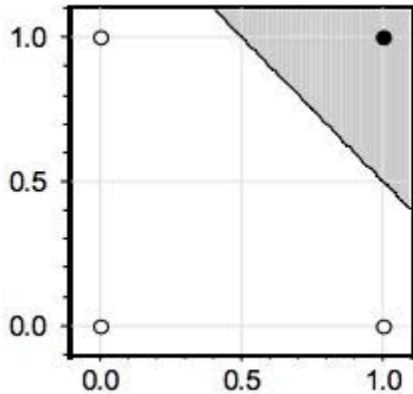


XOR problem

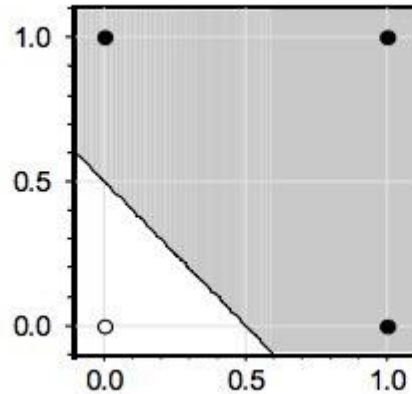


This 2-layer NN (on the left) implements XOR with only x^1 and x^2 features.

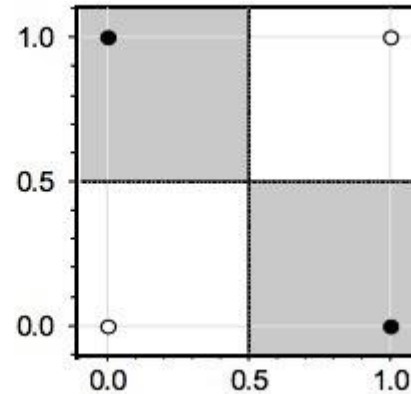
1-layer NN also can succeed, but only with extra feature $x^1 * x^2$.



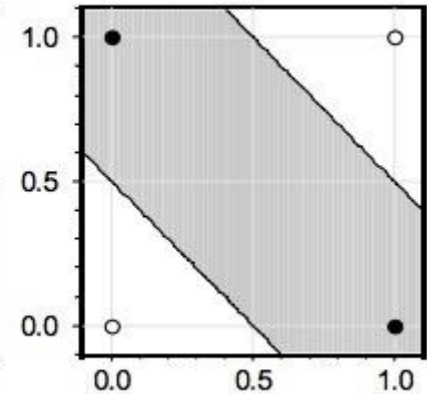
AND



OR

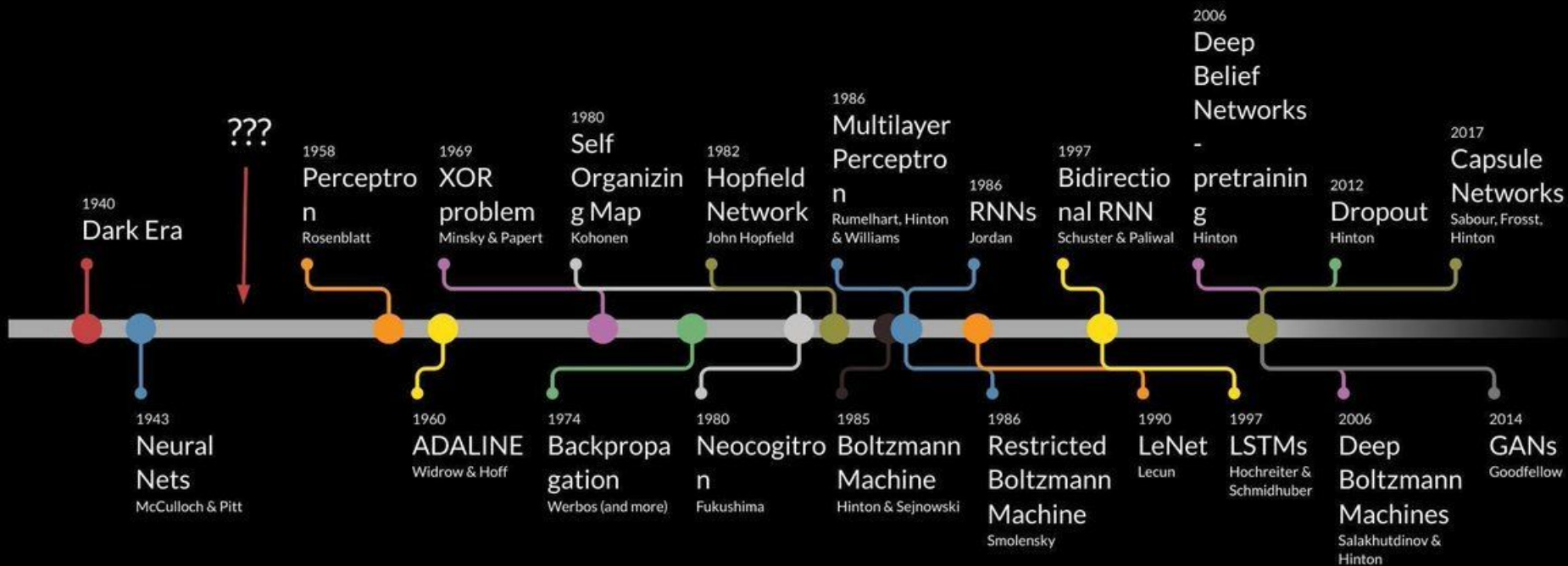


XOR(with $x^1 * x^2$)

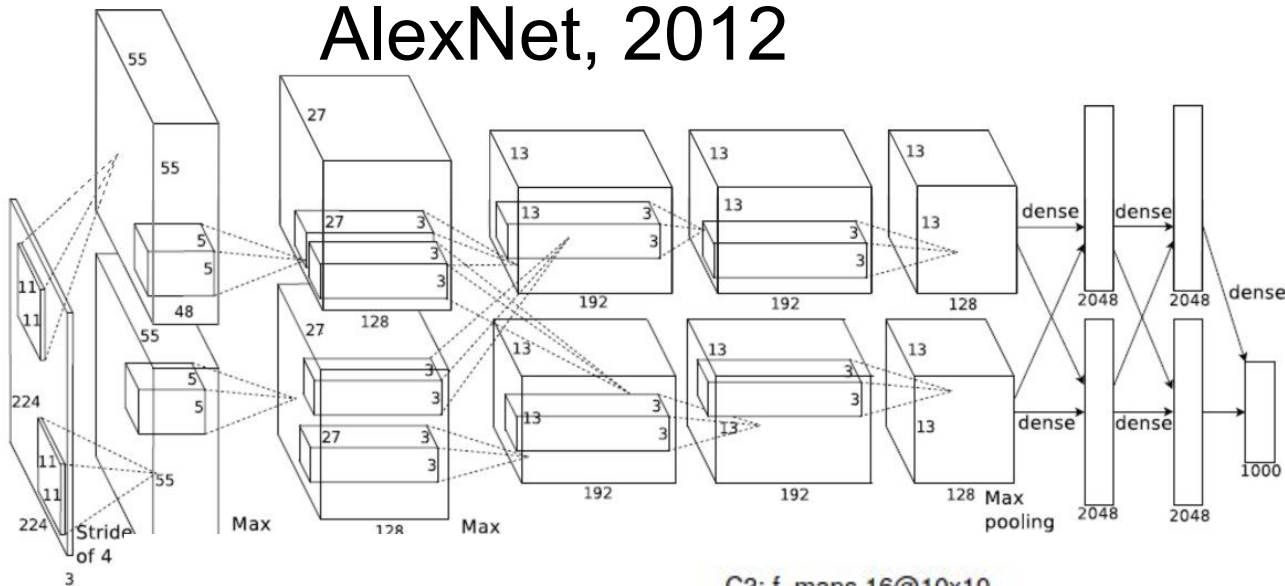


XOR

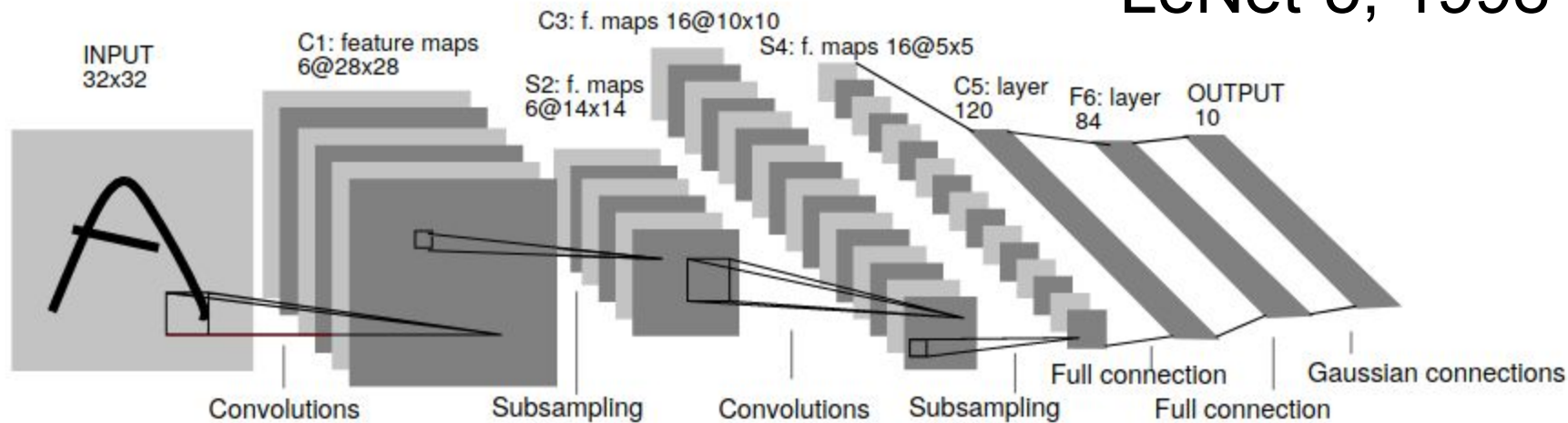
Deep Learning Timeline



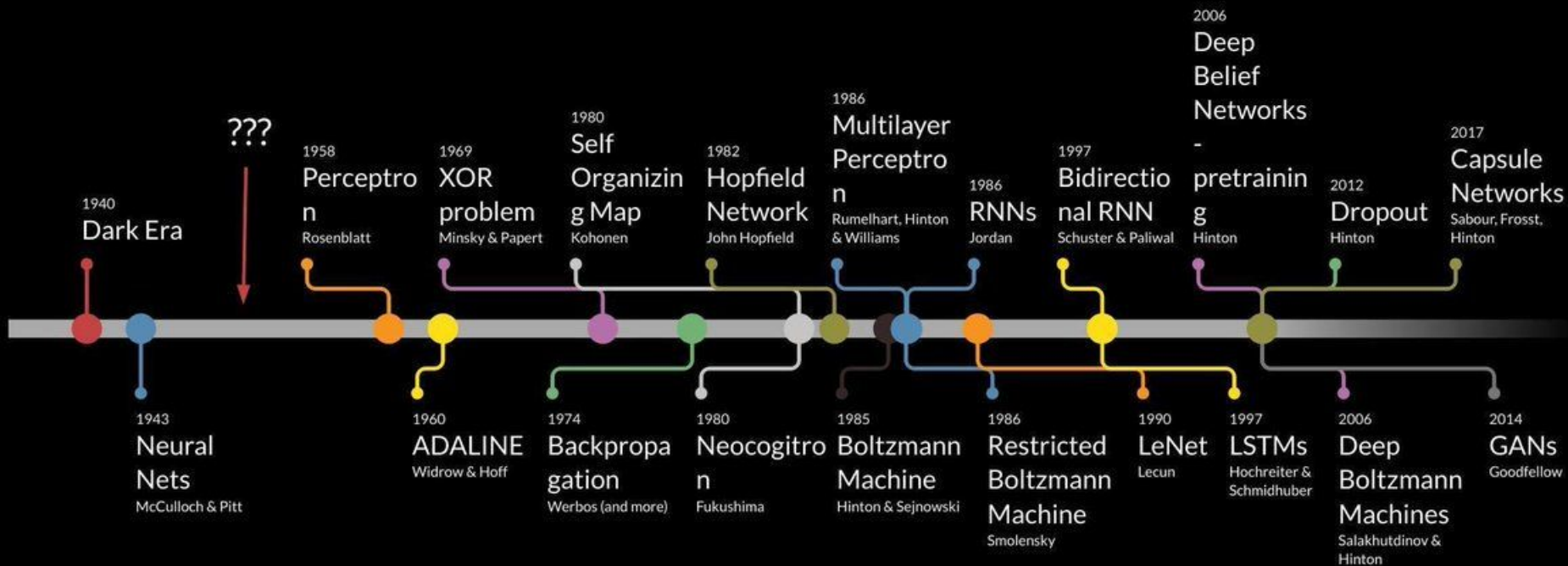
AlexNet, 2012



LeNet-5, 1998



Deep Learning Timeline



GANs, 2014+



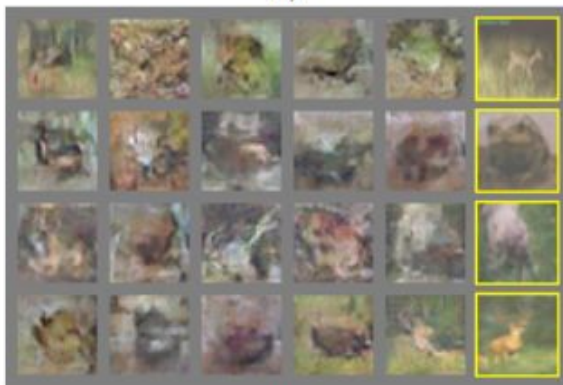
a)



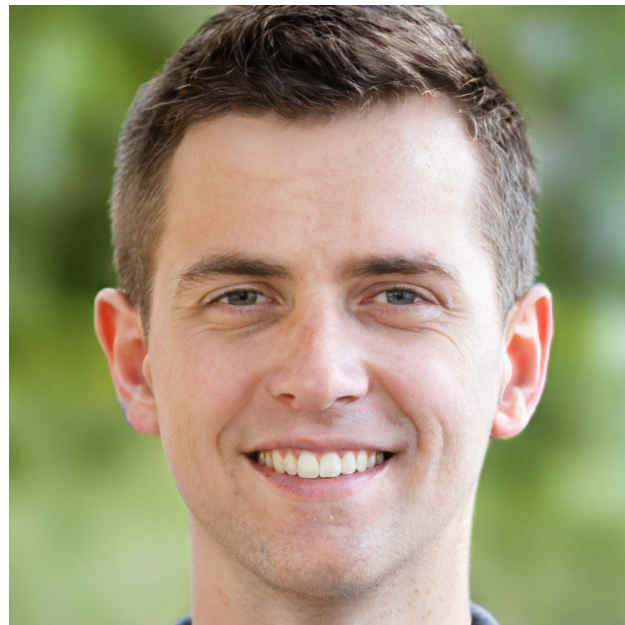
b)



c)

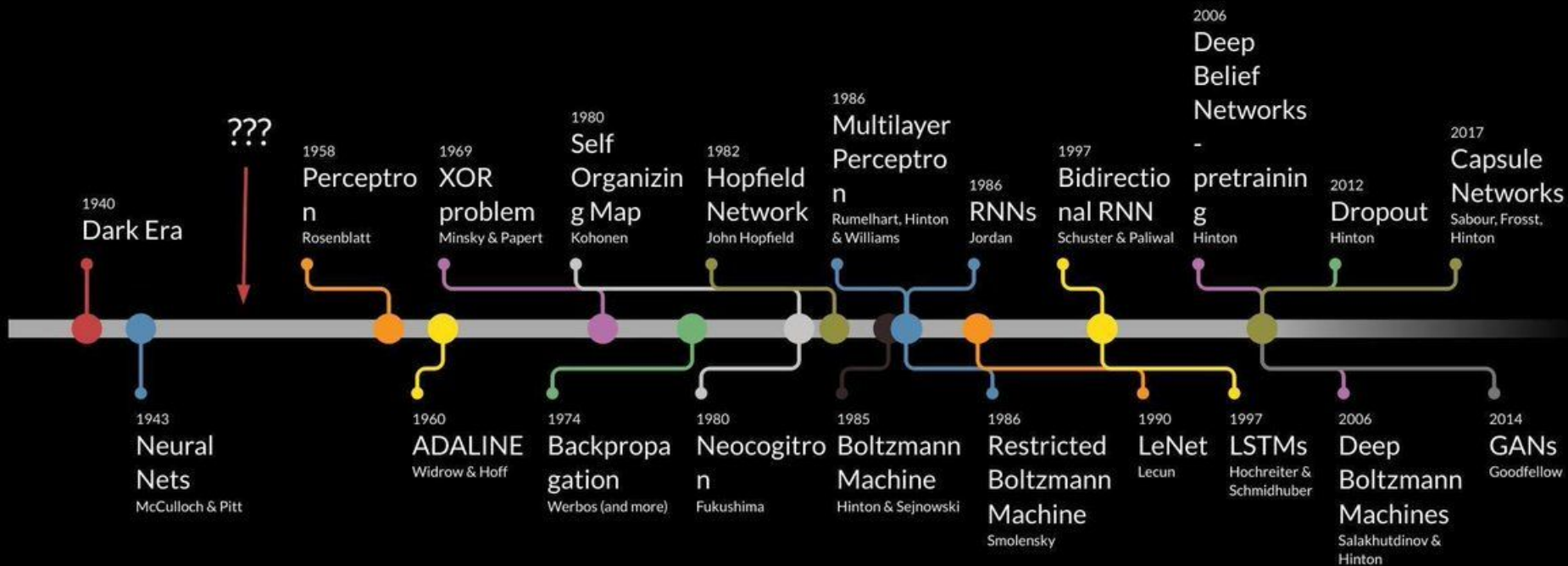


d)

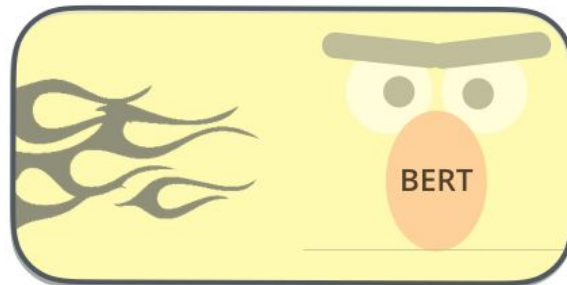
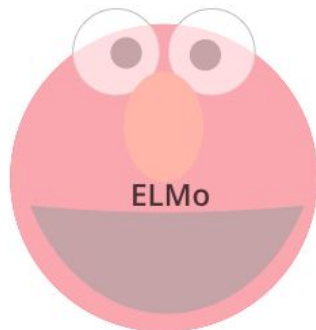
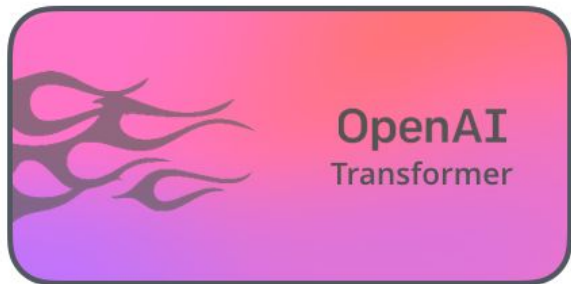


<https://thispersondoesnotexist.com/>

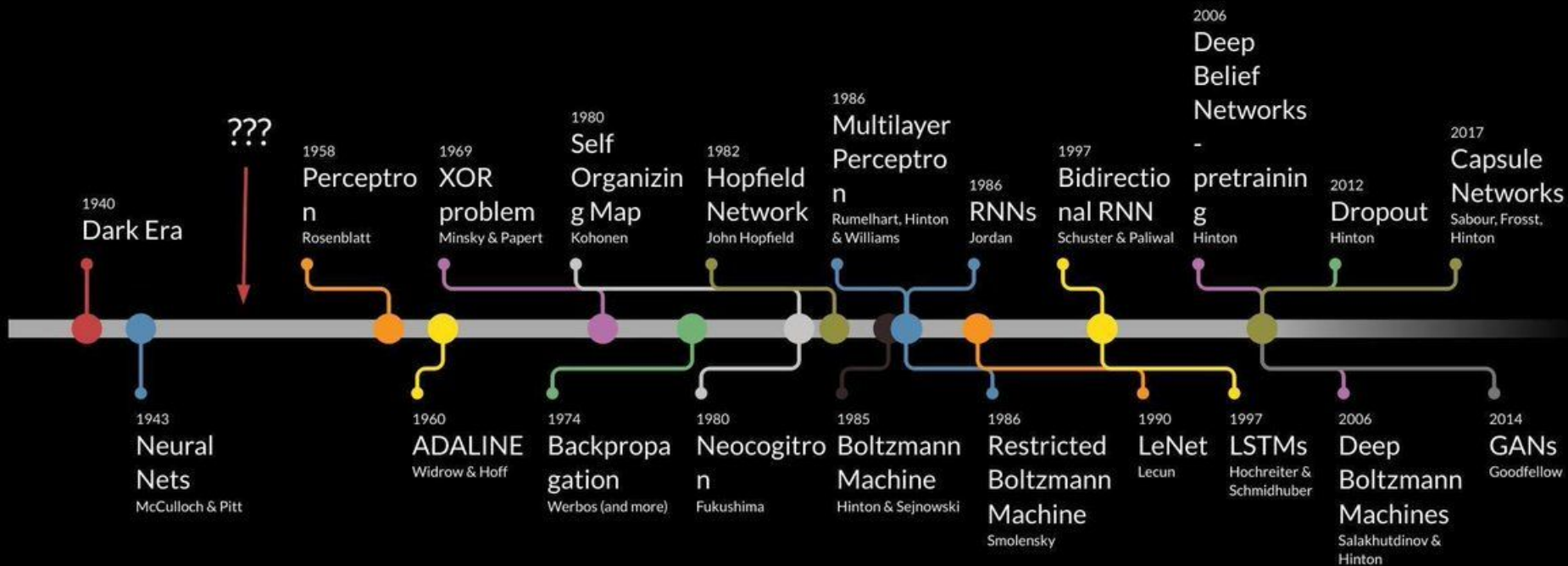
Deep Learning Timeline



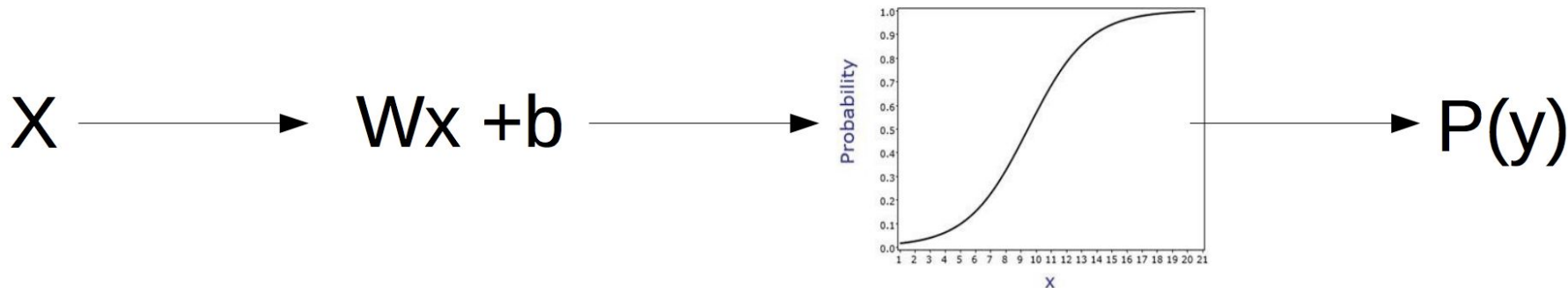
Transformer, BERT, GPT-2 and more, 2017+



Deep Learning Timeline



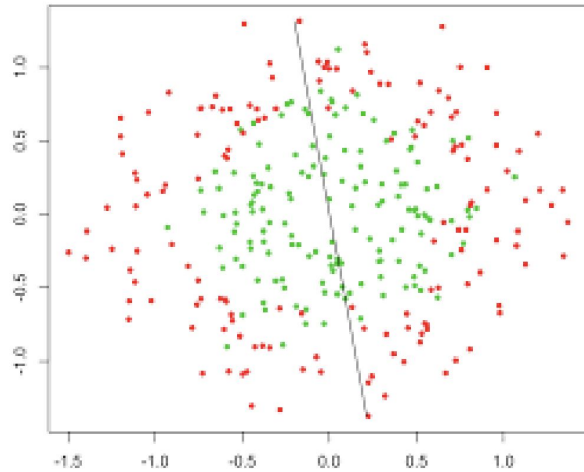
Logistic regression



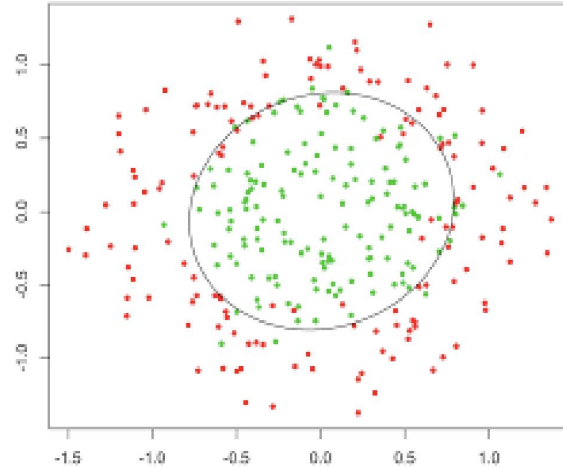
$$P(y|x) = \sigma(w \cdot x + b)$$

$$L = - \sum_i y_i \log P(y|x_i) + (1 - y_i) \log (1 - P(y|x_i))$$

Problem: nonlinear dependencies



What we have

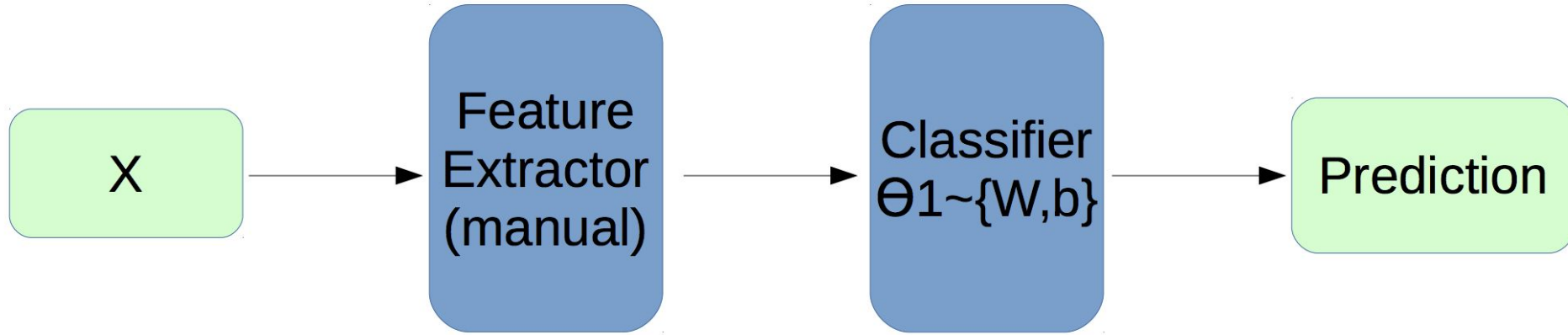


What we want

Logistic regression
(generally, linear model)
need feature engineering
to show good results.

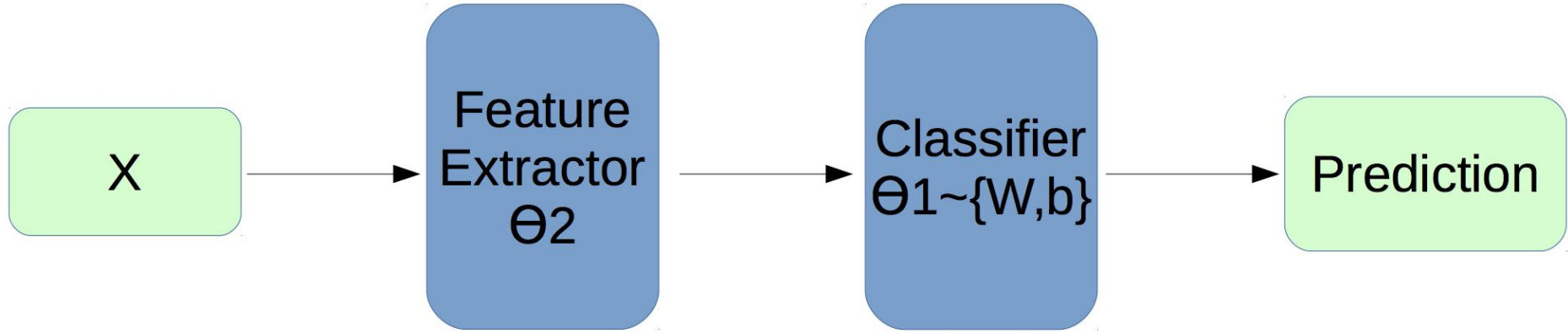
And feature engineering is
an art.

Classic pipeline



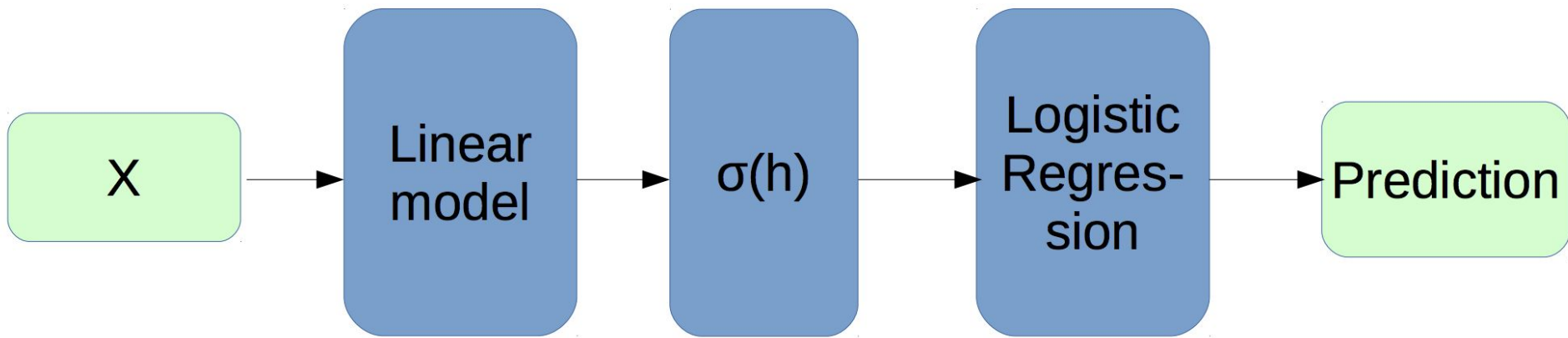
Handcrafted features, generated by experts.

NN pipeline



Automatically extracted features.

NN pipeline: example



E.g. two logistic regressions one after another.

Actually, it's a neural network.

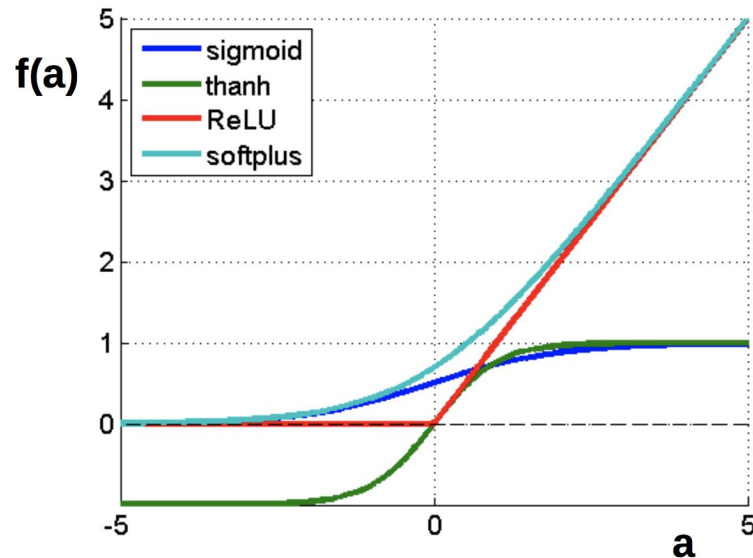
Activation functions: nonlinearities

$$f(a) = \frac{1}{1 + e^a}$$

$$f(a) = \tanh(a)$$

$$f(a) = \max(0, a)$$

$$f(a) = \log(1 + e^a)$$



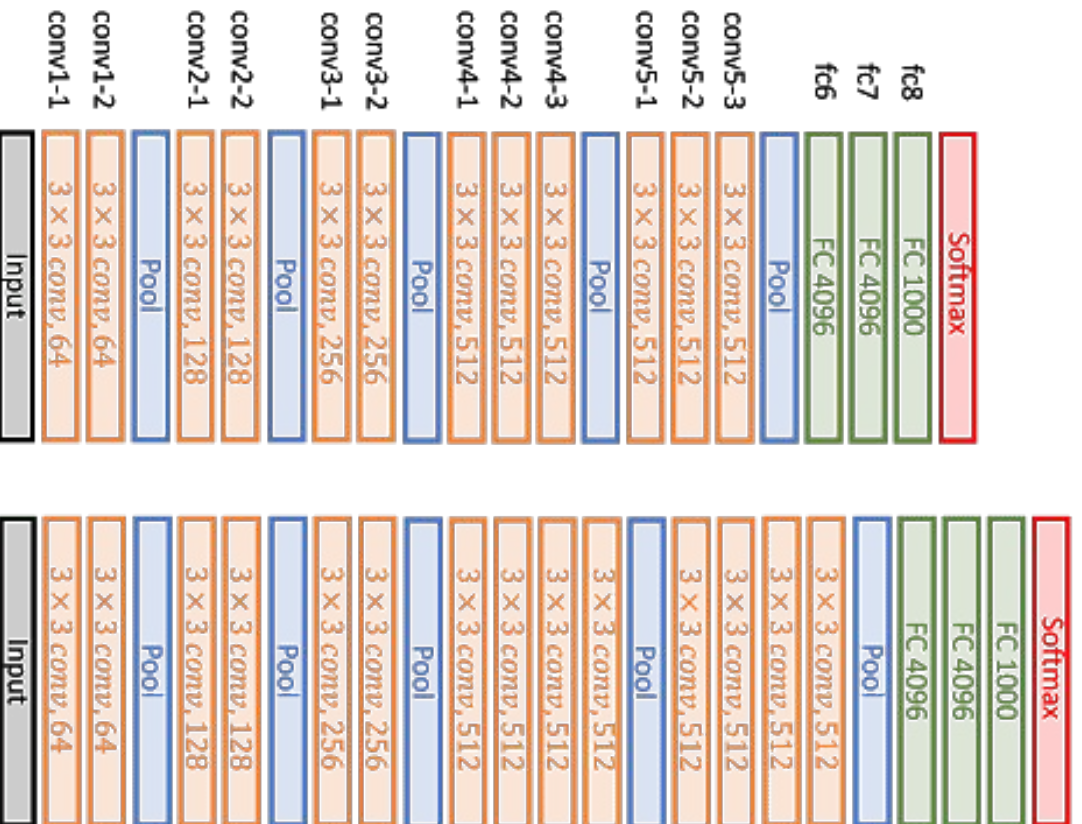
Some generally accepted terms

- Layer – a building block for NNs :
 - Dense/Linear/FC layer: $f(x) = Wx+b$
 - Nonlinearity layer: $f(x) = \sigma(x)$
 - Input layer, output layer
 - A few more we will cover later
- Activation function – function applied to layer output
 - Sigmoid
 - tanh
 - ReLU
 - Any other function to get nonlinear intermediate signal in NN
- Backpropagation – a fancy word for “chain rule”

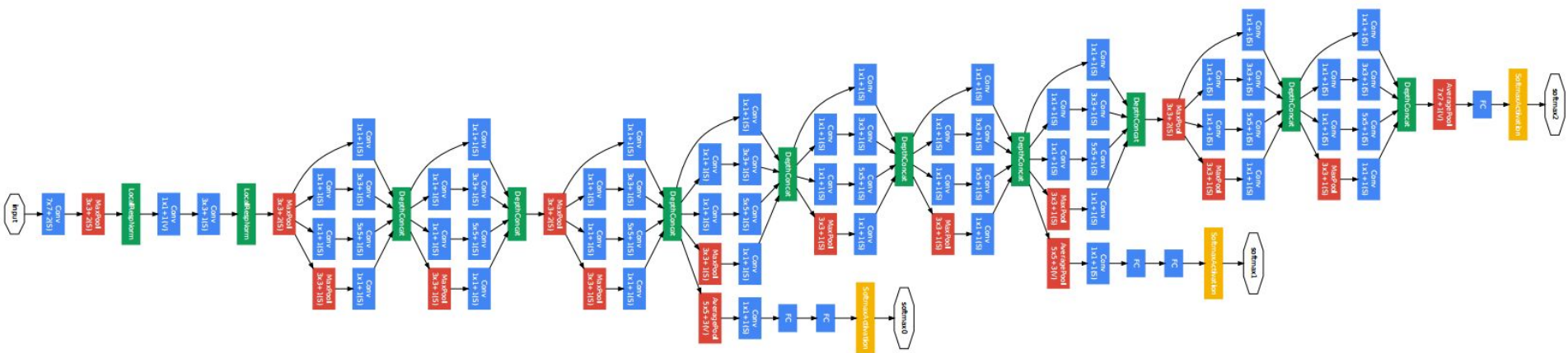
Actually, networks can be deep



And deeper...



Much deeper...

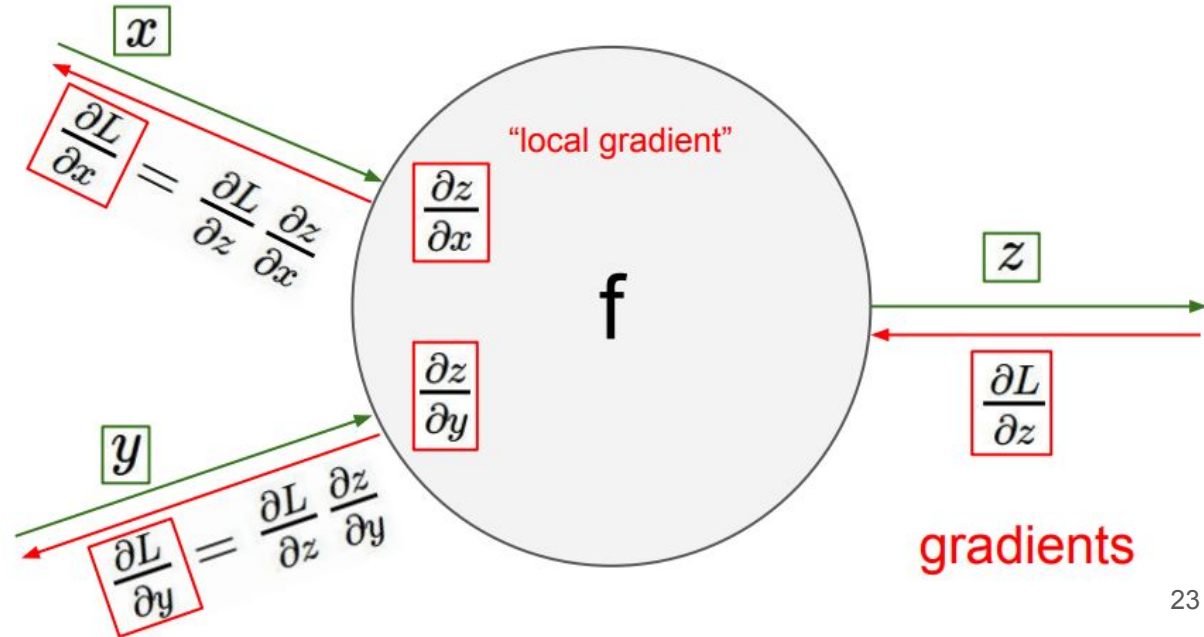


How to train it?

Backpropagation and chain rule

Chain rule is just simple math: $\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z} \frac{\partial z}{\partial x}$

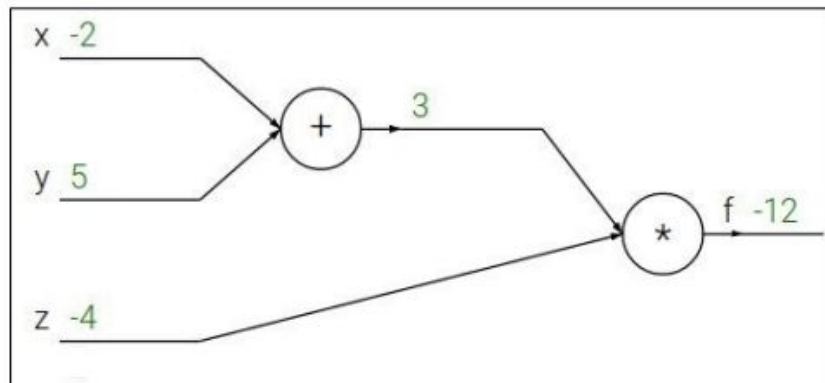
Backprop is just way to use it in NN training.



Backpropagation example

$$f(x, y, z) = (x + y)z$$

e.g. $x = -2, y = 5, z = -4$



Backpropagation example

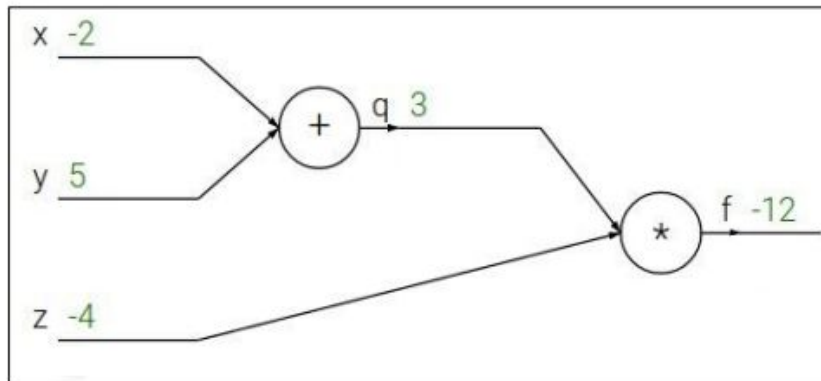
$$f(x, y, z) = (x + y)z$$

e.g. $x = -2, y = 5, z = -4$

$$q = x + y \quad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \quad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Backpropagation example

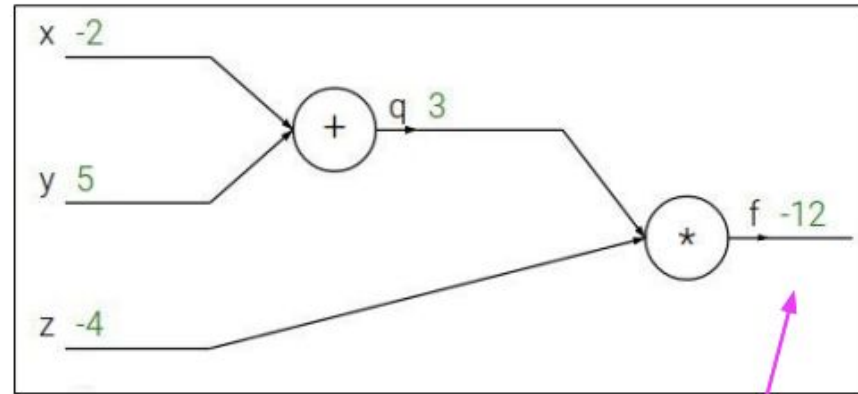
$$f(x, y, z) = (x + y)z$$

e.g. $x = -2, y = 5, z = -4$

$$q = x + y \quad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \quad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$$\frac{\partial f}{\partial f}$$

Backpropagation example

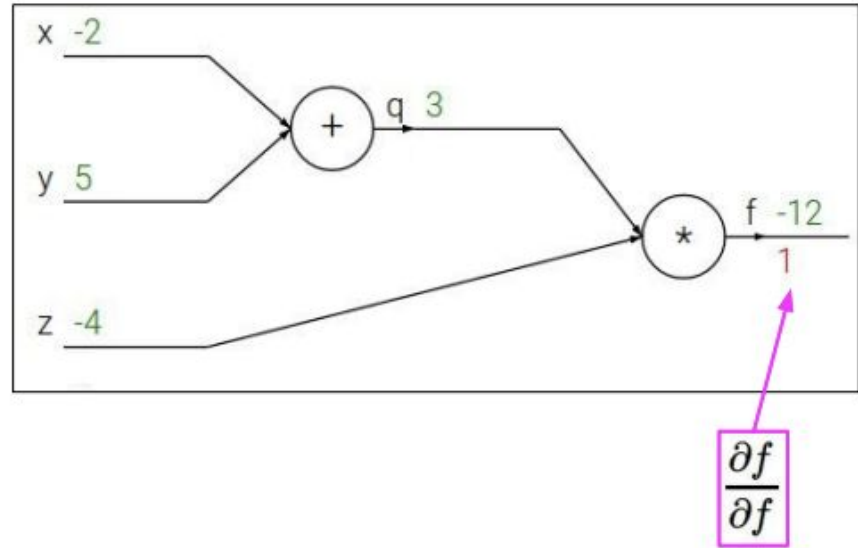
$$f(x, y, z) = (x + y)z$$

e.g. $x = -2, y = 5, z = -4$

$$q = x + y \quad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \quad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Backpropagation example

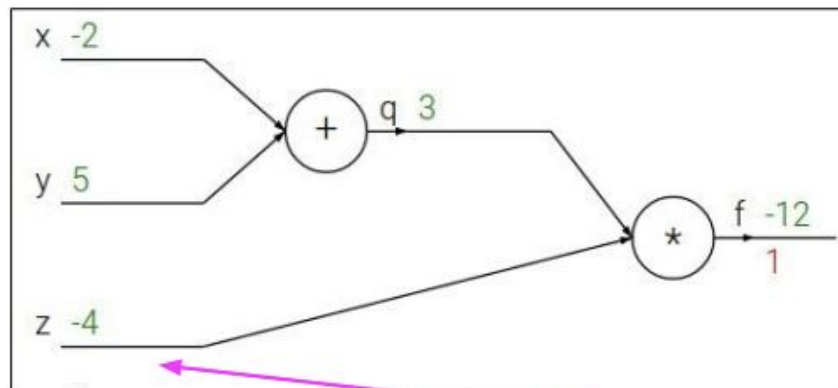
$$f(x, y, z) = (x + y)z$$

e.g. $x = -2, y = 5, z = -4$

$$q = x + y \quad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \quad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$$\frac{\partial f}{\partial z}$$

Backpropagation example

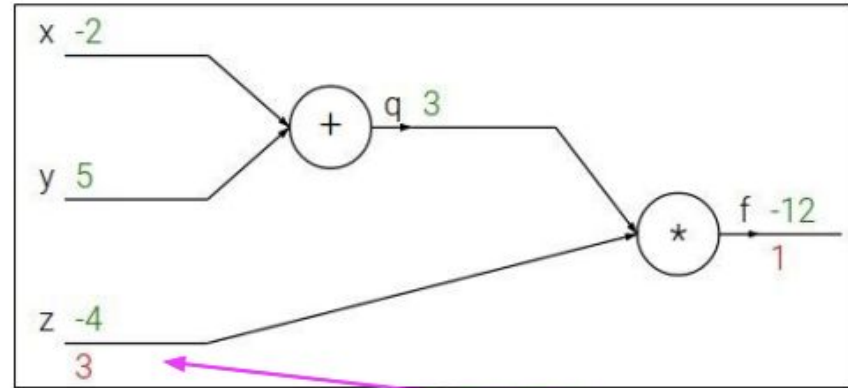
$$f(x, y, z) = (x + y)z$$

e.g. $x = -2, y = 5, z = -4$

$$q = x + y \quad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \quad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$$\frac{\partial f}{\partial z}$$

Backpropagation example

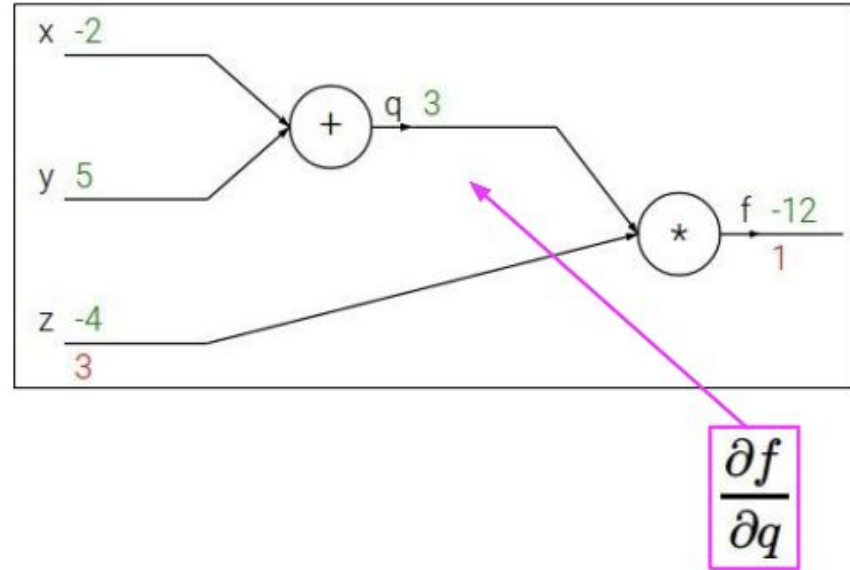
$$f(x, y, z) = (x + y)z$$

e.g. $x = -2, y = 5, z = -4$

$$q = x + y \quad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \quad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Backpropagation example

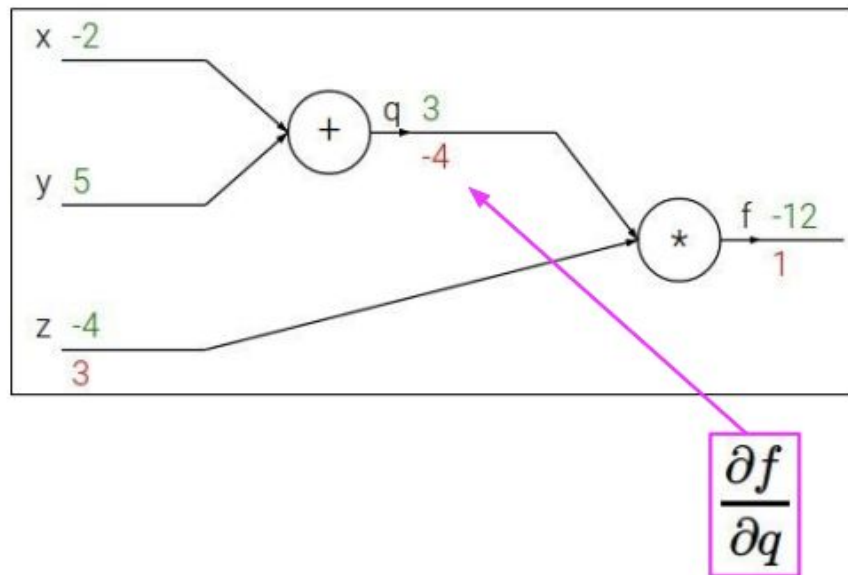
$$f(x, y, z) = (x + y)z$$

e.g. $x = -2, y = 5, z = -4$

$$q = x + y \quad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \quad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Backpropagation example

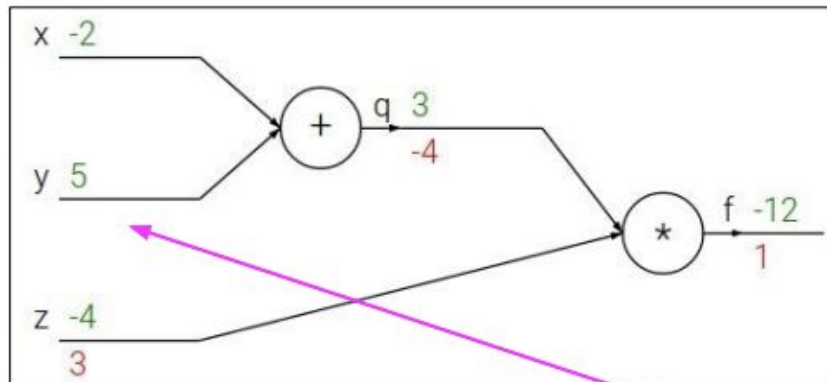
$$f(x, y, z) = (x + y)z$$

e.g. $x = -2$, $y = 5$, $z = -4$

$$q = x + y \quad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \quad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, $\frac{\partial f}{\partial z}$



$$\frac{\partial f}{\partial y}$$

Backpropagation example

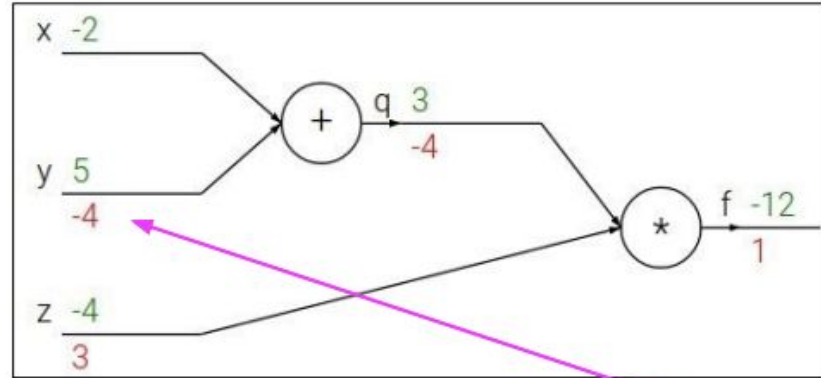
$$f(x, y, z) = (x + y)z$$

e.g. $x = -2, y = 5, z = -4$

$$q = x + y \quad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \quad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Chain rule:

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial y}$$

$$\frac{\partial f}{\partial y}$$

Backpropagation example

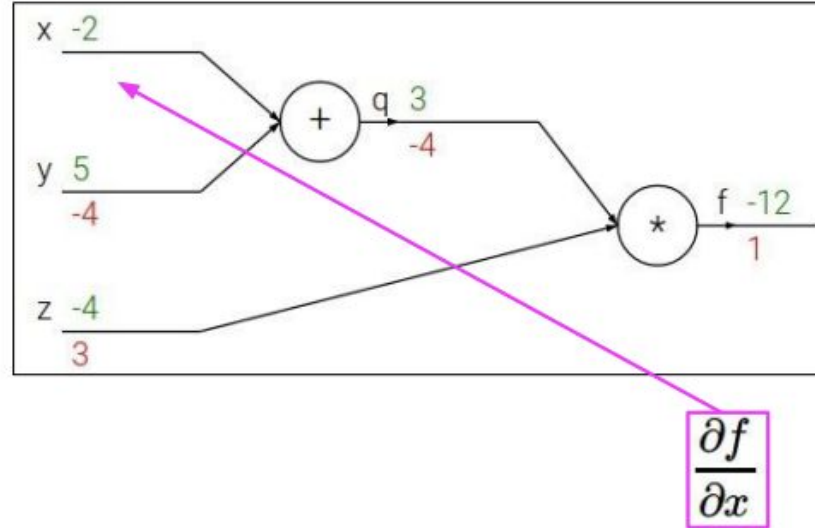
$$f(x, y, z) = (x + y)z$$

e.g. $x = -2, y = 5, z = -4$

$$q = x + y \quad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \quad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Backpropagation example

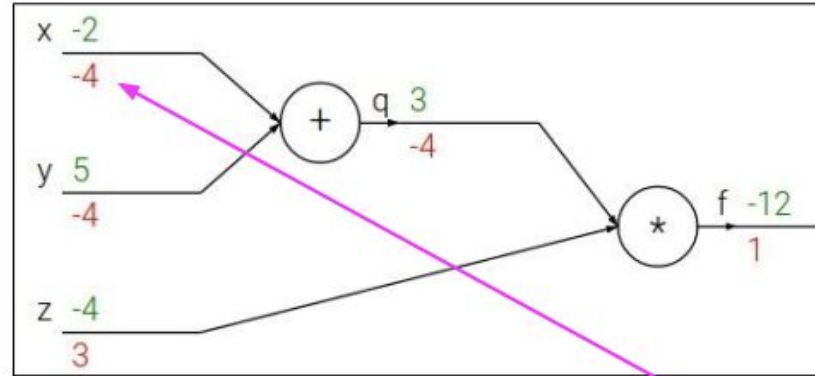
$$f(x, y, z) = (x + y)z$$

e.g. $x = -2, y = 5, z = -4$

$$q = x + y \quad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \quad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Chain rule:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial x}$$

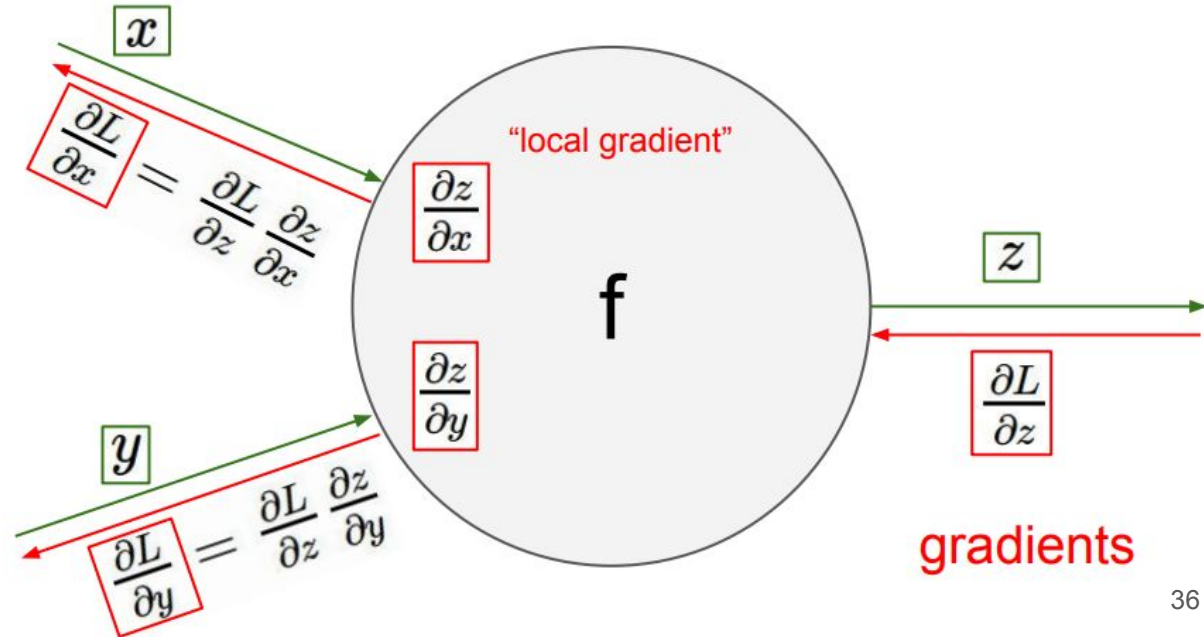
$$\frac{\partial f}{\partial x}$$

Backpropagation and chain rule

Chain rule is just simple math:

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z} \frac{\partial z}{\partial x}$$

Backprop is just way to use it in NN training.



Interactive playground

DATA

Which dataset do you want to use?



Ratio of training to test data: 50%



Noise: 0



Batch size: 10



REGENERATE

FEATURES

Which properties do you want to feed in?



+ - 1 HIDDEN LAYER

+ -

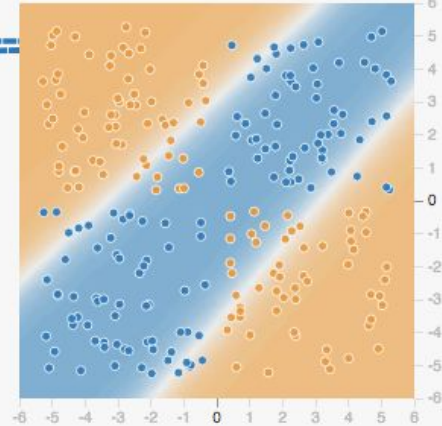
2 neurons

This is the output from one **neuron**.
Hover to see it larger.

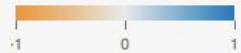
OUTPUT

Test loss 0.208

Training loss 0.207



Colors shows data, neuron and weight values.

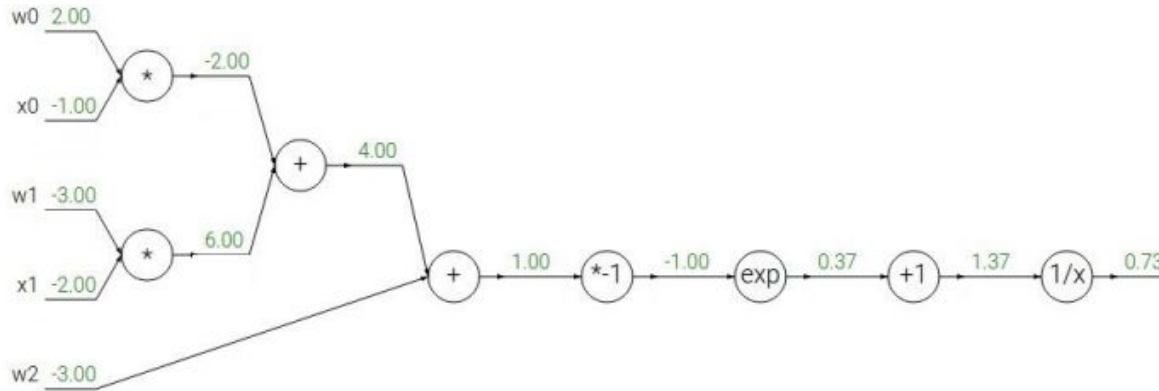


☐ Show test data ☐ Discretize output

<https://playground.tensorflow.org/>

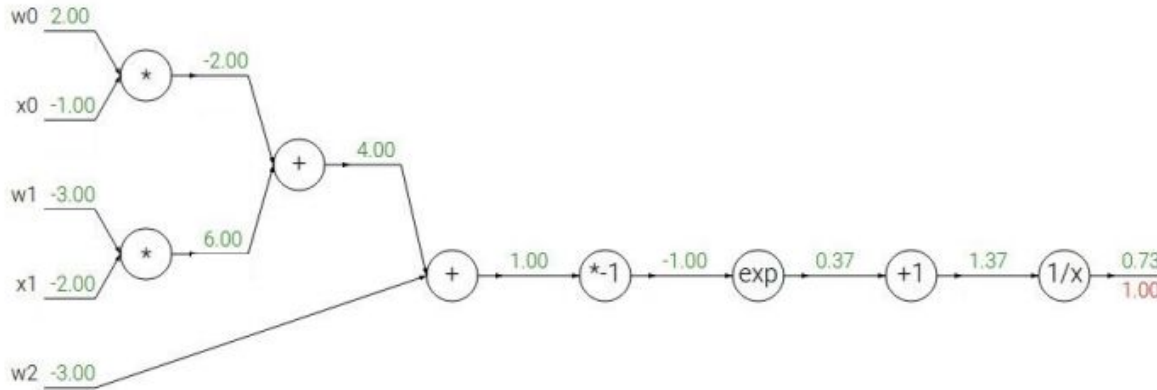
Backpropagation example

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



Backpropagation example

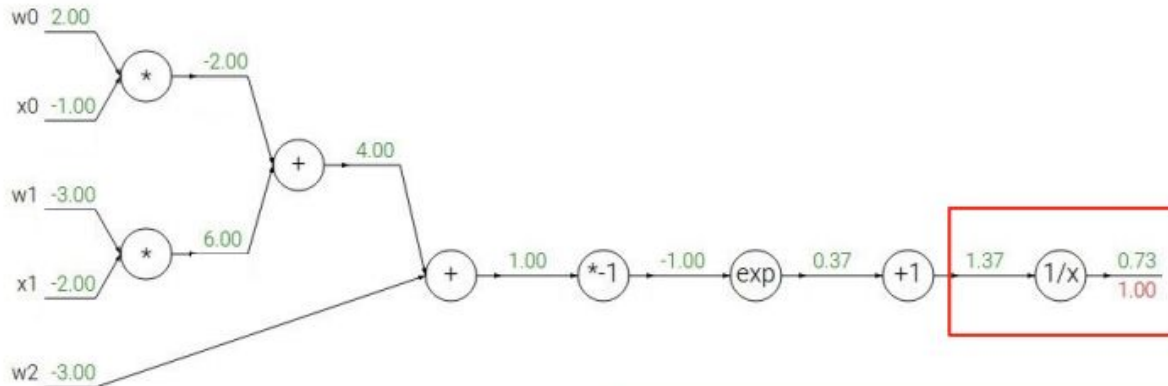
Another example: $f(w, x) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2x_2)}}$



$f(x) = e^x$	\rightarrow	$\frac{df}{dx} = e^x$		$f(x) = \frac{1}{x}$	\rightarrow	$\frac{df}{dx} = -1/x^2$
$f_a(x) = ax$	\rightarrow	$\frac{df}{dx} = a$		$f_c(x) = c + x$	\rightarrow	$\frac{df}{dx} = 1$

Backpropagation example

Another example: $f(w, x) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2x_2)}}$



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

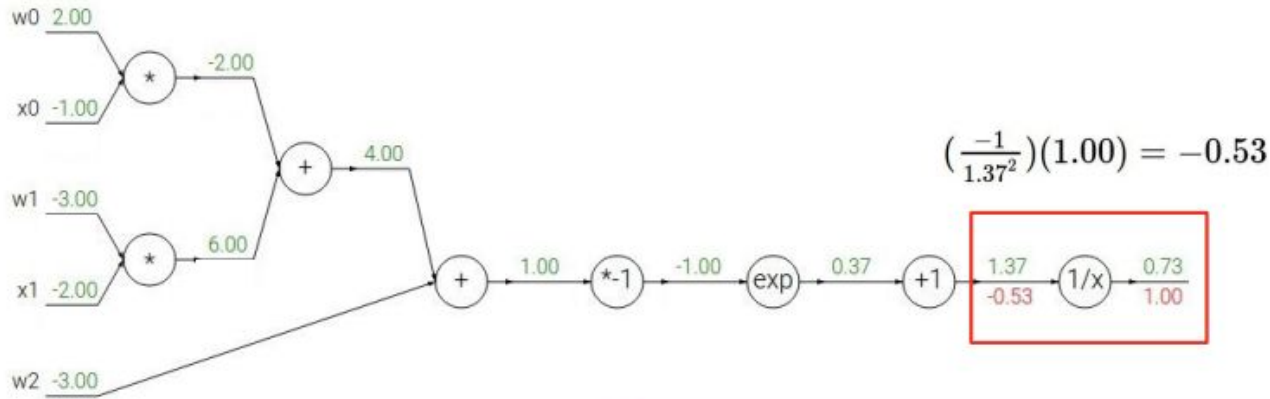
$$f_a(x) = ax \rightarrow \frac{df}{dx} = a$$

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

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

Backpropagation example

Another example: $f(w, x) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2x_2)}}$

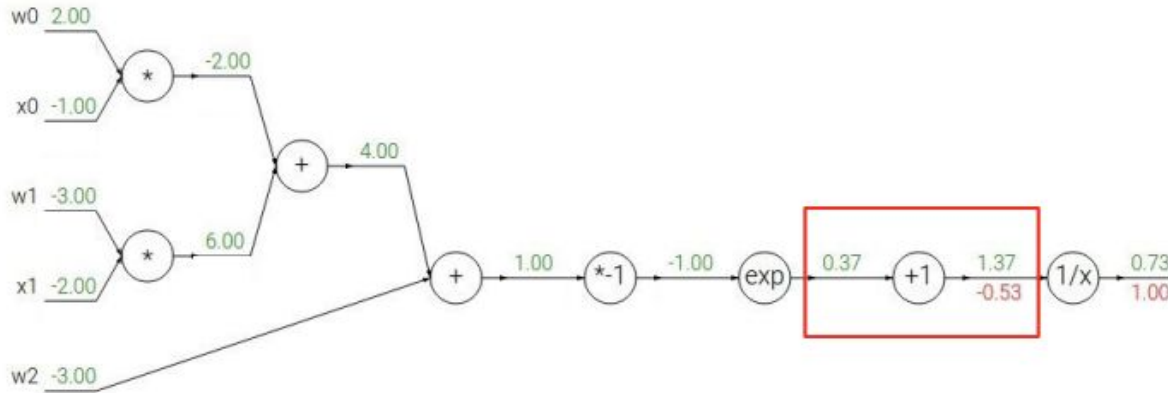


$$\begin{aligned} f(x) = e^x &\rightarrow \frac{df}{dx} = e^x \\ f_a(x) = ax &\rightarrow \frac{df}{dx} = a \end{aligned}$$

$$\begin{aligned} f(x) = \frac{1}{x} &\rightarrow \frac{df}{dx} = -1/x^2 \\ f_c(x) = c + x &\rightarrow \frac{df}{dx} = 1 \end{aligned}$$

Backpropagation example

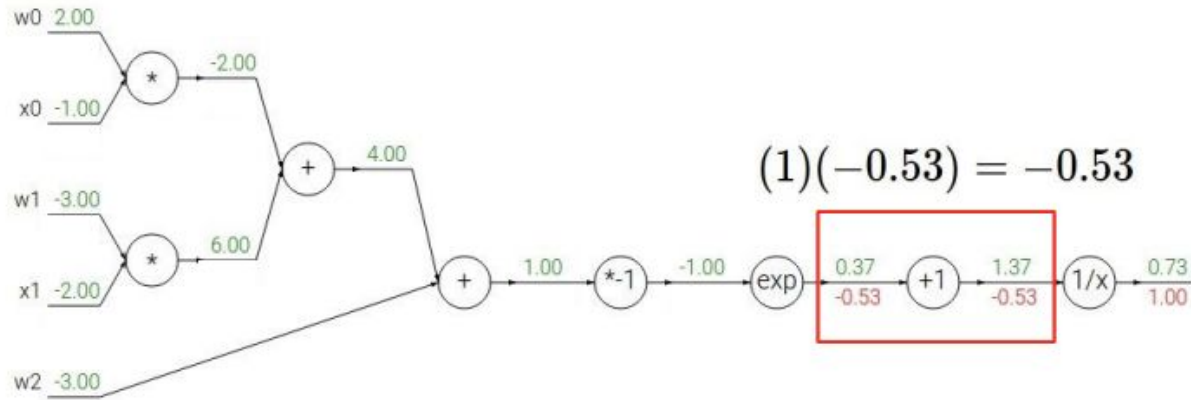
Another example: $f(w, x) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2x_2)}}$



$f(x) = e^x$	\rightarrow	$\frac{df}{dx} = e^x$		$f(x) = \frac{1}{x}$	\rightarrow	$\frac{df}{dx} = -1/x^2$
$f_a(x) = ax$	\rightarrow	$\frac{df}{dx} = a$		<div style="border: 1px solid red; padding: 5px;">$f_c(x) = c + x$</div>	\rightarrow	<div style="border: 1px solid red; padding: 5px;">$\frac{df}{dx} = 1$</div>

Backpropagation example

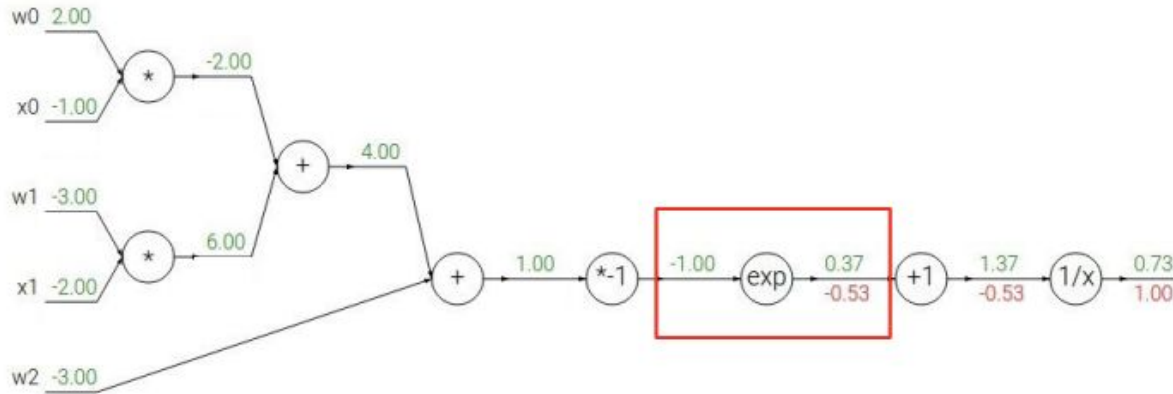
Another example: $f(w, x) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}}$



$f(x) = e^x$	→	$\frac{df}{dx} = e^x$		$f(x) = \frac{1}{x}$	→	$\frac{df}{dx} = -1/x^2$
$f_a(x) = ax$	→	$\frac{df}{dx} = a$		$f_c(x) = c + x$	→	$\frac{df}{dx} = 1$

Backpropagation example

Another example: $f(w, x) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2x_2)}}$

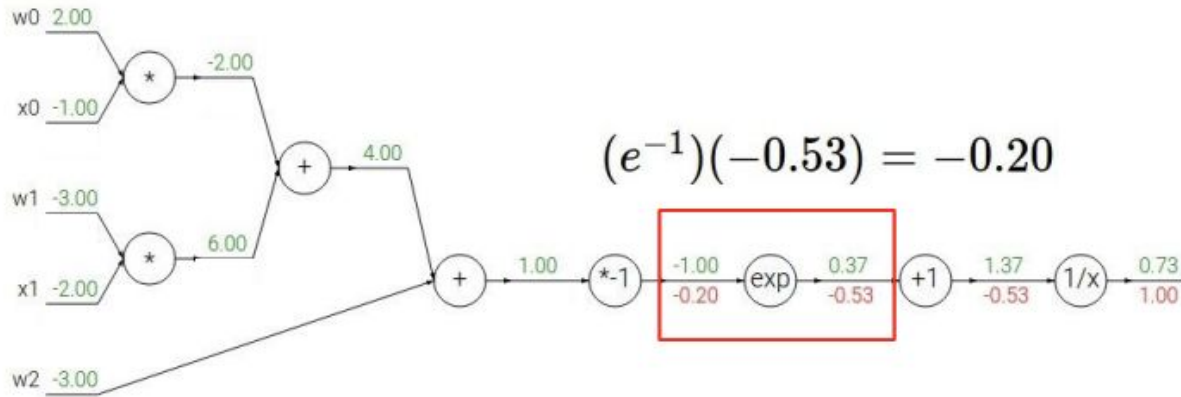


$f(x) = e^x$	\rightarrow	$\frac{df}{dx} = e^x$
$f_a(x) = ax$	\rightarrow	$\frac{df}{dx} = a$

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

Backpropagation example

Another example: $f(w, x) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2x_2)}}$



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

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

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

$$f(x) = \frac{1}{x}$$

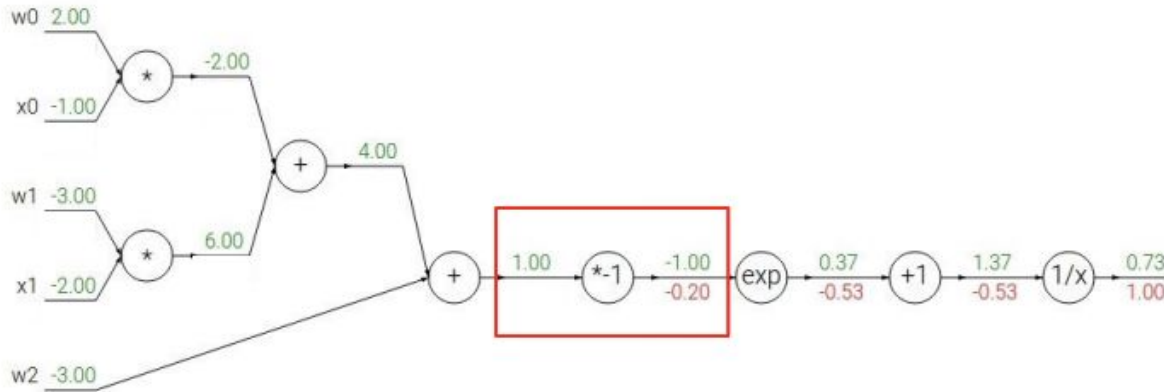
$$f_c(x) = c + x$$

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

$$\rightarrow \frac{df}{dx} = 1$$

Backpropagation example

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



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

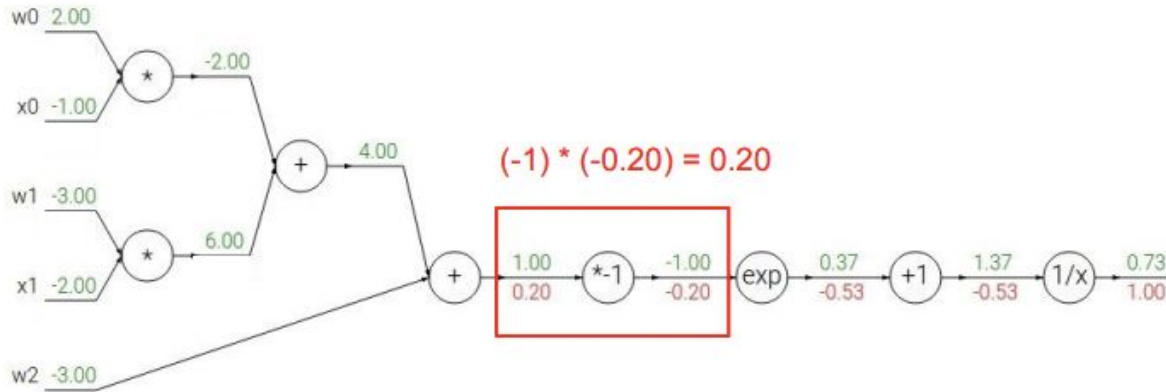
$$f_a(x) = ax \rightarrow \frac{df}{dx} = a$$

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

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

Backpropagation example

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



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

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

$$f(x) = \frac{1}{x}$$

$$f_c(x) = c + x$$

\rightarrow

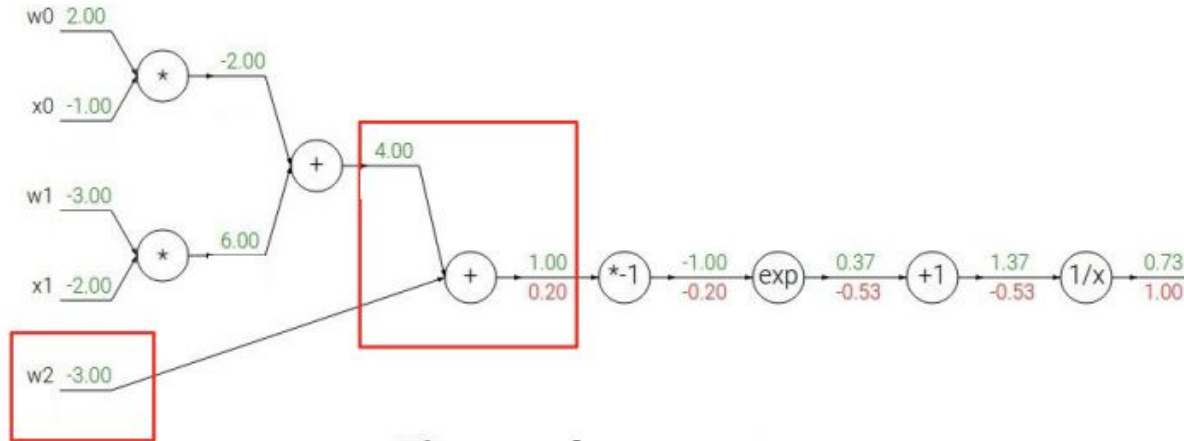
$$\frac{df}{dx} = -1/x^2$$

\rightarrow

$$\frac{df}{dx} = 1$$

Backpropagation example

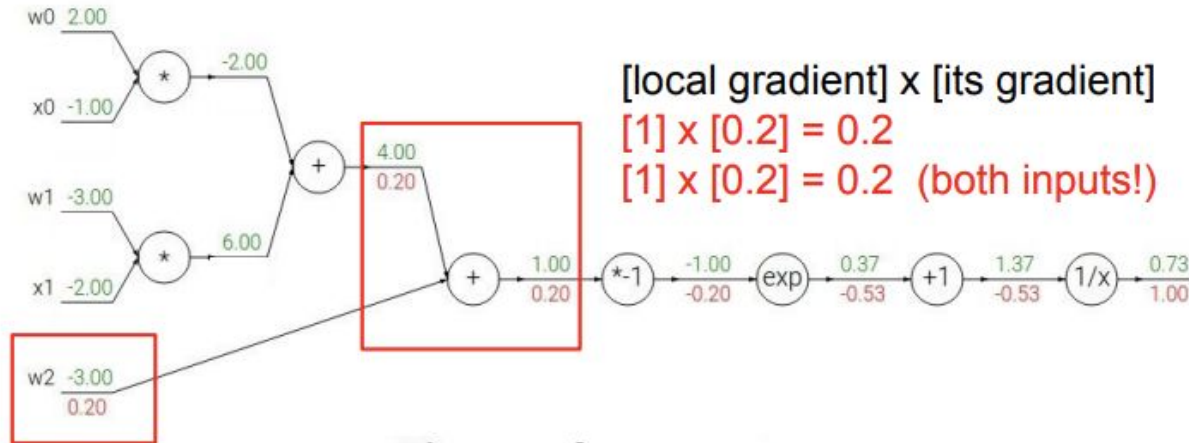
Another example: $f(w, x) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2x_2)}}$



$$\begin{array}{lcl}
 f(x) = e^x & \rightarrow & \frac{df}{dx} = e^x \\
 f_a(x) = ax & \rightarrow & \frac{df}{dx} = a
 \end{array}
 \quad \Bigg| \quad
 \begin{array}{lcl}
 f(x) = \frac{1}{x} & \rightarrow & \frac{df}{dx} = -1/x^2 \\
 f_c(x) = c + x & \rightarrow & \frac{df}{dx} = 1
 \end{array}$$

Backpropagation example

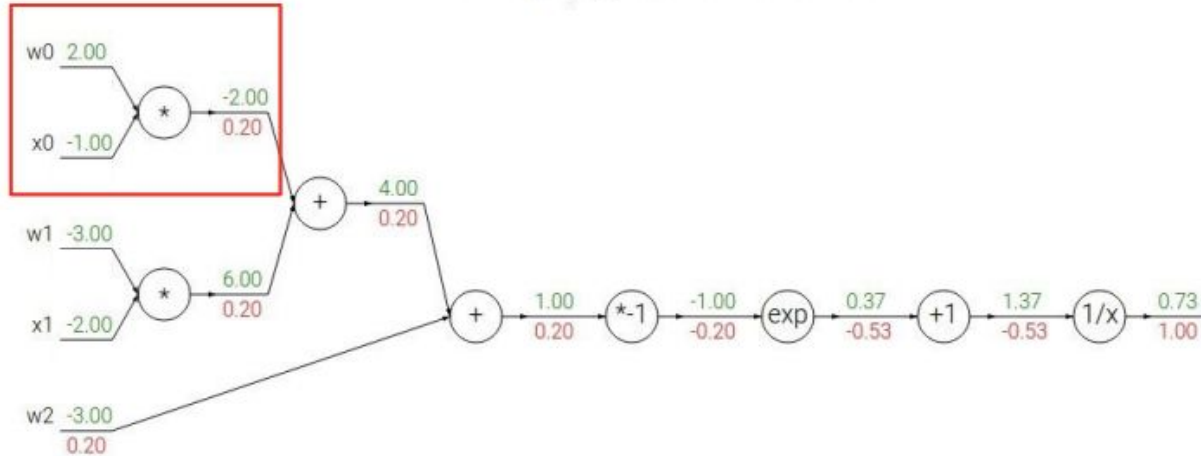
Another example: $f(w, x) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}}$



$$\begin{array}{lcl}
 f(x) = e^x & \rightarrow & \frac{df}{dx} = e^x \\
 f_a(x) = ax & \rightarrow & \frac{df}{dx} = a
 \end{array}
 \quad \Bigg| \quad
 \begin{array}{lcl}
 f(x) = \frac{1}{x} & \rightarrow & \frac{df}{dx} = -1/x^2 \\
 f_c(x) = c + x & \rightarrow & \frac{df}{dx} = 1
 \end{array}$$

Backpropagation example

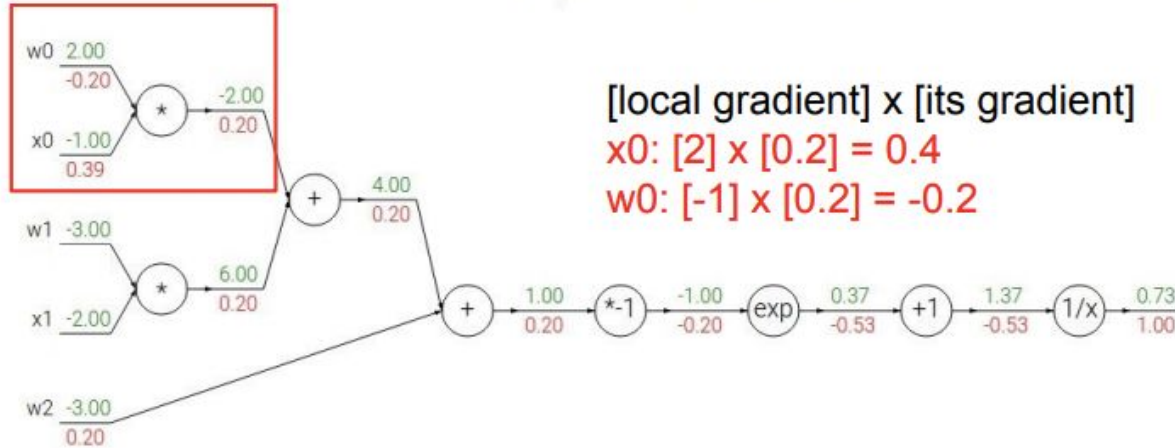
Another example: $f(w, x) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2x_2)}}$



$$\begin{array}{lcl}
 f(x) = e^x & \rightarrow & \frac{df}{dx} = e^x \\
 f_a(x) = ax & \rightarrow & \frac{df}{dx} = a
 \end{array}
 \quad \Bigg| \quad
 \begin{array}{lcl}
 f(x) = \frac{1}{x} & \rightarrow & \frac{df}{dx} = -1/x^2 \\
 f_c(x) = c + x & \rightarrow & \frac{df}{dx} = 1
 \end{array}$$

Backpropagation example

Another example: $f(w, x) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2x_2)}}$



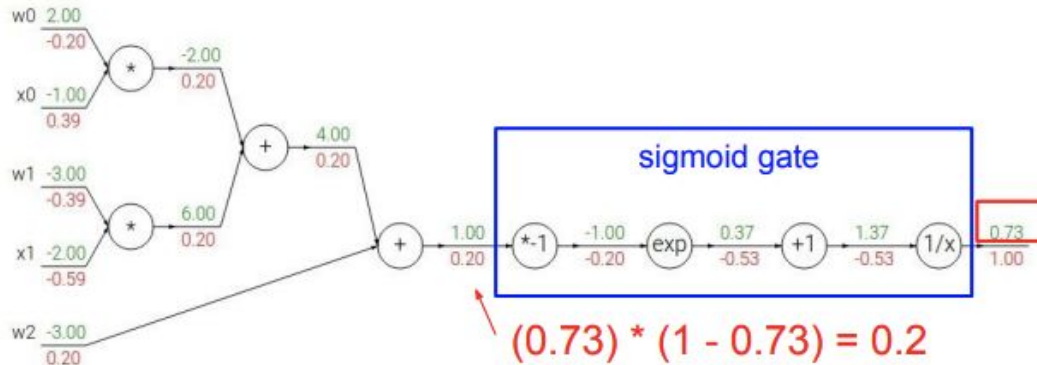
$f(x) = e^x$	\rightarrow	$\frac{df}{dx} = e^x$		$f(x) = \frac{1}{x}$	\rightarrow	$\frac{df}{dx} = -1/x^2$
$f_a(x) = ax$	\rightarrow	$\frac{df}{dx} = a$		$f_c(x) = c + x$	\rightarrow	$\frac{df}{dx} = 1$

Backpropagation example

$$f(w, x) = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad \text{sigmoid function}$$

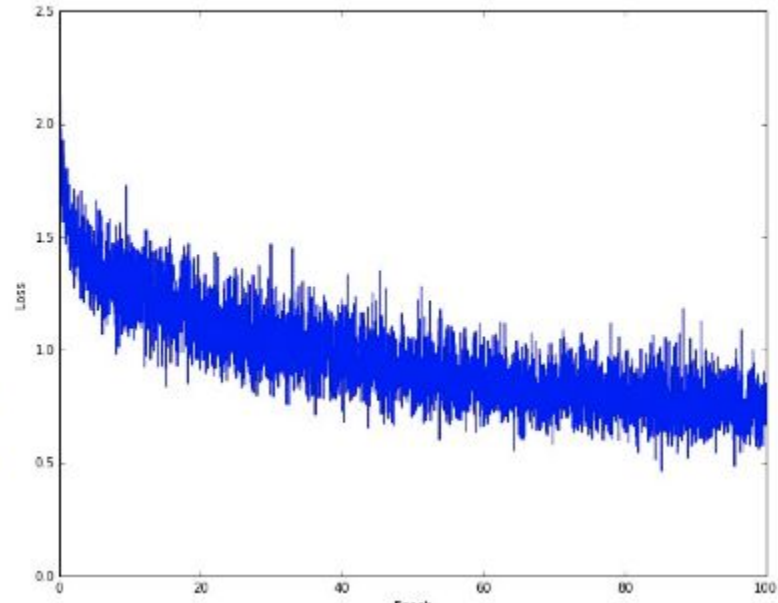
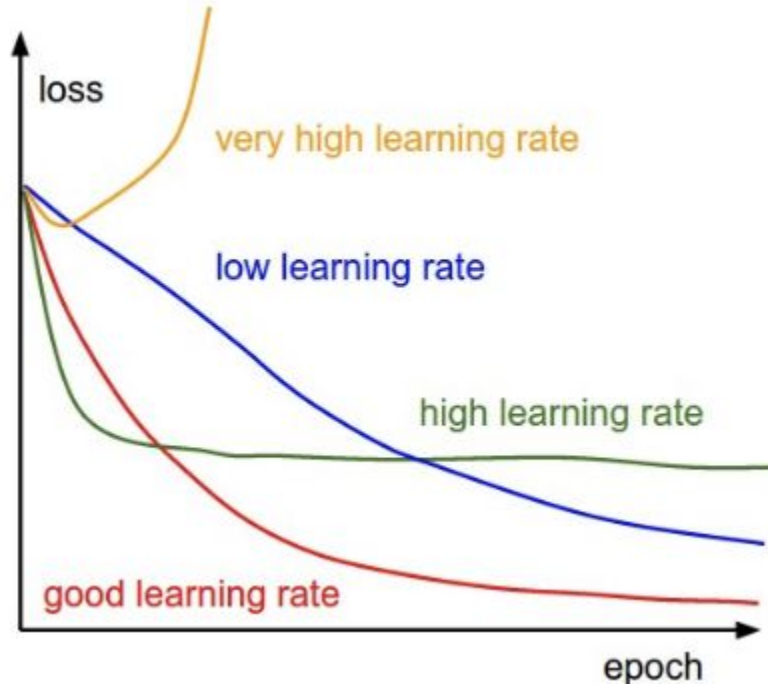
$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1 + e^{-x})^2} = \left(\frac{1 + e^{-x} - 1}{1 + e^{-x}} \right) \left(\frac{1}{1 + e^{-x}} \right) = (1 - \sigma(x)) \sigma(x)$$



Gradient optimization

Stochastic gradient descent (and variations)
is used to optimize NN parameters.

$$x_{t+1} = x_t - \text{learning rate} \cdot dx$$



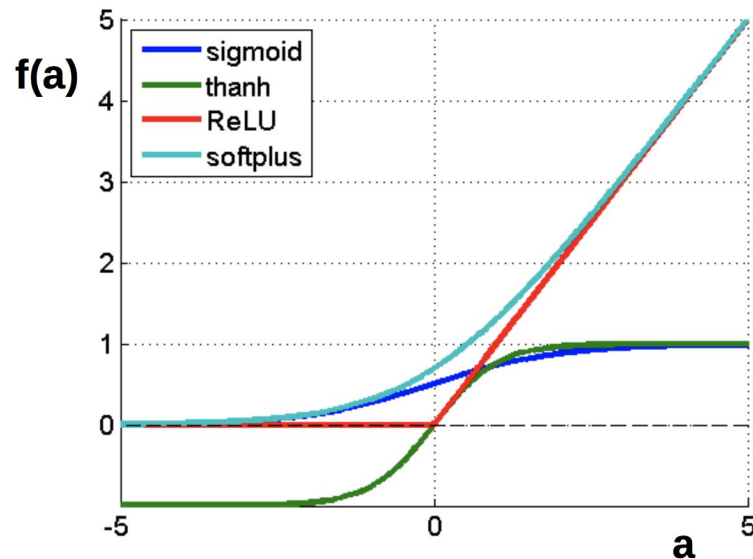
Once more: nonlinearities

$$f(a) = \frac{1}{1 + e^a}$$

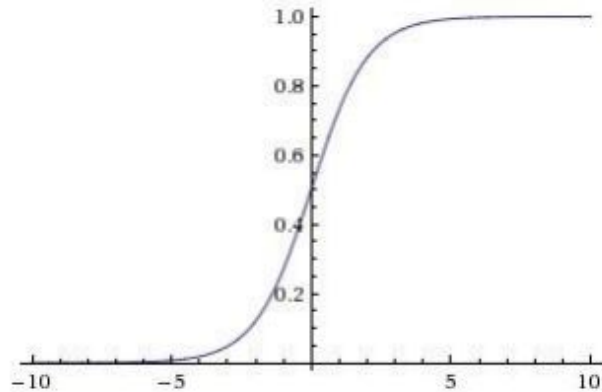
$$f(a) = \tanh(a)$$

$$f(a) = \max(0, a)$$

$$f(a) = \log(1 + e^a)$$



Activation functions



Sigmoid

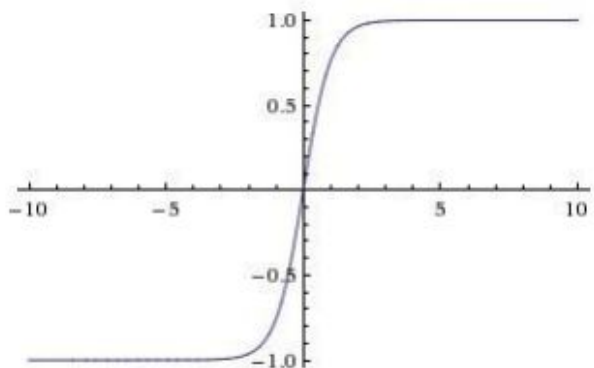
$$f(a) = \frac{1}{1 + e^a}$$

- Squashes numbers to range [0,1]
- Historically popular since they have nice interpretation as a saturating “firing rate” of a neuron

3 problems:

1. Saturated neurons “kill” the gradients
2. Sigmoid outputs are not zero-centered
3. $\exp()$ is a bit compute expensive

Activation functions

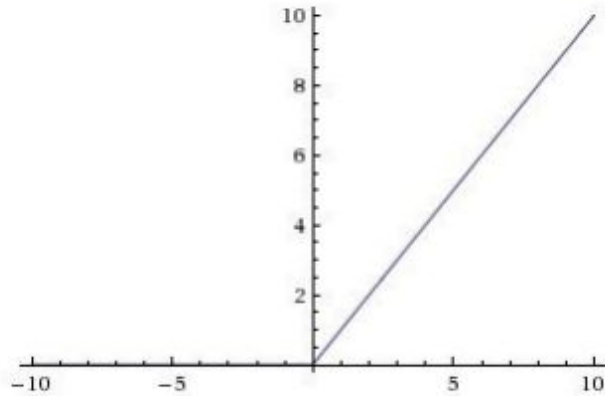


tanh(x)

- Squashes numbers to range $[-1,1]$
- zero centered (nice)
- still kills gradients when saturated :(

$$f(a) = \tanh(a)$$

Activation functions



ReLU

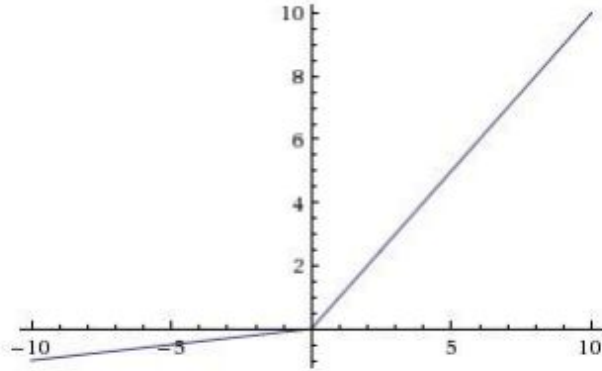
(Rectified Linear Unit)

$$f(a) = \max(0, a)$$

- Does not saturate (in +region)
- Very computationally efficient
- Converges much faster than sigmoid/tanh in practice (e.g. 6x)
- Not zero-centered output
- An annoyance:

hint: what is the gradient when $x < 0$?

Activation functions



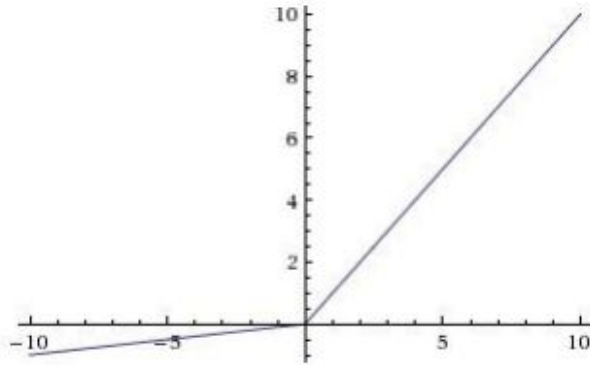
- Does not saturate
- Computationally efficient
- Converges much faster than sigmoid/tanh in practice! (e.g. 6x)
- **will not “die”.**

Leaky ReLU

$$f(x) = \max(0.01x, x)$$

Activation functions

- Does not saturate
- Computationally efficient
- Converges much faster than sigmoid/tanh in practice! (e.g. 6x)
- **will not “die”.**



Leaky ReLU

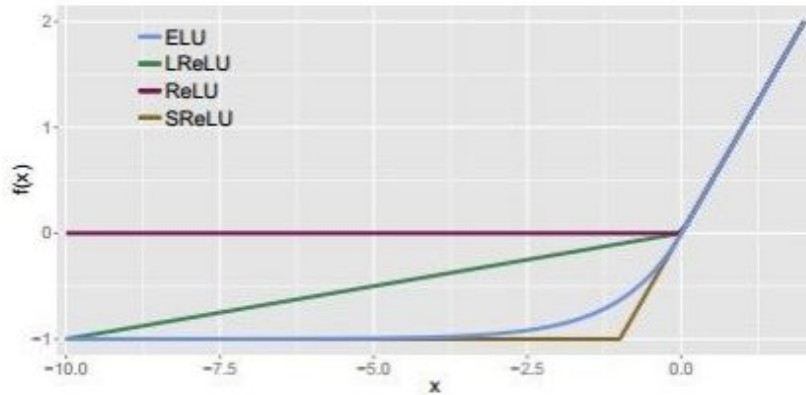
$$f(x) = \max(0.01x, x)$$

Parametric Rectifier (PReLU)

$$f(x) = \max(\alpha x, x)$$

backprop into α
(parameter)

Exponential Linear Units (ELU)



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

- All benefits of ReLU
- Does not die
- Closer to zero mean outputs
- Computation requires $\exp()$

Activation functions: sum up

- Use **ReLU** as baseline approach
- Be careful with the learning rates
- Try out **Leaky ReLU** or **ELU**
- Try out **tanh** but do not expect much from it
- Do not use **Sigmoid**

Don't miss the interactive playground

DATA

Which dataset do you want to use?



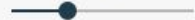
Ratio of training to test data: 50%



Noise: 0



Batch size: 10



REGENERATE

FEATURES

Which properties do you want to feed in?



+ - 1 HIDDEN LAYER

+ -

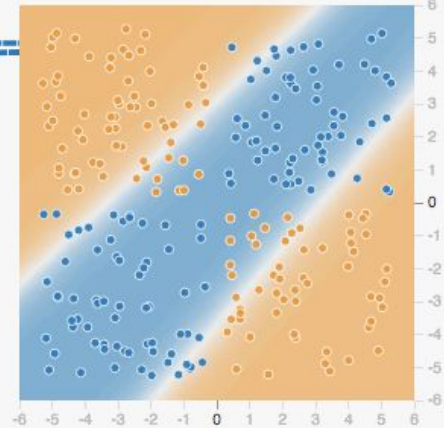
2 neurons

This is the output from one **neuron**.
Hover to see it larger.

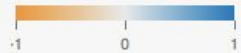
OUTPUT

Test loss 0.208

Training loss 0.207



Colors shows data, neuron and weight values.



☐ Show test data ☐ Discretize output

<https://playground.tensorflow.org/>



WHO'S AWESOME?
You're Awesome