Introduction to backpropagation

In this lesson

- Virginia Tech Hokies
- Matrices
- Gradient descent flavors (full, batch, sgd)
- Correlation
- Overfitting
- Backpropagation

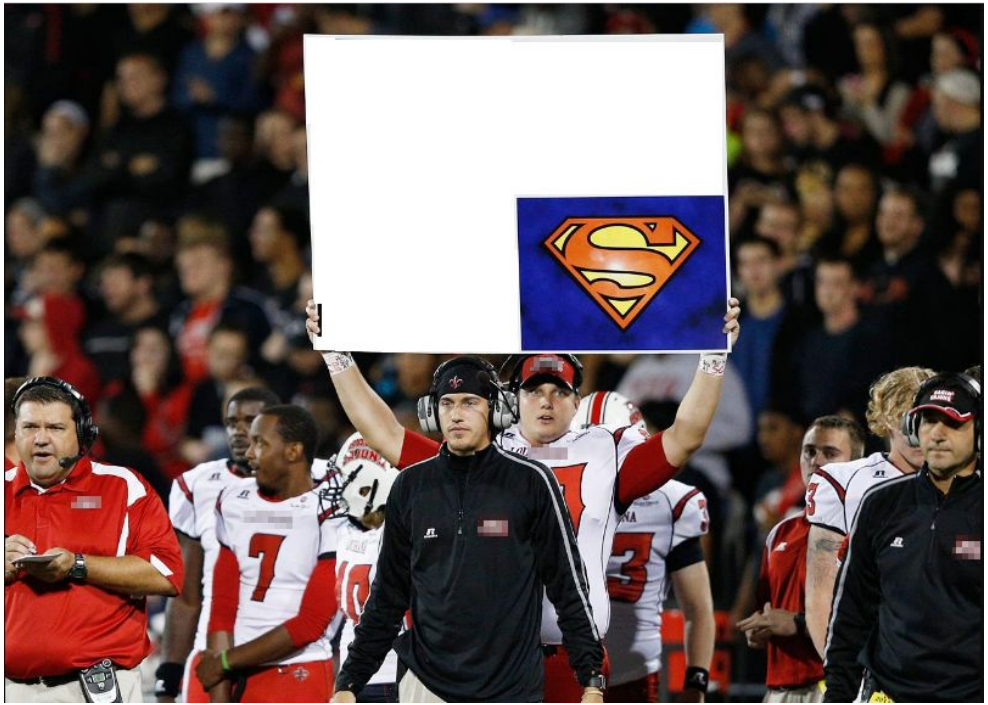
Hokies

Let’s learn how a network learns entire datasets

Run

PASS!!!

Run



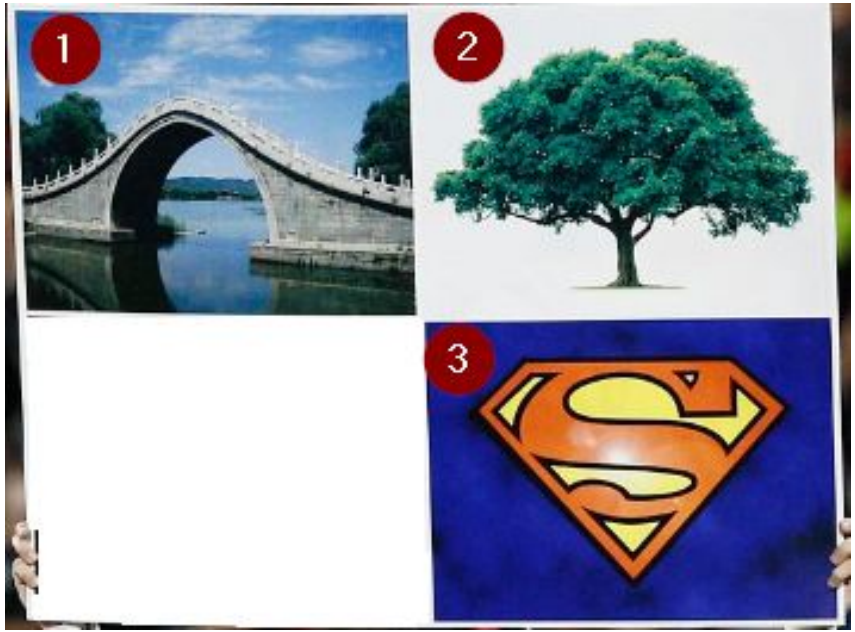


Image 1	Image 2	Image 3	Label
YES	NO	YES	**Run**
NO	YES	YES	**Pass**
NO	NO	YES	**Run**
YES	YES	YES	**Pass**
No	YES	YES	**Pass**
YES	NO	YES	**Run**

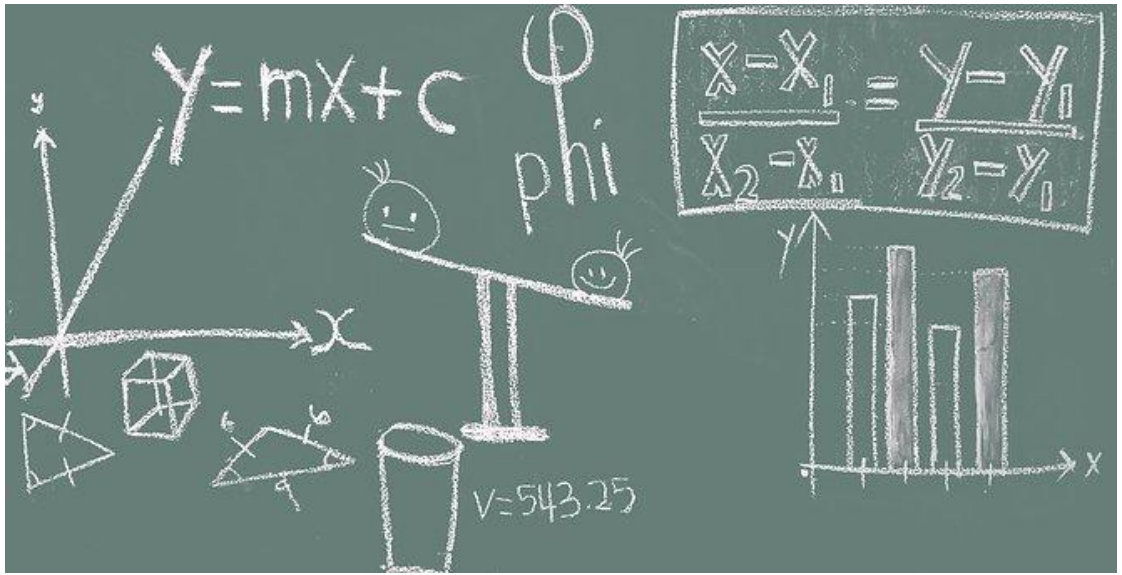
How to think about data

Neural networks don’t read football placards.

What you know			What you wanna know
Image 1	**Image 2**	**Image 3**	**Label**
YES	NO	YES	**Run**
NO	YES	YES	**Pass**
NO	NO	YES	**Run**
YES	YES	YES	**Pass**
No	YES	YES	**Pass**
YES	NO	YES	**Run**

How to think about data

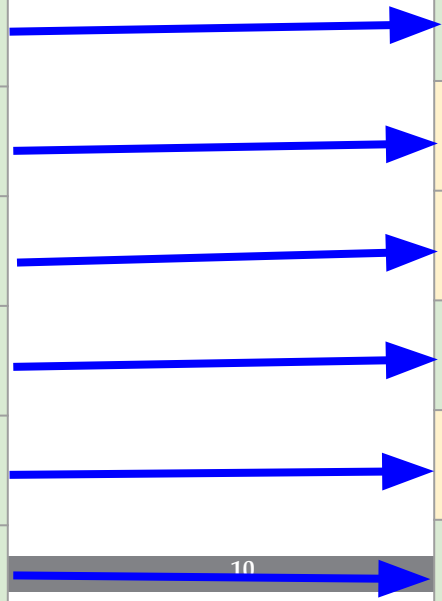
We need to see this in a math context.



How to think about data

Poster Boards

Image 1	Image 2	Image 3
YES	NO	YES
NO	YES	YES
NO	NO	YES
YES	YES	YES
No	YES	YES
YES	NO	YES



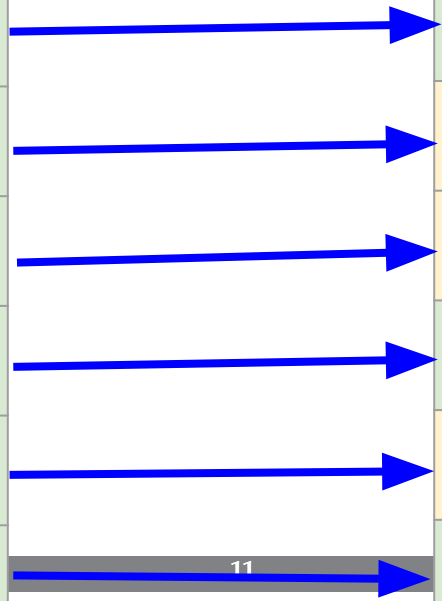
Board Pattern

Image 1	Image 2	Image 3
1	0	1
0	1	1
0	0	1
1	1	1
0	1	1
1	0	1

How to think about data

Poster Boards

Image 1	Image 2	Image 3
YES	BOTH	YES
NO	YES	YES
NO	NO	YES
YES	YES	YES
No	BOTH	YES
YES	NO	YES



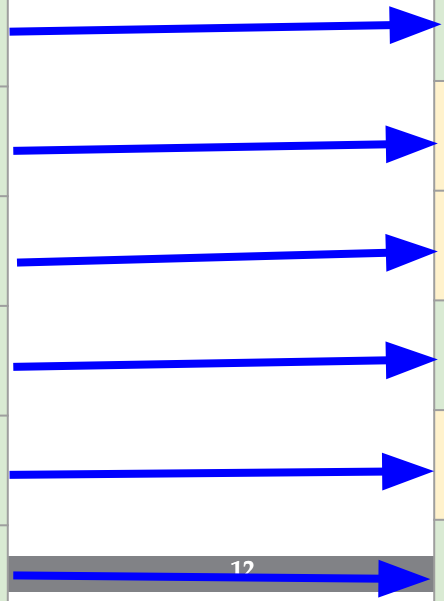
Board Pattern

Image 1	Image 2	Image 3
1	.5	1
0	1	1
0	0	1
1	1	1
0	.5	1
1	0	1

How to think about data

Poster Boards

Image 1	Image 2	Image 3
YES	BOTH	YES
NO	YES	YES
NO	NO	YES
YES	YES	YES
No	BOTH	YES
YES	NO	YES

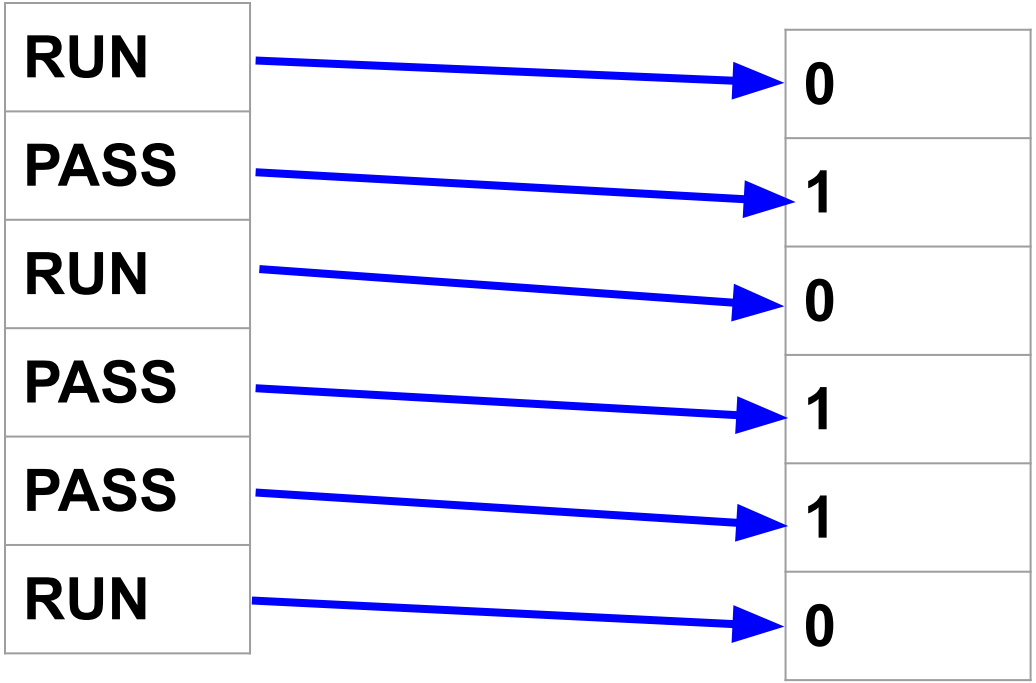


Board Pattern

Image 1	Image 2	Image 3
10	5	10
0	10	10
0	0	10
10	10	10
0	5	10
10	0	10

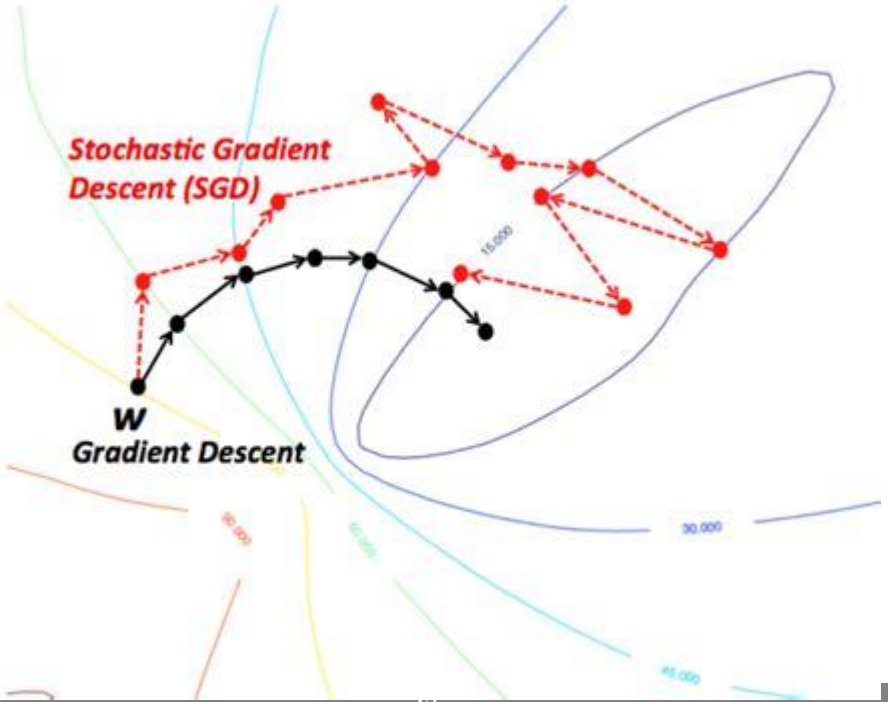
How to think about data

How to think about data



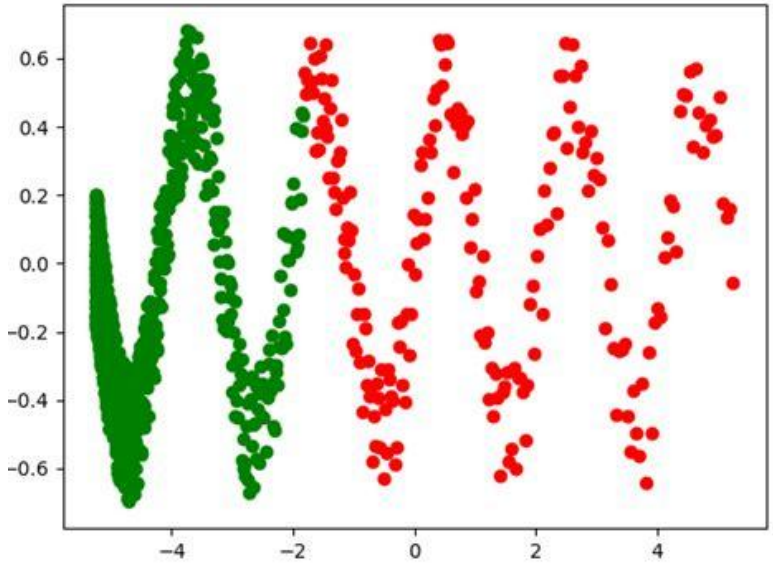
stochastic gradient descent

Stochastic gradient descent updates weights one example



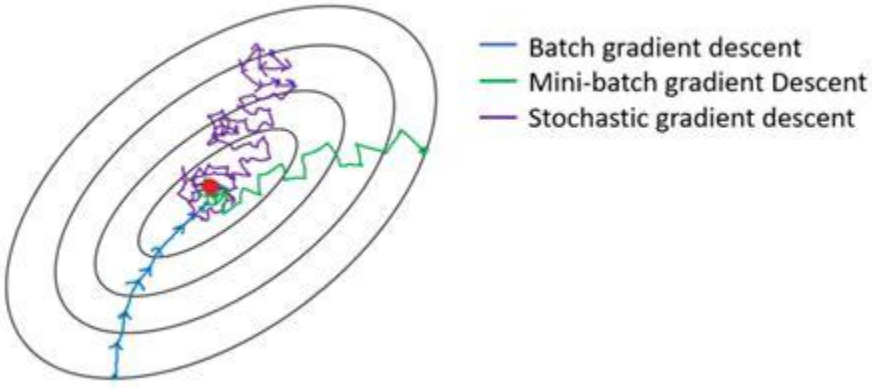
regular gradient descent

(Average/Regular/Full) gradient descent updates weights one dataset at a time.



batch/mini batch gradient descent

(BATCH) gradient descent updates weights one dataset at a time.



TRUISM: Neural Networks Learn Correlations

What did the neural network learn?

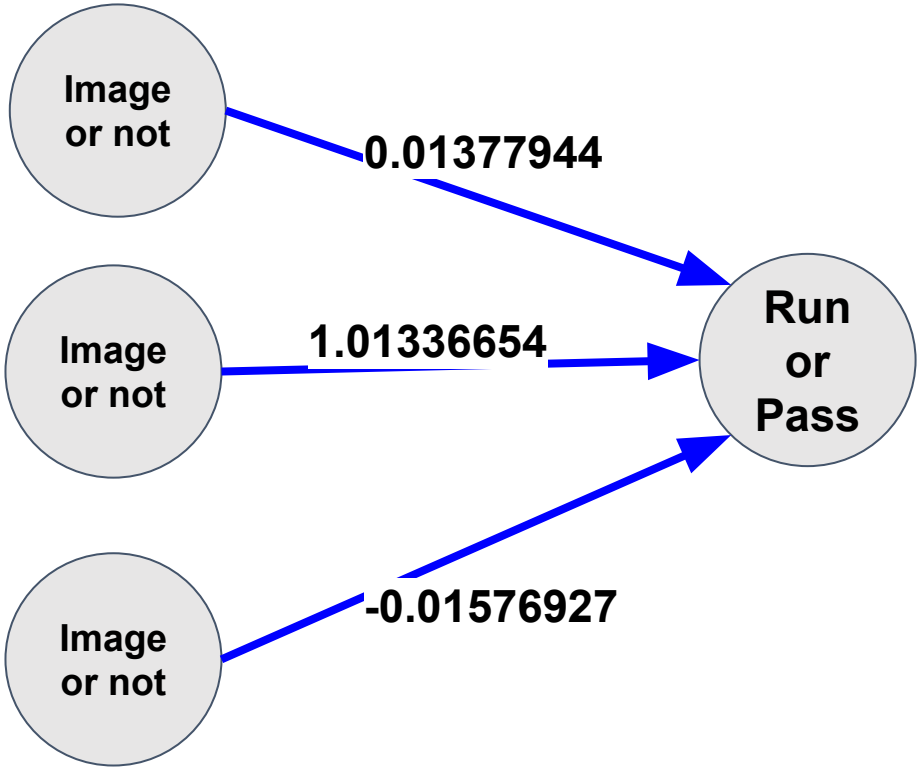


Image or not
Image or not
Image or not
0.01377944
1.01336654
-0.01576927
Run or Pass

Push/Pull of weights comes from the data

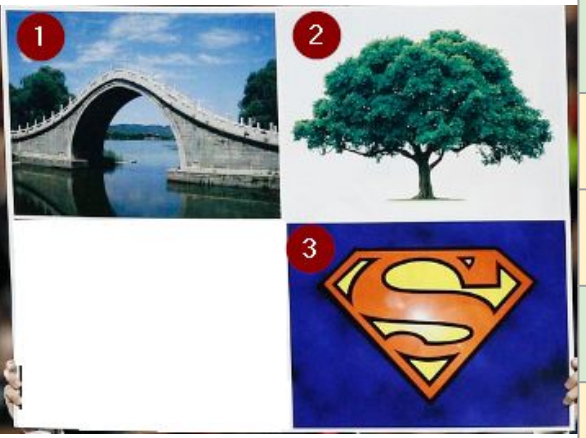
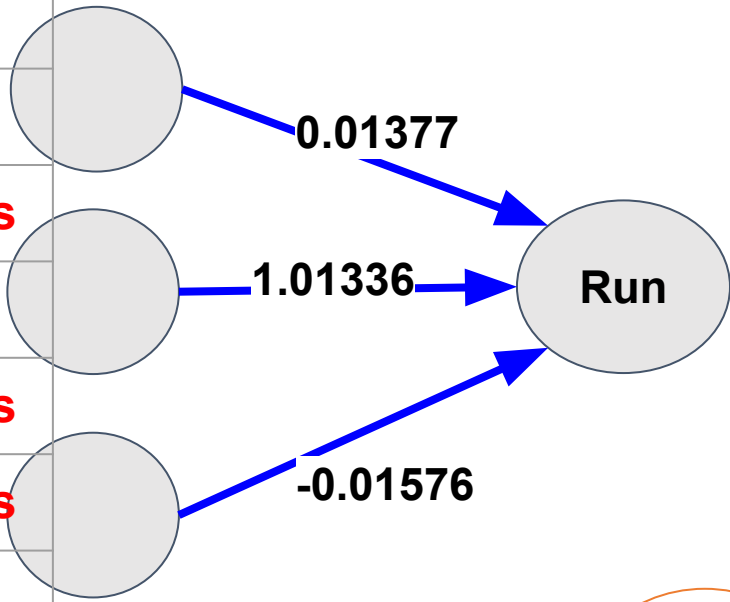


Image 1	Image 2	Image 3	Label
YES	NO	YES	Run
NO	YES	YES	Pass
NO	NO	YES	Run
YES	YES	YES	Pass
No	YES	YES	Pass
YES	NO	YES	Run



Push/Pull of weights comes from the data

Image 1	Image 2	Image 3	Label
1	0	1	Run
0	1	1	Pass
0	0	1	Run
1	1	1	Pass
0	1	1	Pass
1	0	1	Run

Image 1	Image 2	Image 3	Label
-	0	-	Run
0	+	+	Pass
0	0	-	Run
+	+	+	Pass
0	+	+	Pass
-	0	-	Run

Up and down pressure

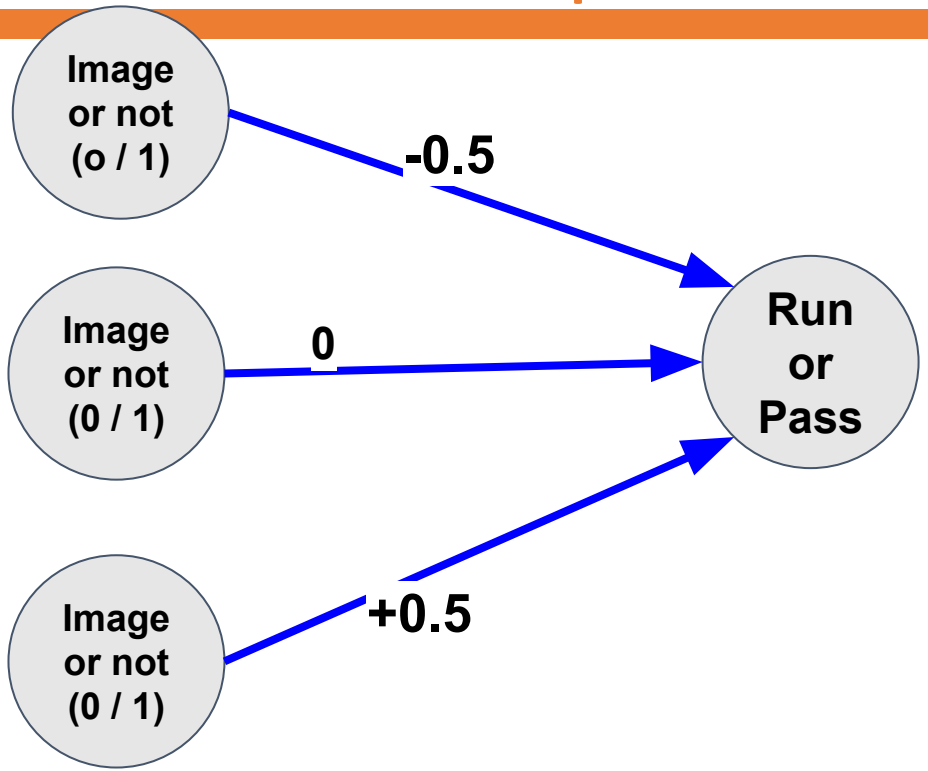
Image 1	Image 2	Image 3	Label
1	0	1	Run
0	1	1	Pass
0	0	1	Run
1	1	1	Pass
0	1	1	Pass
1	0	1	Run

Image 1	Image 2	Image 3	Label
-	0	-	Run
0	+	+	Pass
0	0	-	Run
+	+	+	Pass
0	+	+	Pass
-	0	-	Run

Up and down pressure

Overfitting

Correlation Oops…



Overfitting

How does it learn with all the conflicts?

Know thyself - correlation

Image 1	Image 2	Image 3	Label
1	0	1	Run
0	1	1	Pass
0	0	1	Run
1	1	1	Pass
0	1	1	Pass
1	0	1	Run

Image 1	Image 2	Image 3	Label
-	0	-	Run
0	+	+	Pass
0	0	-	Run
+	+	+	Pass
0	+	+	Pass
-	0	-	Run

Image 1	Image 2	Image 3	Label
1	0	1	Run
0	1	1	Pass
0	0	1	Run
1	1	1	Pass
0	1	1	Pass
1	0	1	Run

Image 1	Image 2	Image 3	Label
-	0	-	Run
0	+	+	Pass
0	0	-	Run
+	+	+	Pass
0	+	+	Pass
-	0	-	Run

Image 1	Image 2	Image 3	Label
1	0	1	Run
0	1	1	Pass
0	0	1	Run
1	1	1	Pass
0	1	1	Pass
1	0	1	Run

Image 1	Image 2	Image 3	Label
-	0	-	Run
0	+	+	Pass
0	0	-	Run
+	+	+	Pass
0	+	+	Pass
-	0	-	Run

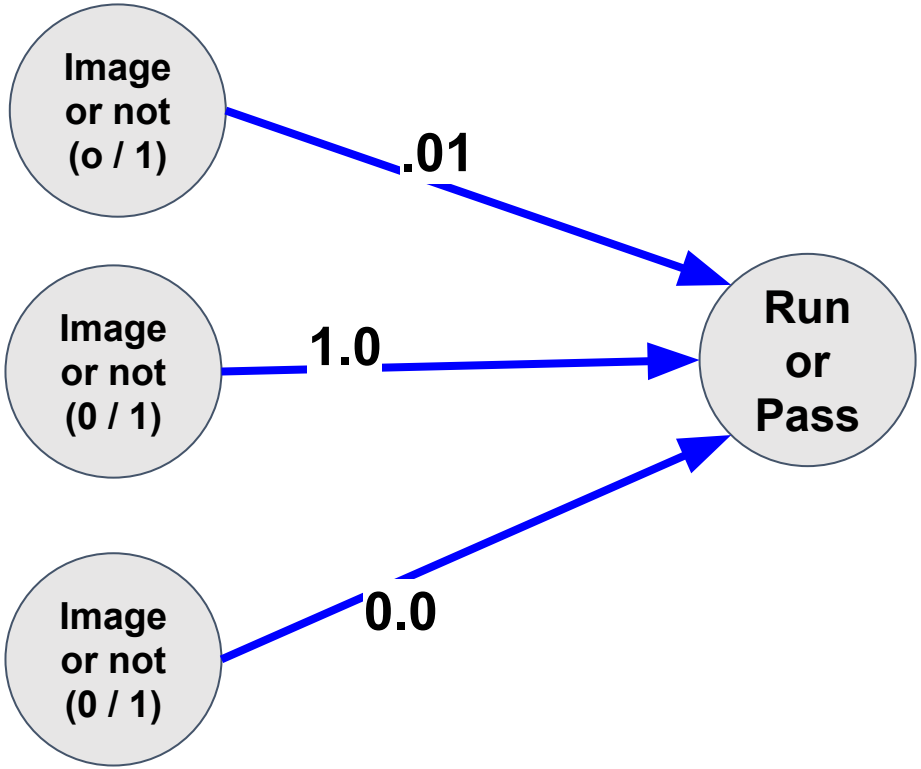
Edge case: Conflicting pressure

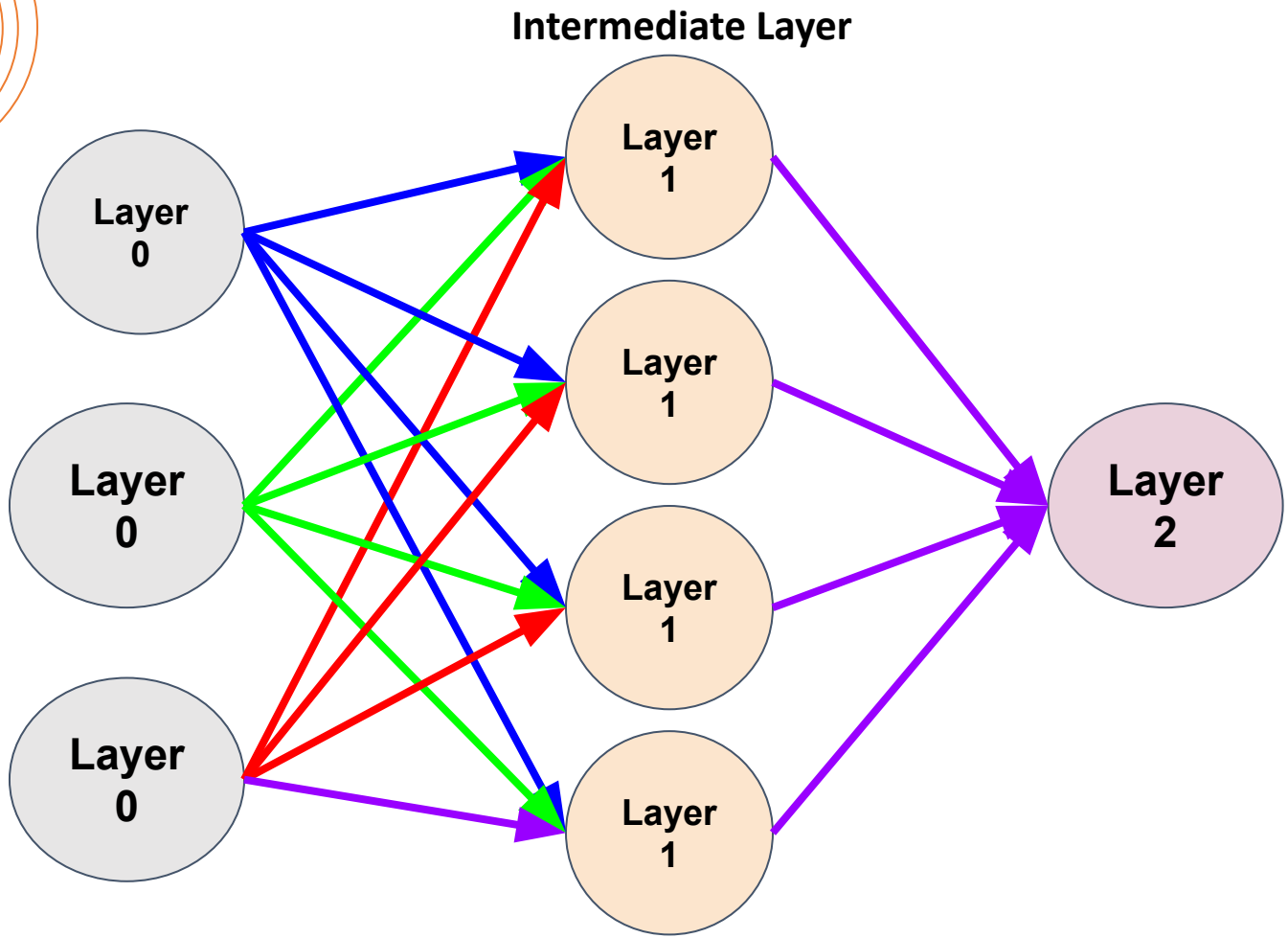
Image 1	Image 2	Image 3	Label
1	0	1	Pass
0	1	1	Pass
0	0	1	Run
1	1	1	Run

Image 1	Image 2	Image 3	Label
+	0	+	Pass
0	+	+	Pass
0	0	-	Run
-	-	-	Run

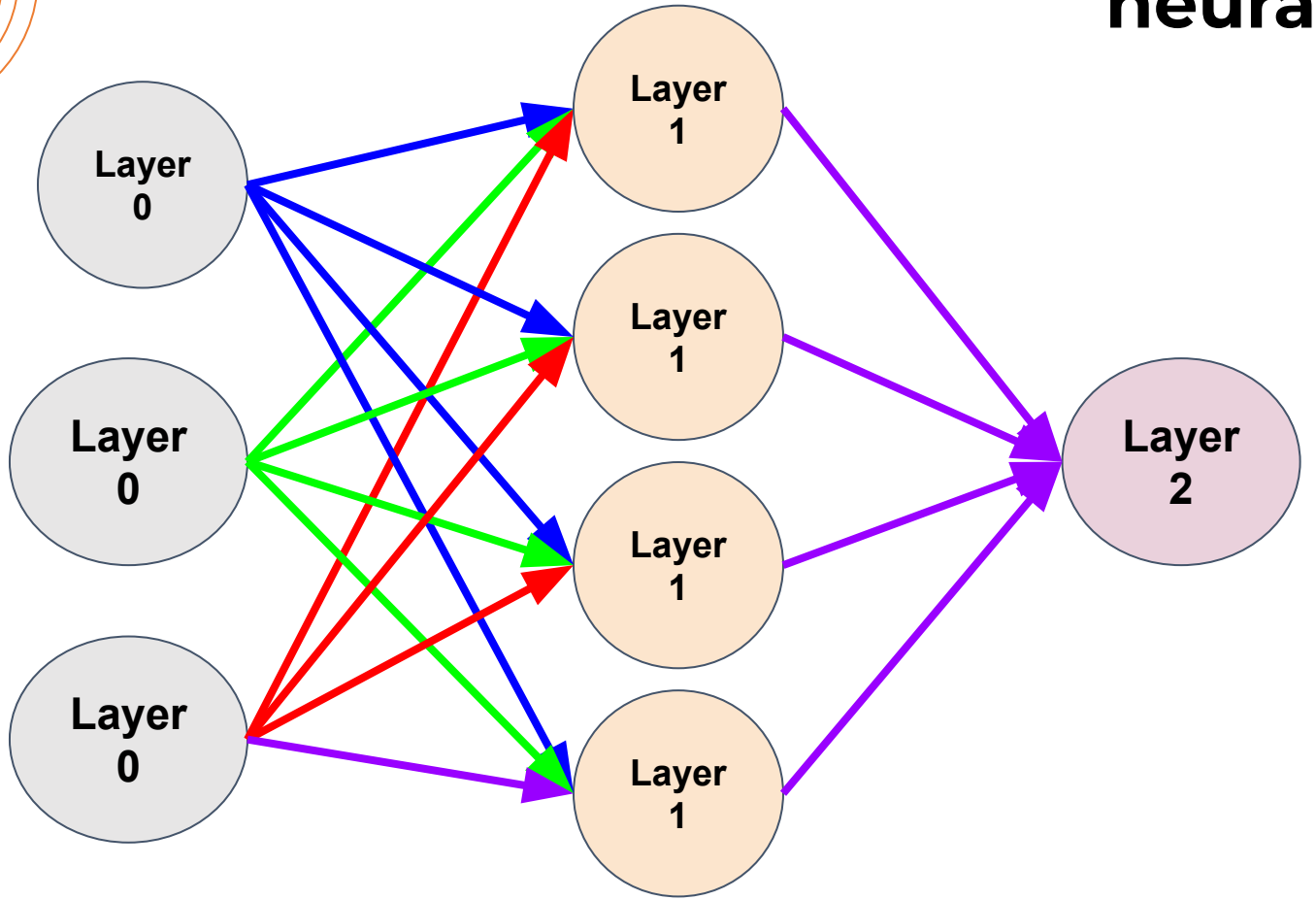
Learning indirect correlation

create an intermediate data set

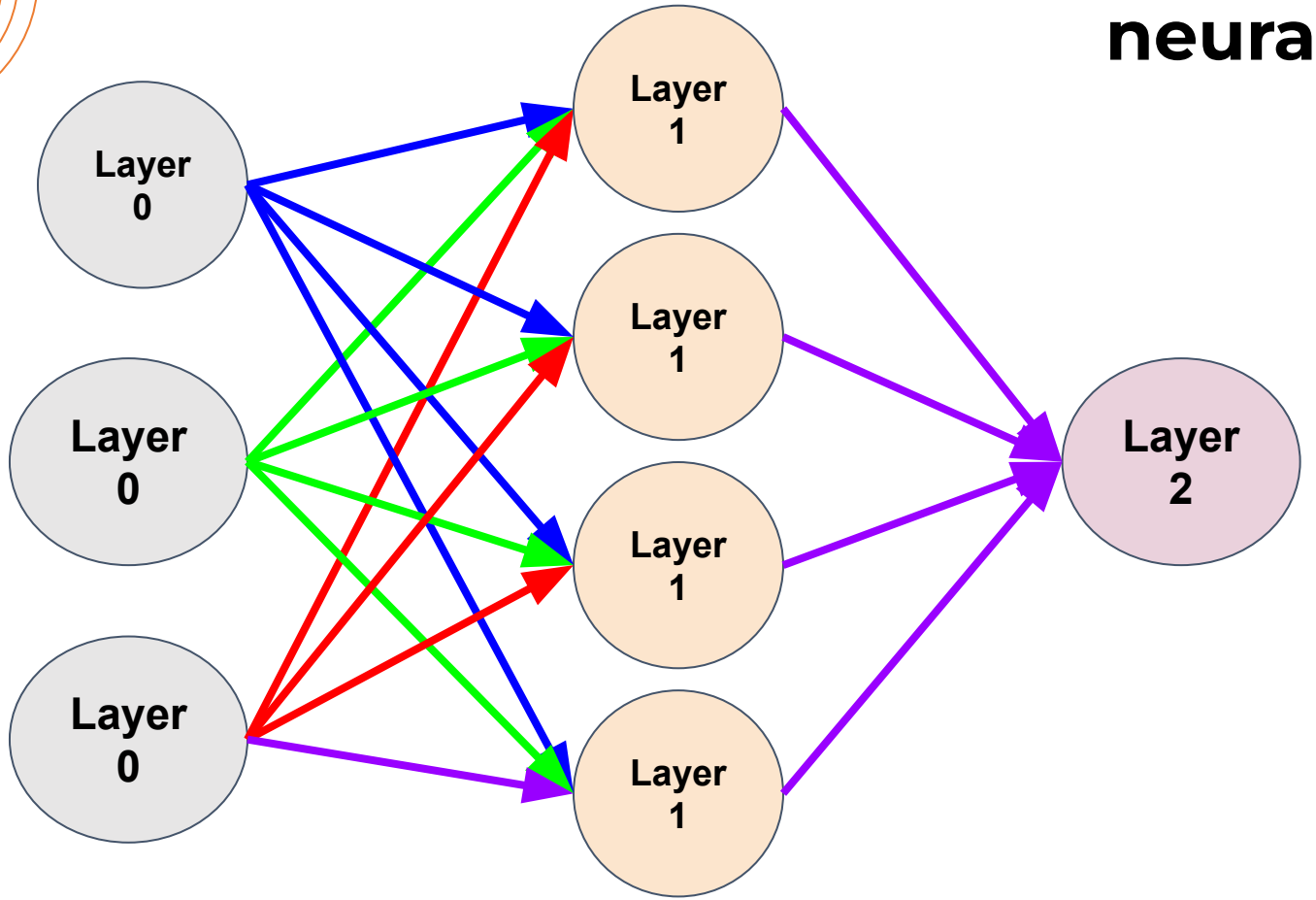



Intermediate Layer
Layer 0
Layer 0
Layer 0
Layer 1
Layer 1
Layer 1
Layer 1
Layer 2

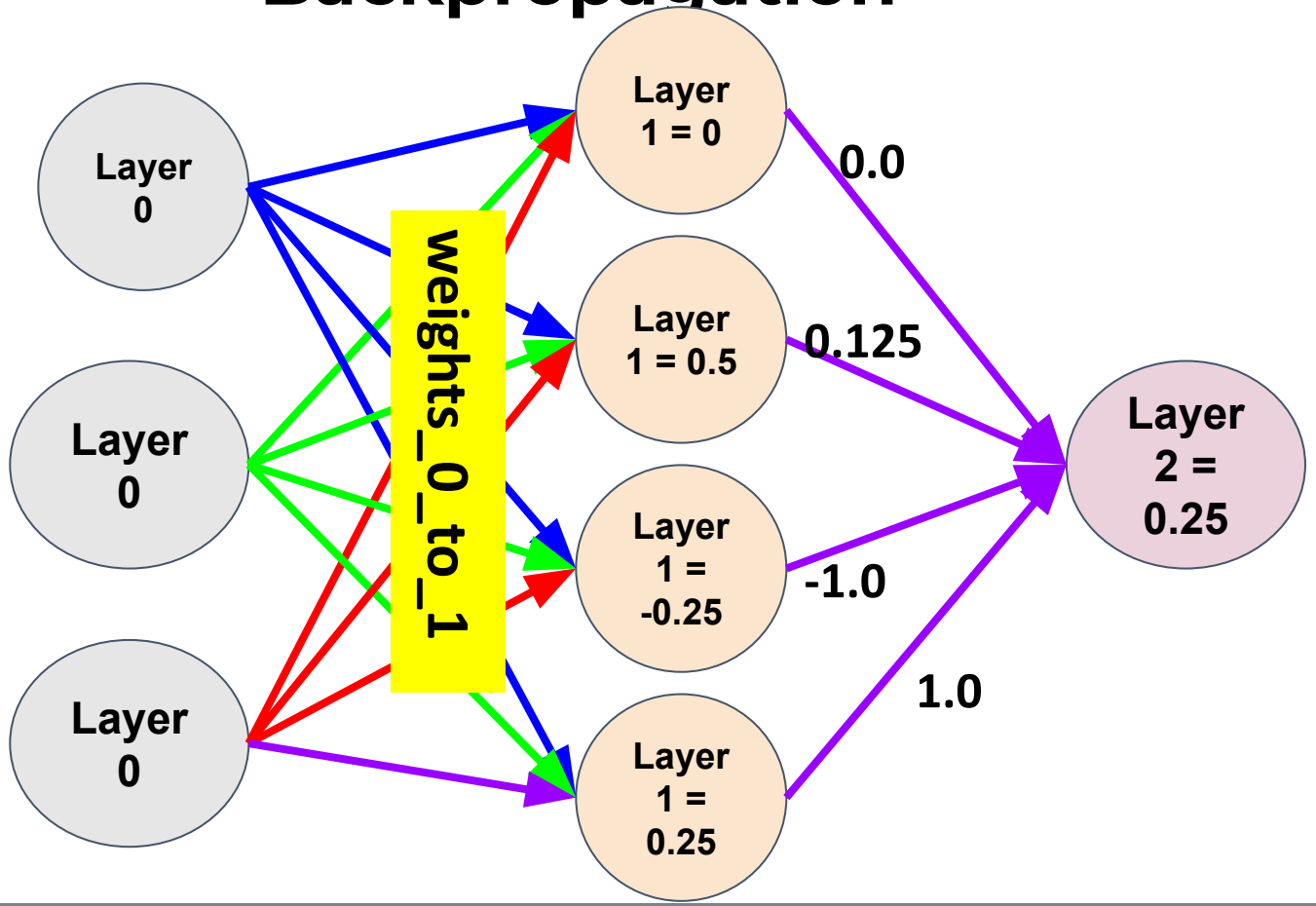
neural networks



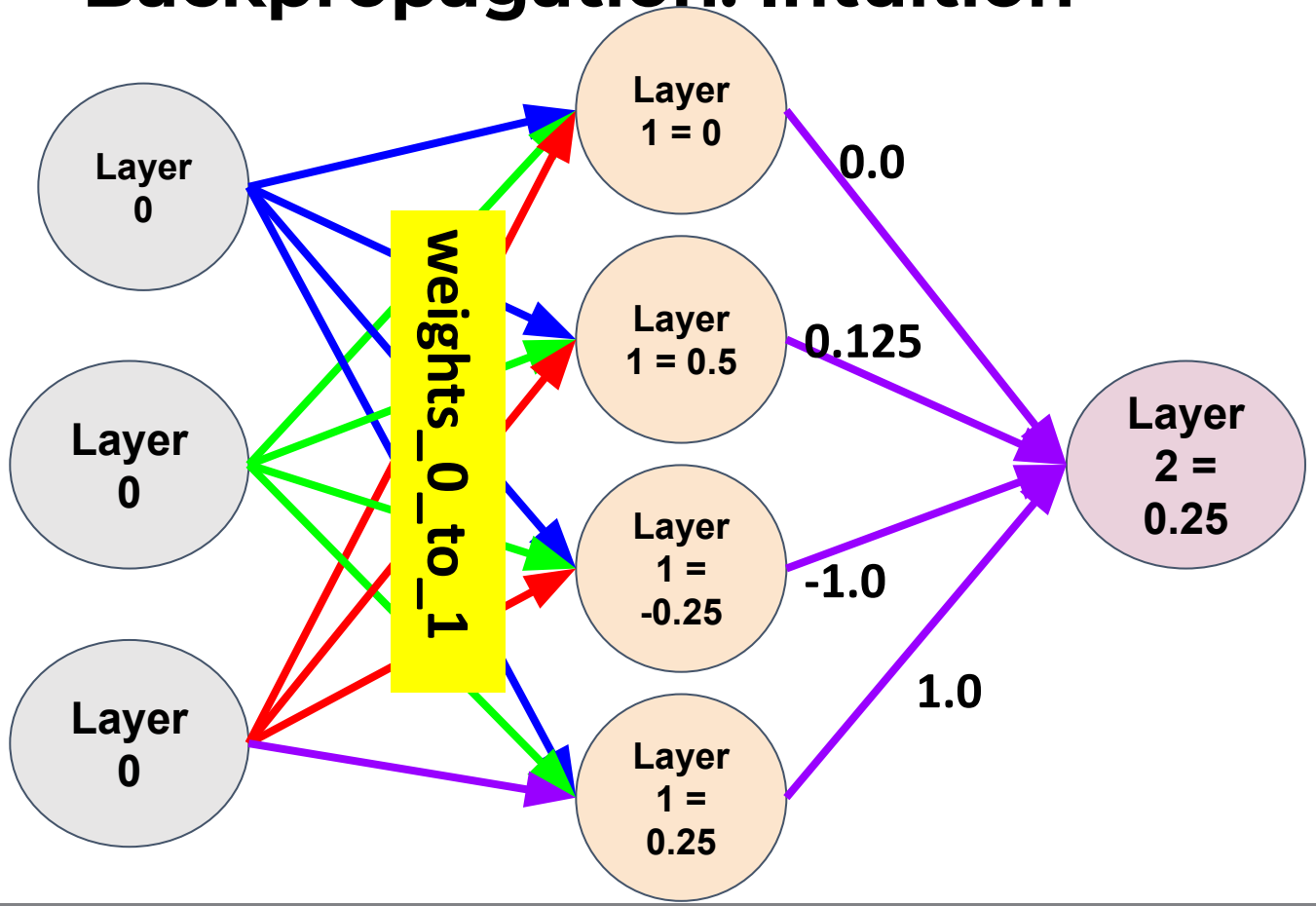
neural networks



Backpropagation



Backpropagation: Intuition

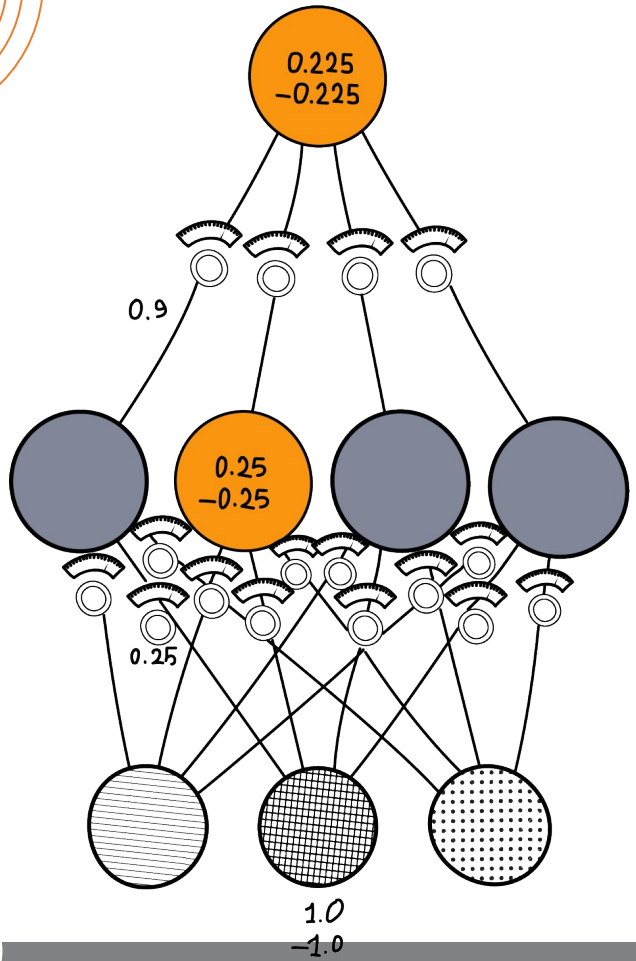


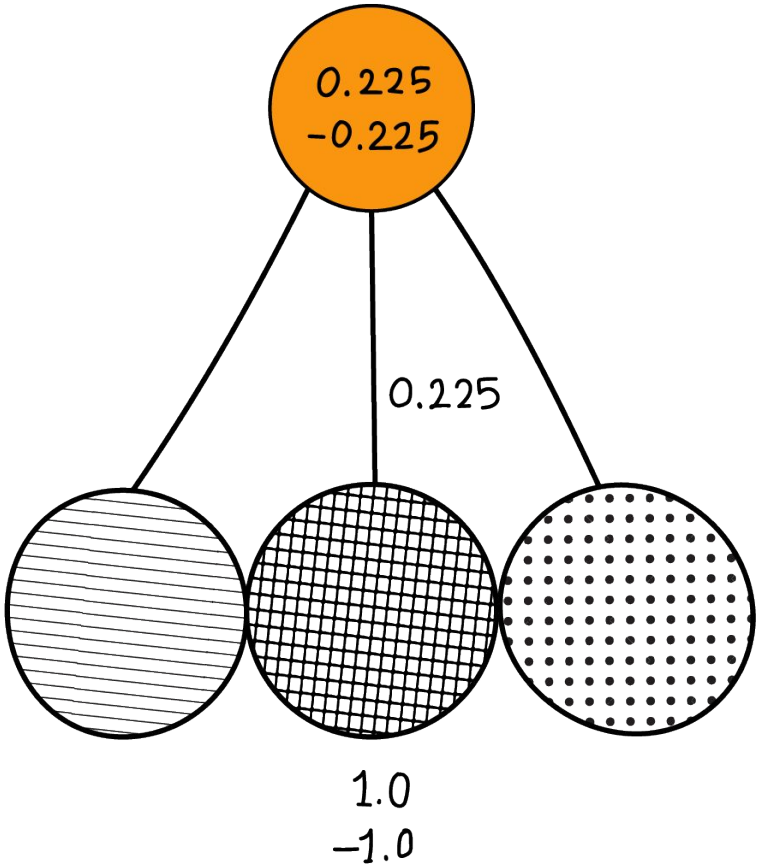
Linear vs. nonlinear

$12 \times 1 = 12 = 2 \times 5 \times 1.2$

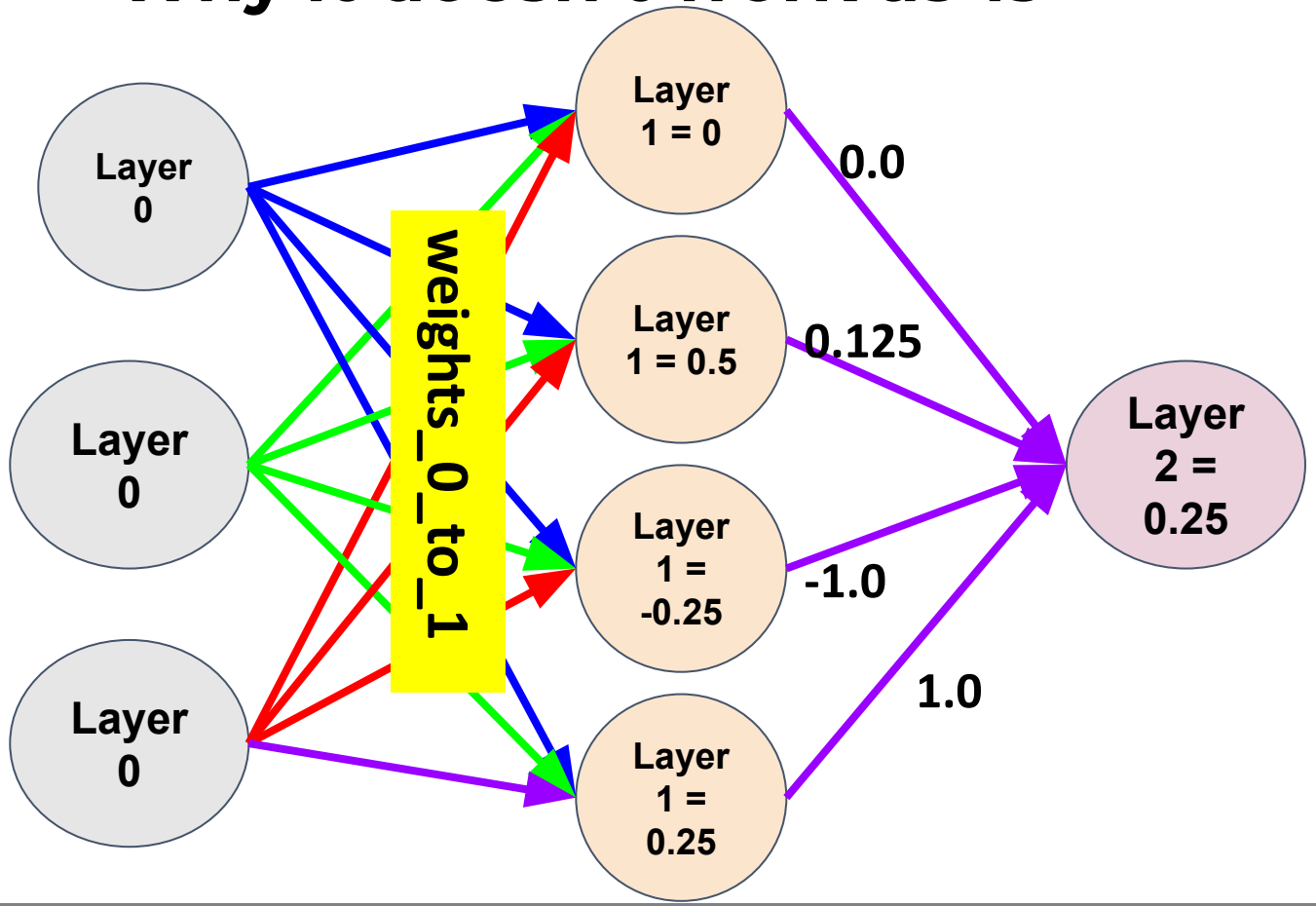
$6 \times 2 = 12 = 9 \times .25 \times 4$

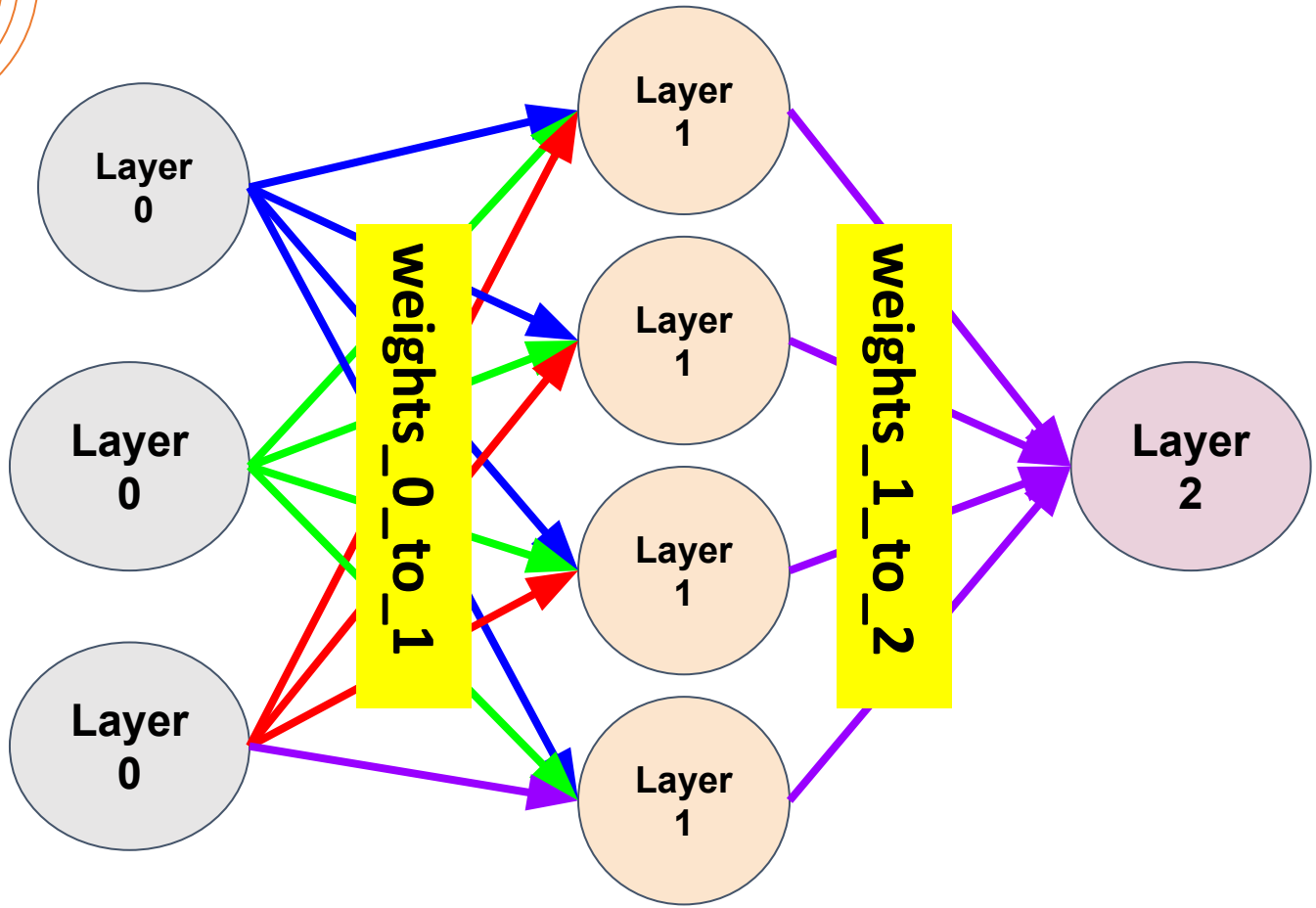
$3 \times 4 = 12 = 6 \times .5 \times 4$


0.225
−0.225
0.9
0.25
−0.25
0.25
1.0
−1.0

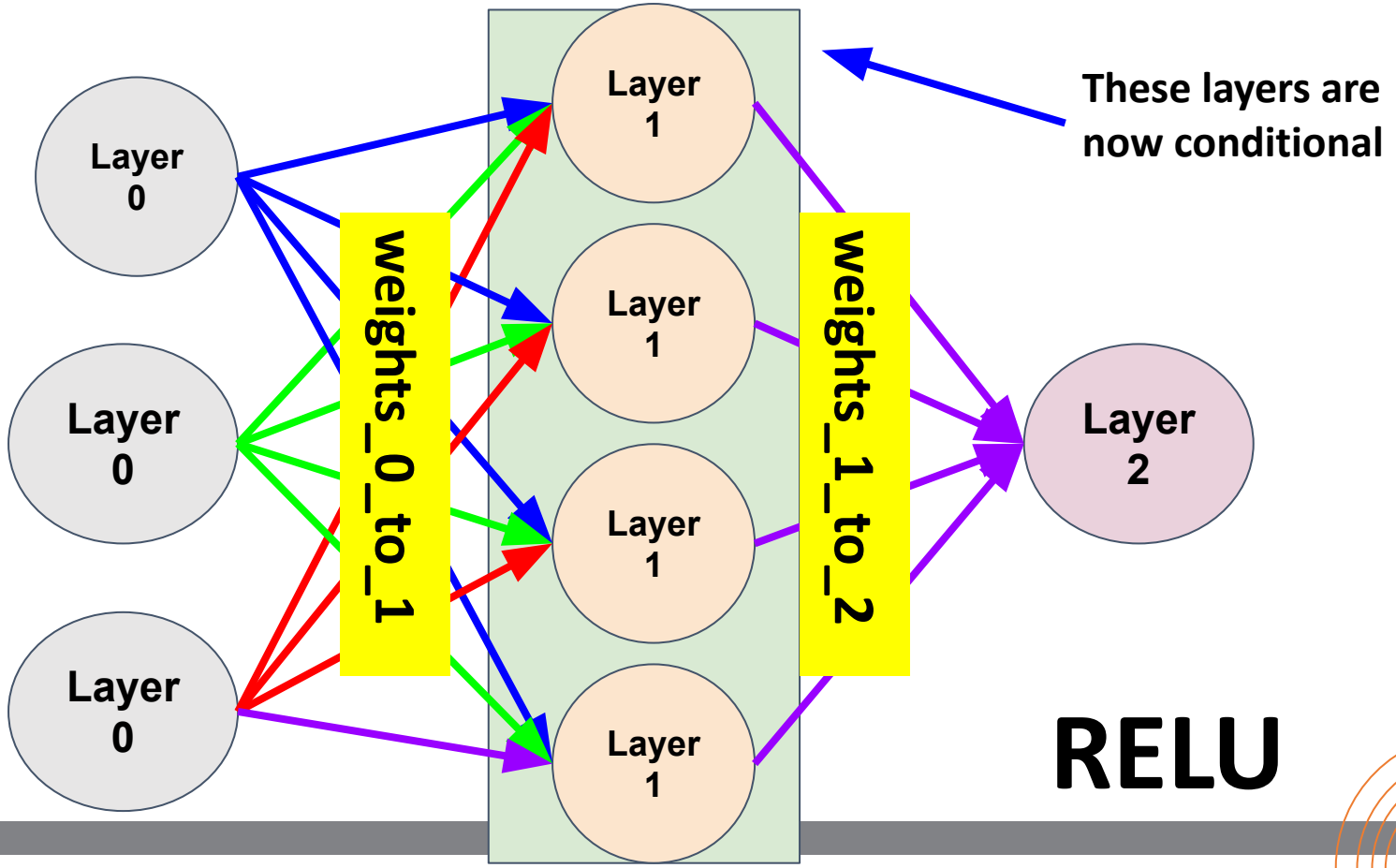

0.225
−0.225
0.225
1.0
−1.0

Why it doesn't work as-is




Layer
0
Layer
0
Layer
0
weights_0_to_1
Layer
1
Layer
1
Layer
1
Layer
1
weights_1_to_2
Layer
2

The Insight


Layer 0
Layer 0
Layer 0
weights_0_to_1
Layer 1
Layer 1
Layer 1
Layer 1
weights_1_to_2
Layer 2
These layers are now conditional
RELU

ABSORB

Why do deep networks matter?

- No individual pixel correlates with whether there's a dog in the picture.
- Only different configurations of pixels correlate with whether Daisy is in the picture.