10701 Recitation 3- Backpropagation CNN

SVM, Kernel, Backpropagation, CNN

Backpropagation and CNN

- Simple neural network with demo of backpropagation
    - XOR (need to search for it)
- Why is backpropagation helpful in neural networks?
- LeNet implementation
    - What are k, s, p, ... in the convolutional layer and pooling layer
    - Demo of lenet in action

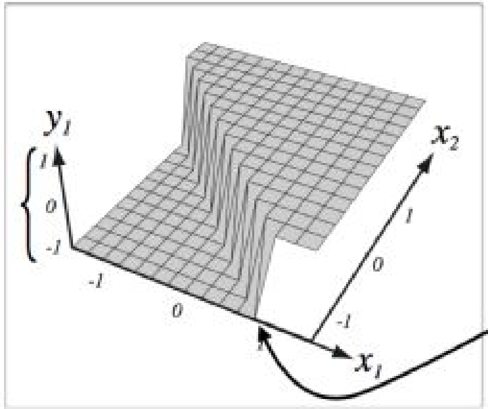
How many layers do you need to construct a neural network that achieves XOR?

Artificial Neuron

- Output activation of the neuron:

$$h(\mathbf{x}) = g(a(\mathbf{x})) = g(b + \sum_i w_i x_i)$$

Range is determined by $g(\cdot)$

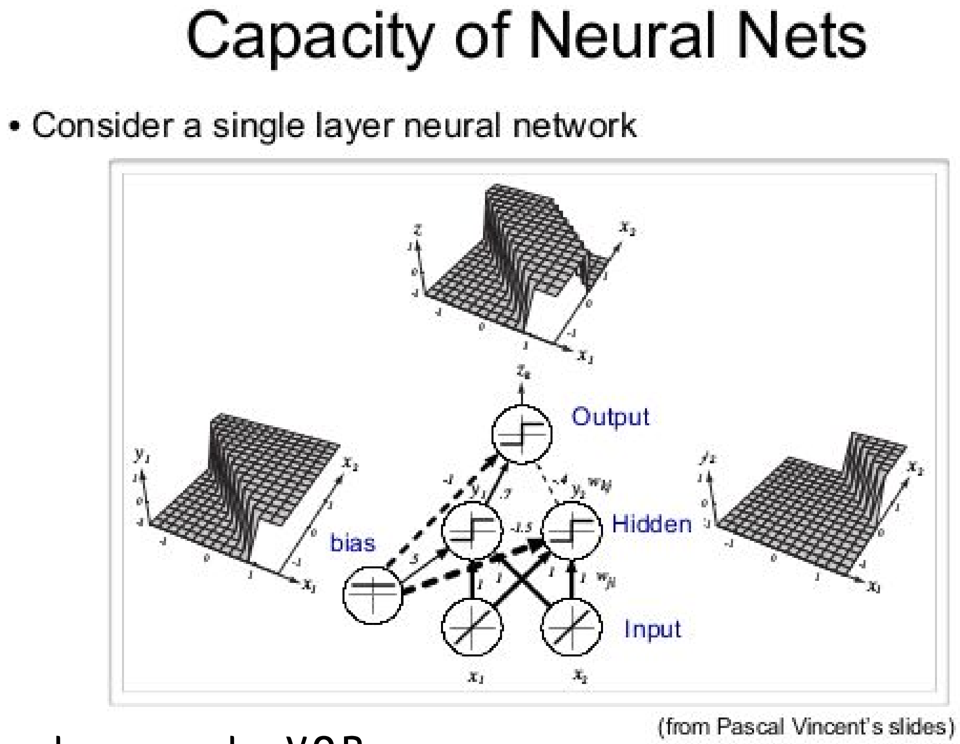


(from Pascal Vincent's slides)

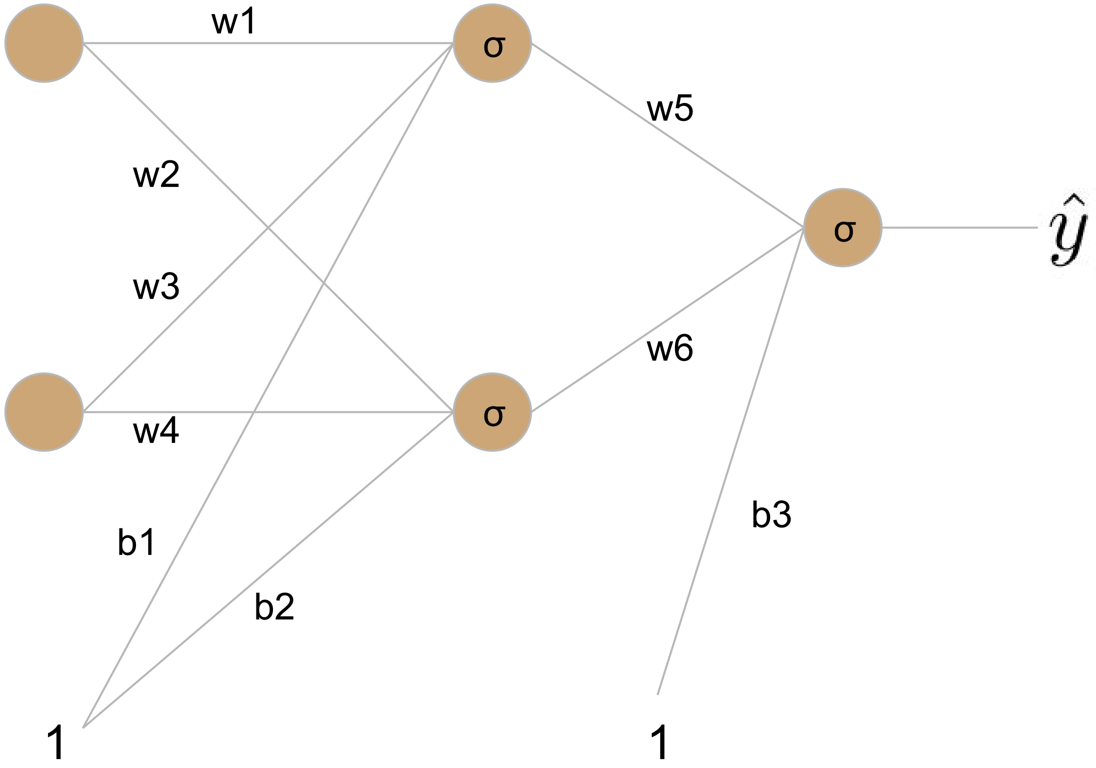
Bias only changes the position of the riff

Backpropagation simple example: XOR

How many layers do you need to construct a neural network that achieves XOR?



Backpropagation simple example: XOR



$$l = (y - \hat{y})^2$$

Backpropagation simple example: XOR

Derivation

Let the loss be $l = (y - \hat{y})^2$

- the special case: $\frac{\partial l}{\partial \hat{y}} = 2(y - \hat{y}) \cdot (-1)$
- layer 2 need to compute these terms

$$\frac{\partial l}{\partial w_5} = \frac{\partial l}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_5}$$
$$\frac{\partial l}{\partial w_6} = \frac{\partial l}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_6}$$
$$\frac{\partial l}{\partial z_1} = \frac{\partial l}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_1}$$
$$\frac{\partial l}{\partial z_2} = \frac{\partial l}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_2}$$

Derivation

The first terms in the above equation are carried over from the previous step, and we need to compute the second term.

$$\begin{aligned}\frac{\partial \hat{y}}{\partial w_5} =& \delta'(w_5 z_1 + w_6 z_2 + b_3) \cdot z_1 \\ =& \delta(w_5 z_1 + w_6 z_2 + b_3) \cdot (1 - \delta(w_5 z_1 + w_6 z_2 + b_3)) \cdot z_1 \\ =& \hat{y} \cdot (1 - \hat{y}) \cdot z_1\end{aligned}$$

Similarly

$$\begin{aligned}\frac{\partial \hat{y}}{\partial w_6} =& \hat{y} \cdot (1 - \hat{y}) \cdot z_2 \\ \frac{\partial \hat{y}}{\partial b} =& \hat{y} \cdot (1 - \hat{y}) \\ \frac{\partial \hat{y}}{\partial z_1} =& \hat{y} \cdot (1 - \hat{y}) \cdot w_5 \\ \frac{\partial \hat{y}}{\partial z_2} =& \hat{y} \cdot (1 - \hat{y}) \cdot w_6\end{aligned}$$

Derivation

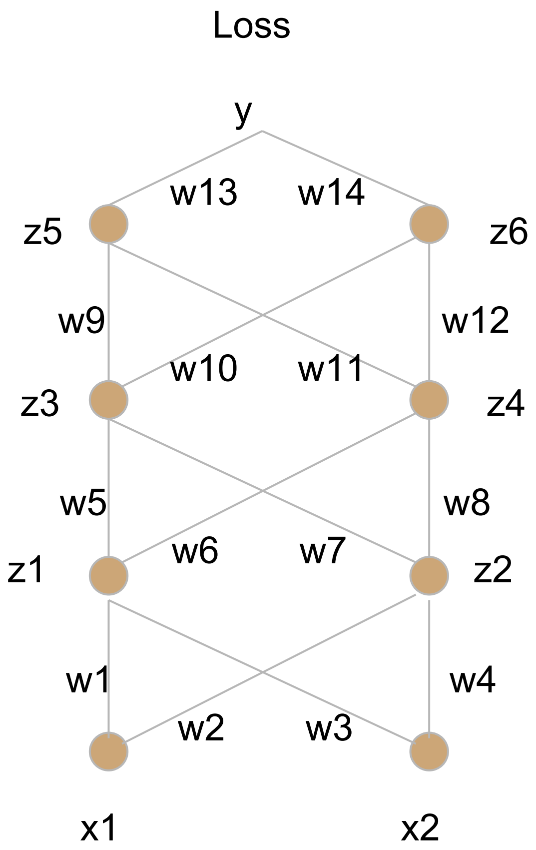
– layer 1 we need to compute these terms

$$\frac{\partial l}{\partial w_1} = \frac{\partial l}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_1}$$
$$\frac{\partial l}{\partial w_2} = \frac{\partial l}{\partial z_2} \cdot \frac{\partial z_2}{\partial w_2}$$
$$\frac{\partial l}{\partial w_3} = \frac{\partial l}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_3}$$
$$\frac{\partial l}{\partial w_4} = \frac{\partial l}{\partial z_2} \cdot \frac{\partial z_2}{\partial w_4}$$
$$\frac{\partial l}{\partial b_1} = \frac{\partial l}{\partial z_1} \cdot \frac{\partial z_1}{\partial b_1}$$
$$\frac{\partial l}{\partial b_2} = \frac{\partial l}{\partial z_2} \cdot \frac{\partial z_2}{\partial b_2}$$

$$\frac{\partial z_1}{\partial w_1} = z_1(1 - z_1)x_1$$
$$\frac{\partial z_2}{\partial w_2} = z_2(1 - z_2)x_1$$
$$\frac{\partial z_1}{\partial w_3} = z_1(1 - z_1)x_2$$
$$\frac{\partial z_2}{\partial w_4} = z_2(1 - z_2)x_2$$
$$\frac{\partial z_1}{\partial b_1} = z_1(1 - z_1)$$
$$\frac{\partial z_2}{\partial b_2} = z_2(1 - z_2)$$

Why backpropagation?

$$l = f_I(w_I, f_{I-1}(w_{I-1}, \ldots))$$



Interpretation 1: since the order of differentiation is from the outer function to the inner function. This corresponds to differentiate upper levels first, thus backpropagation

Interpretation 2: We can see from the toy example that the number of terms computed from the backward propagation is linear in the number of nodes (or weights), but roughly quadratic for the forward path

- Backward path (terms in parenthesis are carried over from previous layer)
  - last layer

$$\frac{\partial l}{\partial \hat{y}}$$

  - layer 4

$$\frac{\partial l}{\partial w_{13}}, \frac{\partial l}{\partial w_{14}}, \frac{\partial l}{\partial z_5}, \frac{\partial l}{\partial z_6} (\frac{\partial l}{\partial \hat{y}})$$

  - layer 3

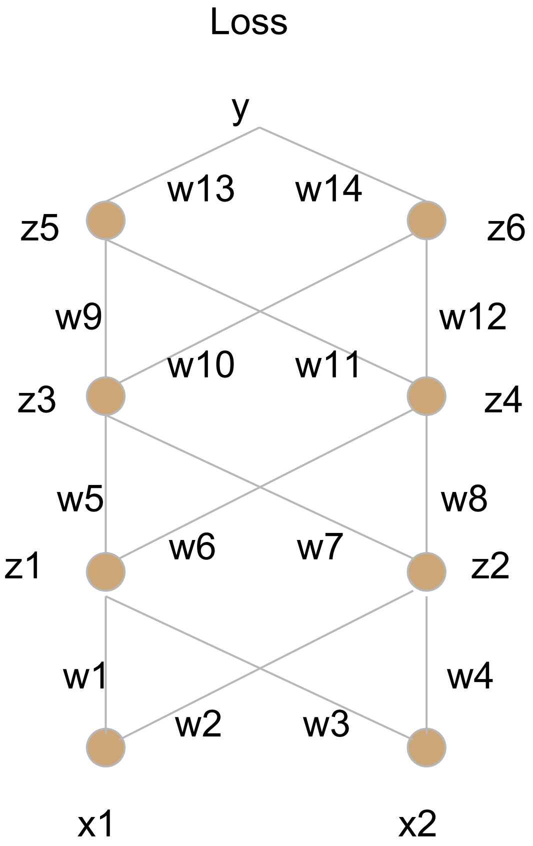
$$\frac{\partial l}{\partial w_9}, \frac{\partial l}{\partial w_{10}}, \frac{\partial l}{\partial w_{11}}, \frac{\partial l}{\partial w_{12}}, \frac{\partial l}{\partial z_3}, \frac{\partial l}{\partial z_4} (\frac{\partial l}{\partial z_5}, \frac{\partial l}{\partial z_6})$$

  - layer 2

$$\frac{\partial l}{\partial w_5}, \frac{\partial l}{\partial w_6}, \frac{\partial l}{\partial w_7}, \frac{\partial l}{\partial w_8}, \frac{\partial l}{\partial z_1}, \frac{\partial l}{\partial z_2} (\frac{\partial l}{\partial z_3}, \frac{\partial l}{\partial z_4})$$

  - layer 1

$$\frac{\partial l}{\partial w_1}, \frac{\partial l}{\partial w_2}, \frac{\partial l}{\partial w_3}, \frac{\partial l}{\partial w_4}, (\frac{\partial l}{\partial z_1} \frac{\partial l}{\partial z_2})$$



Why backpropagation?

Each layer compute some constant number of terms (including carried over terms)

- Forward path
  - layer 1 (4)

$$\frac{\partial z_1}{\partial w_1}, \frac{\partial z_2}{\partial w_2}, \frac{\partial z_1}{\partial w_3}, \frac{\partial z_2}{\partial w_4}$$

  - layer 2 (12)

$$\frac{\partial z_3}{\partial w_5}, \frac{\partial z_3}{\partial w_7}, \frac{\partial z_4}{\partial w_6}, \frac{\partial z_4}{\partial w_8}$$

$$\frac{\partial z_3}{\partial w_1}, \frac{\partial z_3}{\partial w_2}, \frac{\partial z_3}{\partial w_3}, \frac{\partial z_3}{\partial w_4}, \frac{\partial z_4}{\partial w_1}, \frac{\partial z_4}{\partial w_2}, \frac{\partial z_4}{\partial w_3}, \frac{\partial z_4}{\partial w_4}$$

  - layer 3 (28)

$$\frac{\partial z_5}{\partial w_9}, \frac{\partial z_5}{\partial w_{11}}, \frac{\partial z_6}{\partial w_{10}} \frac{\partial z_6}{\partial w_{12}}$$

$$\frac{\partial z_5}{\partial w_1}, \frac{\partial z_5}{\partial w_2}, \frac{\partial z_5}{\partial w_3}, \frac{\partial z_5}{\partial w_4}, \frac{\partial z_5}{\partial w_5}, \frac{\partial z_5}{\partial w_6}, \frac{\partial z_5}{\partial w_7}, \frac{\partial z_5}{\partial w_8}, \frac{\partial z_5}{\partial w_9}, \frac{\partial z_5}{\partial w_{10}}, \frac{\partial z_5}{\partial w_{11}} \frac{\partial z_5}{\partial w_{12}}$$

$$\frac{\partial z_6}{\partial w_1}, \frac{\partial z_6}{\partial w_2}, \frac{\partial z_6}{\partial w_3}, \frac{\partial z_6}{\partial w_4}, \frac{\partial z_6}{\partial w_5}, \frac{\partial z_6}{\partial w_6}, \frac{\partial z_6}{\partial w_7}, \frac{\partial z_6}{\partial w_8}, \frac{\partial z_6}{\partial w_9}, \frac{\partial z_6}{\partial w_{10}}, \frac{\partial z_6}{\partial w_{11}}, \frac{\partial z_6}{\partial w_{12}}$$

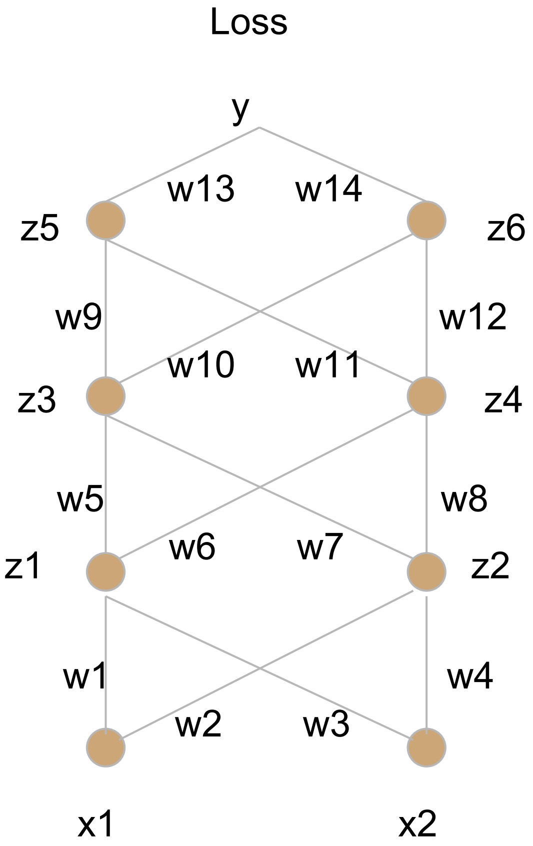
  - layer 4 (14)

$$\frac{\partial \hat{y}}{\partial w_{13}} \frac{\partial \hat{y}}{\partial w_{14}}$$

$$\frac{\partial \hat{y}}{\partial w_1}, \frac{\partial \hat{y}}{\partial w_2}, \frac{\partial \hat{y}}{\partial w_3}, \frac{\partial \hat{y}}{\partial w_4}, \frac{\partial \hat{y}}{\partial w_5}, \frac{\partial \hat{y}}{\partial w_6}, \frac{\partial \hat{y}}{\partial w_7}, \frac{\partial \hat{y}}{\partial w_8}, \frac{\partial \hat{y}}{\partial w_9}, \frac{\partial \hat{y}}{\partial w_{10}}, \frac{\partial \hat{y}}{\partial w_{11}}, \frac{\partial \hat{y}}{\partial w_{12}}$$

  - last layer (14)

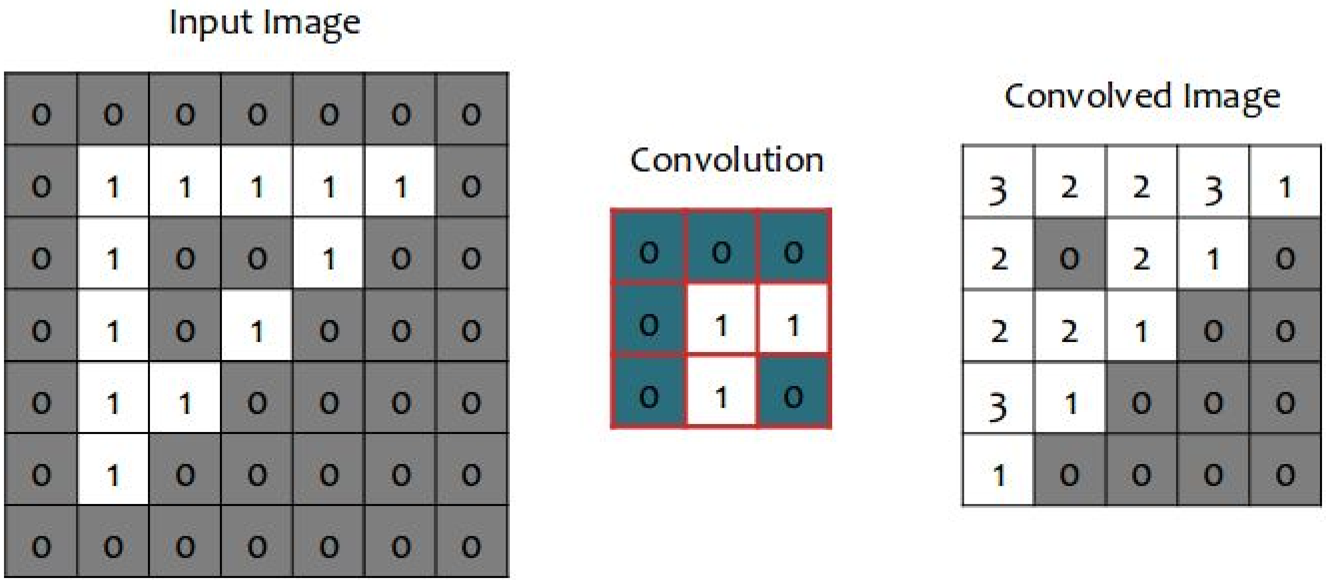
$$\frac{\partial l}{\partial w_1}, \frac{\partial l}{\partial w_2}, \frac{\partial l}{\partial w_3}, \frac{\partial l}{\partial w_4}, \frac{\partial l}{\partial w_5}, \frac{\partial l}{\partial w_6}, \frac{\partial l}{\partial w_7}, \frac{\partial l}{\partial w_8}, \frac{\partial l}{\partial w_9}, \frac{\partial l}{\partial w_{10}}, \frac{\partial l}{\partial w_{11}}, \frac{\partial l}{\partial w_{12}}, \frac{\partial l}{\partial w_{13}}, \frac{\partial l}{\partial w_{14}}$$



Why backpropagation?

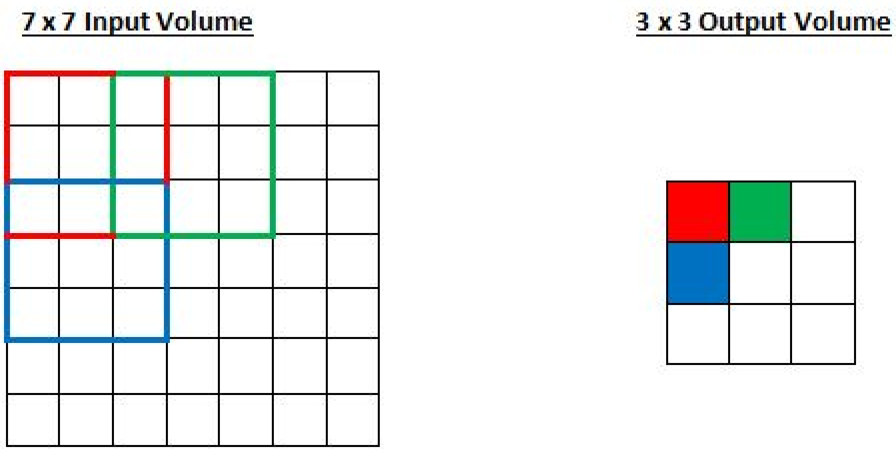
Each layer compute 8 more terms than the previous layer

Demon of convolution operation

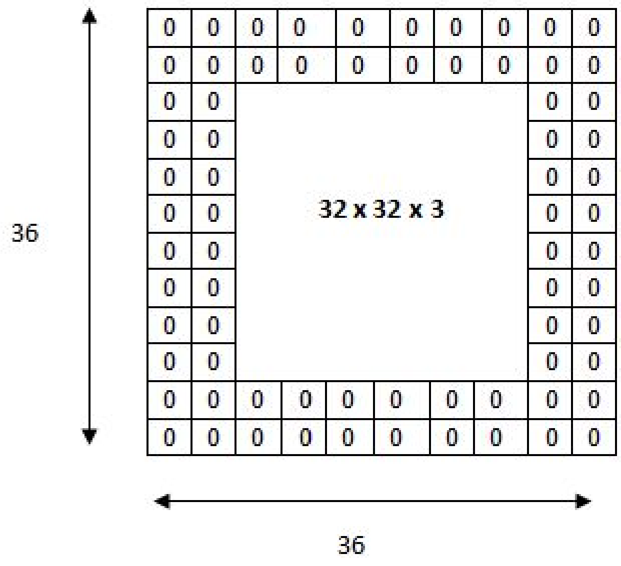


Stolen from f16 10601 slides

What are the stride, padding, size of the receptive fields



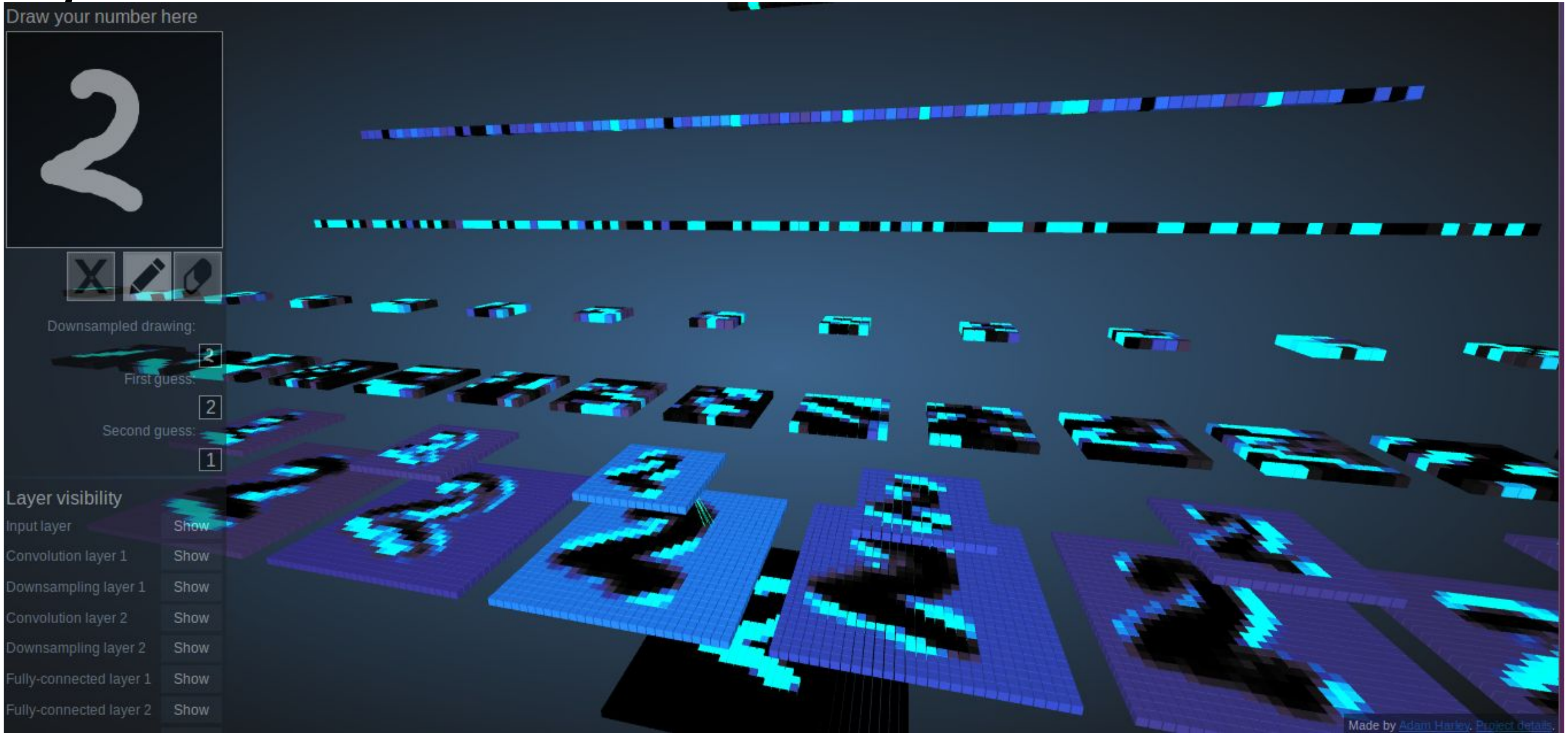
Stride: the step size your receptive field moves



padding

https://adeshpande3.github.io/A-Beginner's-Guide-To-Understanding-Convolutional-Neural-Networks-Part-2/

Layer structure



http://scs.ryerson.ca/~aharley/vis/conv/