

CONVOLUTIONAL NEURAL NETWORKS

1	0	1
0	1	0
1	0	1

•

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

→

4	3	4
2	4	3
2	3	4

filter / kernel

image

convolved feature

4	3	4
2	4	3
2	3	4

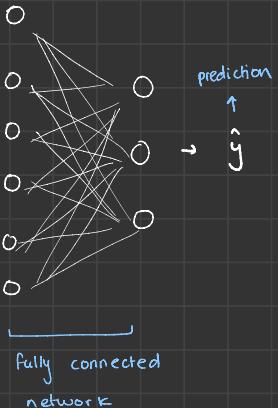
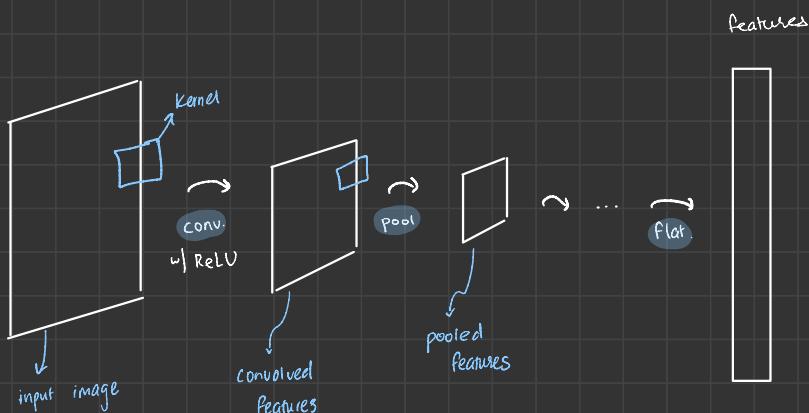
↓
2x2 filter

max pooling →

4	4
4	4

N=3 P=0 F=2 stride = 1
 $\Rightarrow \text{output size} = \frac{N + 2P - F}{\text{stride}} + 1 = \frac{3 + 2(0) - 2}{1} + 1 = 2$ ✓

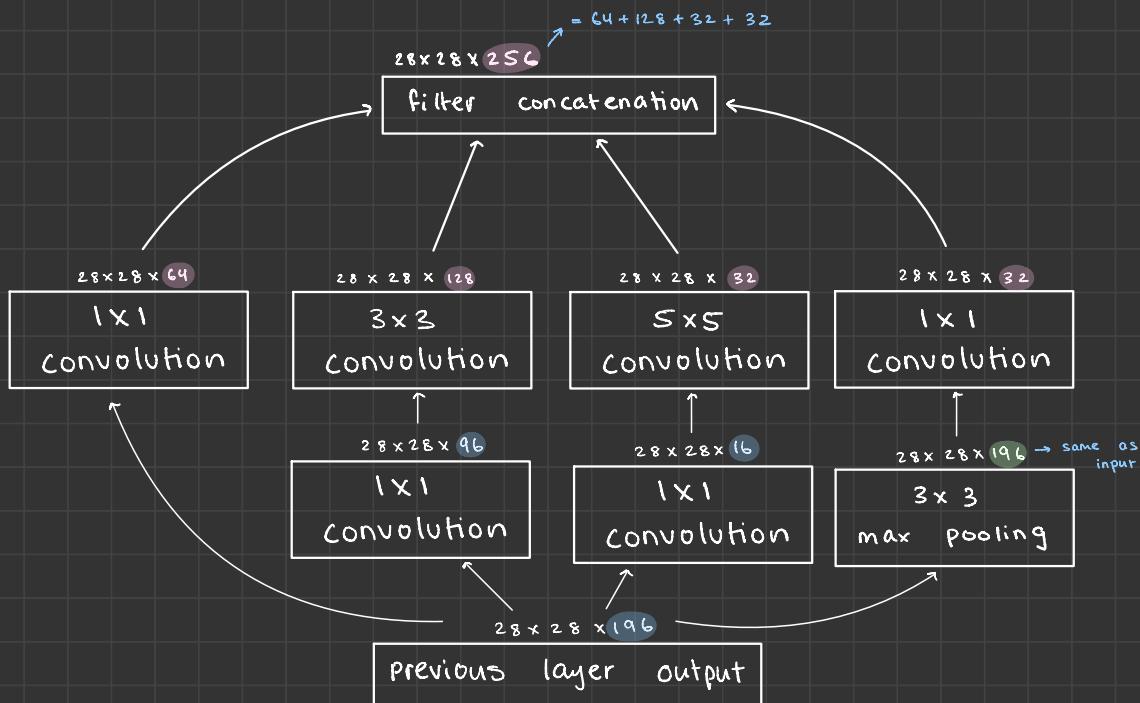
for any convolution / pooling, output size = $\frac{N + 2P - F}{\text{stride}} + 1$



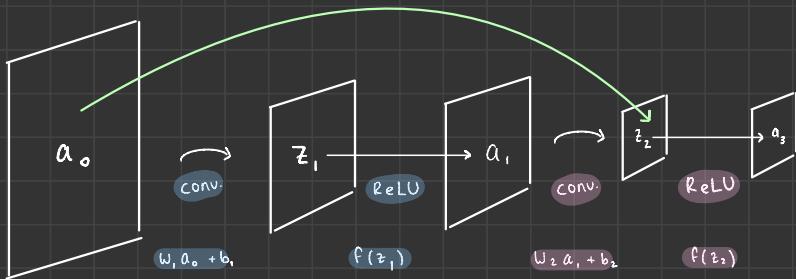
- example:
- 3×3 input
 - 1 row / column padding
 - 2×2 kernel
 - stride = 1

$$\Rightarrow \text{output} = \frac{3 + 2(1) - 2}{1} + 1 = 4 \Rightarrow 4 \times 4 \text{ output}$$

Inception module



Skip connection in ResNet



w/out skip

$$z_1 = w_1 a_0 + b_1$$

$$a_1 = f(z_1)$$

$$z_2 = w_2 a_1 + b_2$$

$$a_3 = f(z_2)$$

$$\Rightarrow a_3 = f(w_2 a_1 + b_2)$$

w/ skip

$$z_1 = w_1 a_0 + b_1$$

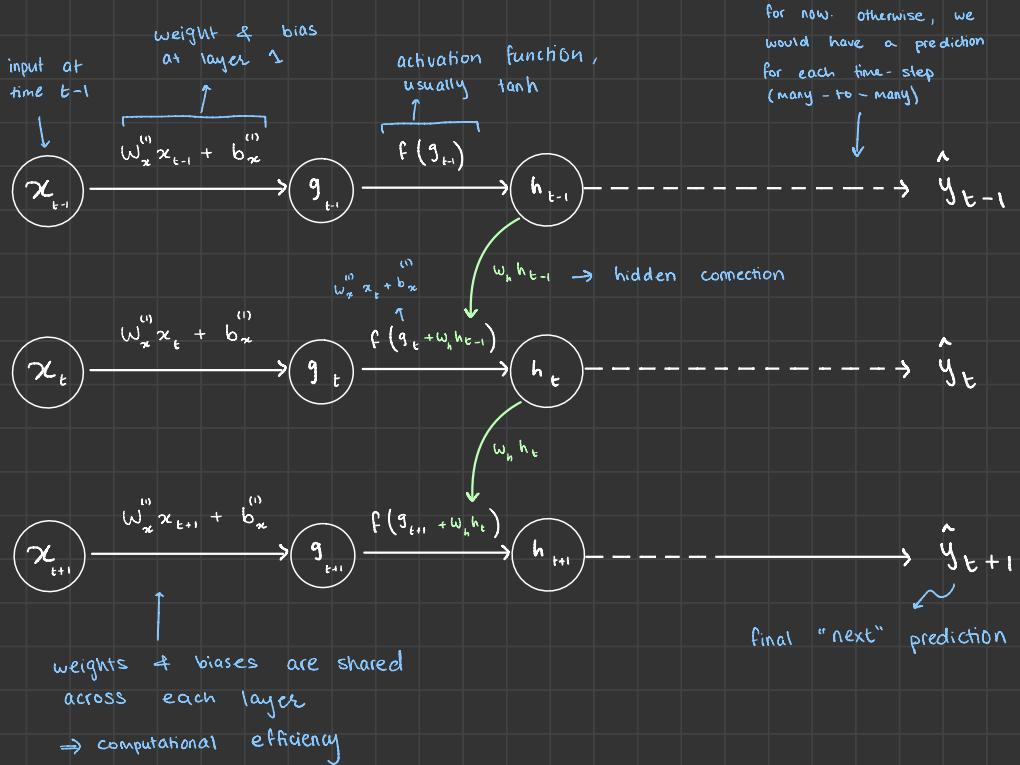
$$a_1 = f(z_1)$$

$$z_2 = w_2 a_1 + b_2$$

$$a_3 = f(z_2 + a_0)$$

$$\Rightarrow a_3 = f(w_2 a_1 + b_2 + a_0)$$

recurrent neural networks



many-to-one:

forward pass

$$h_t = f(w_x x_t + w_h h_{t-1} + b)$$

backpropagation through time → to update w_h

$$w_h := w_h - \alpha \frac{\partial L}{\partial w_h}$$

$$\frac{dL}{dW_n} = \sum_{t=1}^T \boxed{\frac{dL_t}{W_n}}$$

$$\frac{dL_T}{W_n} = \frac{dL_T}{d\hat{y}_T} \cdot \frac{d\hat{y}_T}{dh_T} \cdot \boxed{\frac{dh_T}{dh_t}} \cdot \frac{dh_t}{dW_n}$$

$$\frac{dh_T}{dh_t} = \frac{dh_T}{dh_{T-1}} \cdot \frac{dh_{T-1}}{dh_{T-2}} \dots \cdot \boxed{\frac{dh_{t+1}}{dh_t}} = \prod_{a=t}^{T-1} \frac{dh_{a+1}}{dh_a}$$

$$\frac{dh_{t+1}}{dh_t} = \frac{d}{dh_t} \tanh (W_x x_{t+1} + W_h h_t + b)$$

$$= \text{diag} \left(\tanh (W_x x_{t+1} + W_h h_t + b) \circ W_h \right)$$

since each $\frac{dh_{t+1}}{dh_t}$ in $\prod_{a=t}^{T-1} \frac{dh_{a+1}}{dh_a}$ has a W_h :

$$W_h^{(1)} \cdot W_h^{(2)} \cdot W_h^{(3)} \dots = W_h^\alpha$$

taking eigenvalue decomposition of W_h :

$$W_h = T \Sigma T^{-1}$$

$$W_h^\alpha = T \boxed{\Sigma}^\alpha T^{-1}$$

if $\Sigma > 1 \Rightarrow$ exploding gradient
 if $\Sigma < 1 \Rightarrow$ vanishing gradient] this directly affects $\frac{dL}{W_h}$
 in the update rule

gated recurrent unit

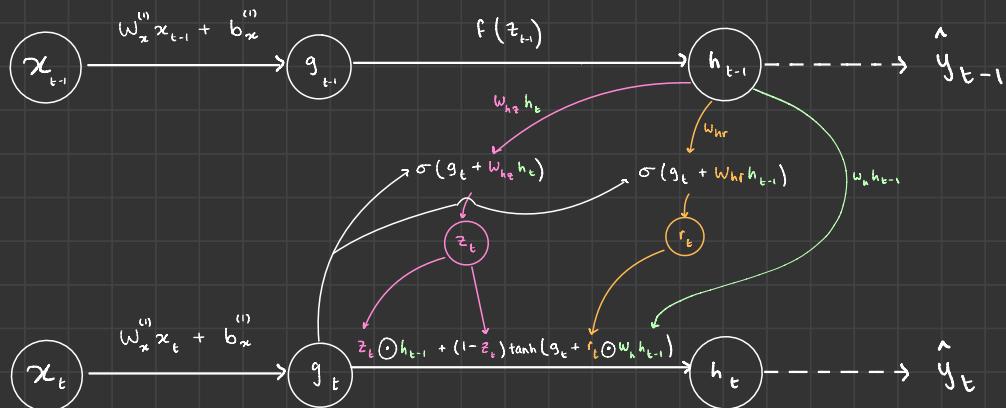
previously: $h_t = \tanh(w_x x_t + w_h h_{t-1} + b)$

now: element-wise multiplication (hadamard product)

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \cdot \tanh(w_x x_t + r_t \odot w_h h_{t-1} + b)$$

$$z_t = \sigma(w_x x_t + w_{hz} h_{t-1} + b) \quad \text{long-term dependancies}$$

$$r_t = \sigma(w_x x_t + w_{hr} h_{t-1} + b) \quad \text{short-term dependancies}$$



three cases:

- 1 if $z_t = 0 \Rightarrow$ only $r_t \Rightarrow$ only short-term memory
- 2 if $r_t = 0 \Rightarrow$ only $z_t \rightarrow$ only long-term memory
- 3 if $z_t = 1 \rightarrow h_t = h_{t-1} \Rightarrow$ no memory