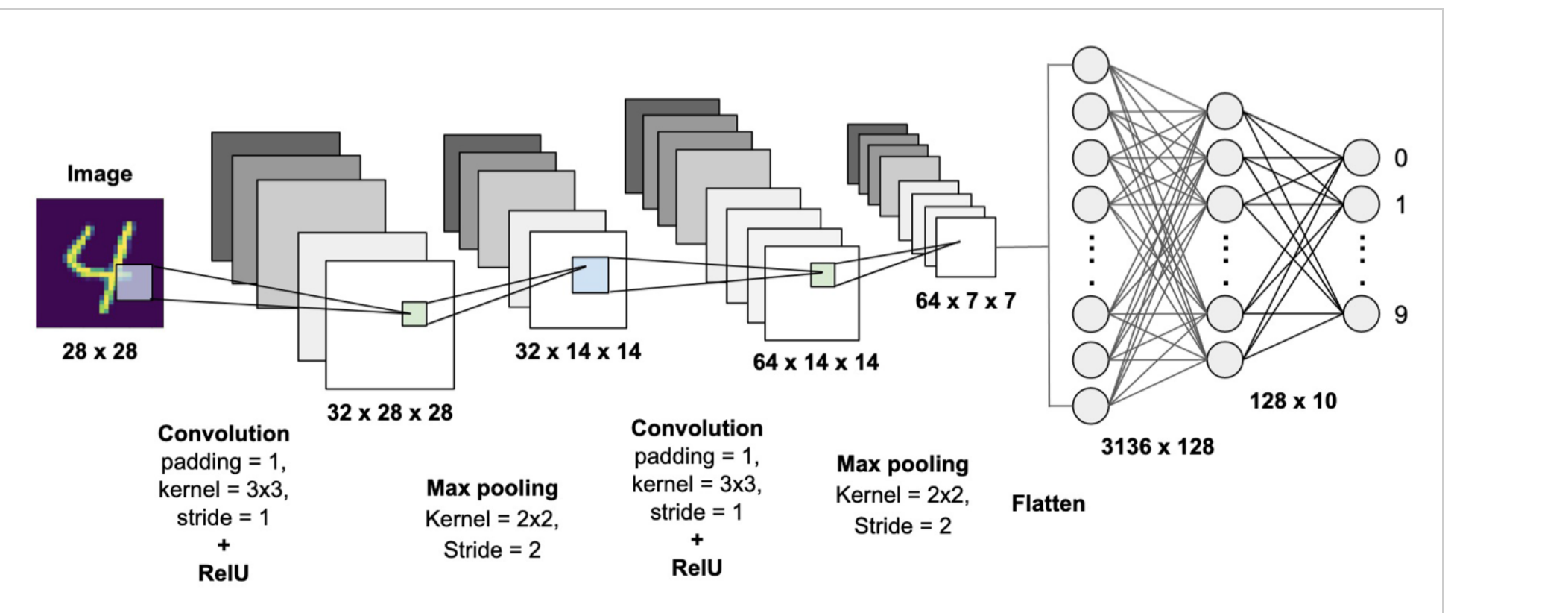
Derivation

Monday, 3 August 2020 4:23 PM



For FC layers:

Forward pass –

Input layer = $\hat{x}$

$z_1 := \hat{x}$
for $i = 1, 2, 3, \ldots N-1$:
$\quad z_{i+1} := f_i(W_i z_i + b_i)$

(also, let $a_i = W_i z_i + b_i$)

Backward pass –

Output vector – z_N

Loss – $L(\hat{y}, \theta)$

$$\frac{\partial L(\hat{y}, \theta)}{\partial z_i^n} = \delta_i^n$$

eg. $\frac{\partial L(\hat{y}, \theta)}{\partial z_N^1} = \delta_N^1$,

and $\frac{\partial L(\hat{y}, \theta)}{\partial z_N} = [\delta_N^1, \delta_N^2, \ldots \delta_N^n]$

To prove

$\varepsilon_i = \frac{\partial L(\hat{y}, \theta)}{\partial a_i} = \delta_i \circ f'(a_i)$ – ①

$\delta_{i-1} = W_i^T \cdot \varepsilon_i$ – ②

$dB_i = \varepsilon_i$ – ③

$dW_i = \varepsilon_i \otimes z_{i+1}$ – ④

where,

$\nabla L(\hat{y}, \theta) = [dB_1, dB_2, dB_3, \ldots dB_{N-1}, dW_1, dW_2, dW_3 \ldots dW_{N-1}]$

and the backprop algorithm,

compute $\delta_{N-1} := \frac{\partial L(\hat{y}, \theta)}{\partial z_N}$
$\varepsilon_{N-1} := \delta_{N-1} \circ f'_{N-1}(a_{N-1})$
for $i = N-1, N-2, \ldots 1$:
$\quad dB_i := \varepsilon_i$
$\quad dW_i := \varepsilon_i \otimes z_{i+1}$
$\quad \delta_{i-1} := W_i^T \cdot \varepsilon_i$
$\quad \varepsilon_{i-1} := \delta_{i-1} \circ f'(a_{i-1})$

① $\frac{\partial L(\hat{y}, \theta)}{\partial a_i} = \frac{\partial L(\hat{y}, \theta)}{\partial z_i} \cdot \frac{\partial z_i}{\partial a_i}$

$= \delta_i f'(a_i)$ ∎

② ε_i is the derivative of loss function w.r.t input of neurons in i^{th} layers.

We want δ_{i-1} which is derivative of loss function with respect to output of neurons in $i-1^{th}$ layer.

Consider z_{i-1}^1 (first neuron in layer i) and a_i.

$$a_i = W_{i-1} z_{i-1} + b_{i-1}$$

$$a_i^1 = W_{i-1}^1 \cdot z_{i-1} + b_{i-1}^1$$

$$\frac{\partial a_i^1}{\partial z_{i-1}^1} = \frac{\partial W_{i-1}^1 z_{i-1} + b_{i-1}^1}{\partial z_{i-1}^1}$$

$\frac{\partial a_i^1}{\partial z_{i-1}^1} = W_{i-1}^{11}$ and in general,

$$\frac{\partial a_i^n}{\partial z_{i-1}^k} = W_{i-1}^{nk}$$

$$\frac{\partial L(\hat{y}, \theta)}{\partial z_{i-1}^k} = \sum_{j=1}^{n} \frac{\partial L(\hat{y}, \theta)}{\partial a_i^j} \cdot \frac{\partial a_i^j}{\partial z_{i-1}^k}$$

$$= \sum_{j=1}^{n} \varepsilon_i^j W_{i-1}^{jk}$$

$$\frac{\partial L(\hat{y}, \theta)}{\partial z_{i-1}} = \begin{bmatrix} \sum_{j=1}^{n} \varepsilon_i^j W_{i-1}^{j1} \\ \sum_{j=1}^{n} \varepsilon_i^j W_{i-1}^{j2} \\ \sum_{j=1}^{n} \varepsilon_i^j W_{i-1}^{j3} \\ \vdots \\ \sum_{j=1}^{n} \varepsilon_i^j W_{i-1}^{jk} \end{bmatrix}$$

where $\sum_{j=1}^{n} \varepsilon_i^j W_{i-1}^{jk}$ is the dot product of ε_i and k^{th} column of W_{i-1} $\Rightarrow \sum_{j=1}^{n} \varepsilon_i^j W_{i-1}^{jk} = \varepsilon_i \cdot W_{i-1}^{T\,1}$

$\therefore \frac{\partial L(\hat{y}, \theta)}{\partial z_{i-1}} = W^T \varepsilon_i$ ∎

③ $\frac{\partial L(\hat{y}, \theta)}{b_i} = \frac{\partial L(\hat{y}, \theta)}{\partial a_i} \cdot \frac{\partial a_i}{\partial b_i}$

$= \varepsilon_i \cdot 1$

$= \varepsilon_i$ ∎

④ $\frac{\partial L(\hat{y}, \theta)}{\partial W_{i-1}^{ab}} = \varepsilon_i^b \cdot z_{i-1}^a$

Expanding gives the following:

$$\frac{\partial L(\hat{y}, \theta)}{\partial W_{i-1}} = \begin{bmatrix} \varepsilon_i^1 z_{i-1}^1, & \varepsilon_i^2 z_{i-1}^1, & \cdots & \varepsilon_i^b z_{i-1}^1 \\ \varepsilon_i^1 z_{i-1}^2 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \varepsilon_i^1 z_{i-1}^a & \cdots & \cdots & \varepsilon_i^b \cdot z_{i-1}^a \end{bmatrix}$$

$= \varepsilon_i \otimes z_{i-1}$ ∎