2.

$$\frac{\partial y_L}{\partial b_i} = \frac{\partial y_L}{\partial y_{L-1}} \cdot \frac{\partial y_{L-1}}{\partial y_{L-2}} \cdots \frac{\partial y_{i+1}}{\partial y_i} \cdot \frac{\partial y_i}{\partial \tilde{y}_i} \cdot \frac{\partial \tilde{y}_i}{\partial b_i}$$

$$= A_L \prod_{k=i+1}^{L-1} \left[\mathrm{diag}\left(\sigma'(\tilde{y}_k)\right) A_k \right] \cdot \mathrm{diag}\left(\sigma'(y_i)\right)$$

$$\leq A_L \cdot A_{L-1} \cdots A_j \cdots A_{i+1} \qquad \text{"}\leq\text{" denotes all elements have smaller absolute value than RHS}$$

$$\frac{\partial y_L}{\partial A_i} = \frac{\partial y_L}{\partial y_{L-1}} \cdot \frac{\partial y_{L-1}}{\partial y_{L-2}} \cdots \frac{\partial y_{i+1}}{\partial y_i} \cdot \frac{\partial y_i}{\partial \tilde{y}_i} \cdot \frac{\partial \tilde{y}_i}{\partial A_i}$$

$$= A_L \prod_{k=i+1}^{L-1} \left[\mathrm{diag}\left(\sigma'(\tilde{y}_k)\right) A_k \right] \cdot \mathrm{diag}(\sigma'(\tilde{y}_i)) \; y_{i-1}^T$$

$$\leq A_L \cdot A_{L-1} \cdots A_j \cdots A_{i+1} \; y_{i-1}^T$$

Since A_j is small and other matrices are not too large,

the two gradients become small.

$$\frac{\partial y_L}{\partial b_i} = \frac{\partial y_L}{\partial y_{L-1}} \cdot \frac{\partial y_{L-1}}{\partial y_{L-2}} \cdots \frac{\partial y_{i+1}}{\partial y_i} \cdot \frac{\partial y_i}{\partial \tilde{y}_i} \cdot \frac{\partial \tilde{y}_i}{\partial b_i}$$

$$= A_L \prod_{k=i+1}^{L-1} \left[\mathrm{diag}\left(\sigma'(\tilde{y}_k)\right) A_k \right] \cdot \mathrm{diag}\left(\sigma'(y_i)\right)$$

$$\leq A_L \cdots A_{j+1} \, \mathrm{diag}(\sigma'(\tilde{y}_j)) A_j \; A_{j-1} \cdots A_{i+1}$$

$$\frac{\partial y_L}{\partial A_i} = \frac{\partial y_L}{\partial y_{L-1}} \cdot \frac{\partial y_{L-1}}{\partial y_{L-2}} \cdots \frac{\partial y_{i+1}}{\partial y_i} \cdot \frac{\partial y_i}{\partial \tilde{y}_i} \cdot \frac{\partial \tilde{y}_i}{\partial A_i}$$

$$= A_L \prod_{k=i+1}^{L-1} \left[\mathrm{diag}\left(\sigma'(\tilde{y}_k)\right) A_k \right] \cdot \mathrm{diag}(\sigma'(\tilde{y}_i)) \; y_{i-1}^T$$

$$\leq A_L \cdot A_{L-1} \cdots A_{j+1} \, \mathrm{diag}(\sigma' y_j) \, A_j \, A_{j-1} \cdots A_{i+1} \; y_{i-1}^T$$

$|\tilde{y}_j|$ large $\Rightarrow$ $\sigma'(\tilde{y}_j)$ small

and the other matrices are not to large.

Hence, the two gradients become small.

□

3.

$$\theta_I^1 = \theta^0 - \alpha g^0 + \beta(\theta^0 - \theta^{-1})$$

$$= \theta^0 - \alpha g^0 = \theta^0 - \alpha(g^0 + \beta v^0) = \theta_{II}^1$$

Suppose that $\theta_I^K = \theta_{II}^K$ for some $K \geq 1$.

That is, $\theta_I^K = \theta^{K-1} - \alpha g^{K-1} + \beta(\theta^{K-1} - \theta^{K-2})$

$$= \theta^{K-1} - \alpha v^K = \theta_{II}^K$$

Then, $\theta^K - \theta^{K-1} = -\alpha v^K$

$$\theta_I^{K+1} = \theta^K - \alpha g^K + \beta(\theta^K - \theta^{K-1})$$

$$= \theta^K - \alpha g^K + \beta(-\alpha v^K)$$

$$= \theta^K - \alpha(g^K + \beta v^K) = \theta_{II}^{K+1}$$

$\therefore$ By induction, form I and II are identical.

□

4.

i) $y_1[K, i, j]$ depends on $X[:, i-2:i+2, j-2:j+2]$.

$i, j \leq 0$ or $i, j \geq 225$ is due to zero padding and can be ignored.

ii) $y_2[K, i, j]$ depends on $y_1[K, 2i-1:2i, 2j-1:2j]$,

So on $X[:, 2i-3:2i+2, 2j-3:2j+2]$ by i).

iii) $y_3[K, i, j]$ depends on $y_2[:, 2i-3:2i+2, 2j-3:2j+2]$ by ii),

So on $X[:, 4i-9:4i+6, 4j-9:4j+6]$ by ii).

□

5.

(i) First module has

$$((1\times1)\times256+1)\times128$$
$$+\ ((3\times3)\times256+1)\times128$$
$$+\ ((5\times5)\times256+1)\times128 \quad = 1,147,264 \text{ trainable parameters.}$$

Second module has

$$((1\times1)\times256+1)\times128$$
$$+\ ((1\times1)\times256+1)\times64 + ((3\times3)\times64+1)\times192$$
$$+\ ((1\times1)\times256+1)\times64 + ((5\times5)\times64+1)\times96$$
$$+\ ((1\times1)\times256+1)\times64 \quad = 346,720 \text{ trainable parameters.}$$

The second module has much less trainable parameters.

(ii)

	module 1	module 2
addition	$256\times32\times32\times128$ $+\ 256\times3\times3\times32\times32\times192$ $+\ 256\times5\times5\times32\times32\times96$ $= 1,115,684,864$	$256\times32\times32\times128$ $+\ 256\times32\times32\times64 + 64\times3\times3\times32\times32\times192$ $+\ 256\times32\times32\times64 + 64\times5\times5\times32\times32\times96$ $+\ 256\times32\times32\times64$ $= 354,418,688$
multiplication	$256\times32\times32\times128$ $+\ 256\times3\times3\times32\times32\times192$ $+\ 256\times5\times5\times32\times32\times96$ $= 1,115,684,864$	$256\times32\times32\times128$ $+\ 256\times32\times32\times64 + 64\times3\times3\times32\times32\times192$ $+\ 256\times32\times32\times64 + 64\times5\times5\times32\times32\times96$ $+\ 256\times32\times32\times64$ $= 354,418,688$
Activation function evaluation	$32\times32\times128$ $+\ 32\times32\times192$ $+\ 32\times32\times96$ $= 425,984$	$32\times32\times128$ $32\times32\times64 + 32\times32\times192$ $32\times32\times64 + 32\times32\times96$ $32\times32\times64$ $= 622,592$

There are much more add, multiplying in module 1.
Activation function evaluation is slightly more in module 2. □