Q1 Given (B, W, H, input),

output should be $\left(B, \frac{W - k_1 + 2p_1}{s_1} + 1, \frac{H - k_2 + 2p_2}{s_2} + 1,\right.$ output_channels)

(By formula of $\text{out} = \frac{\text{in} - K - 2P}{S} + 1$)

Q2.

$$\frac{\partial L}{\partial \gamma} = \sum_i \frac{\partial L}{\partial y_i} \frac{\partial y_i}{\partial \gamma}$$

$$\frac{\partial L}{\partial \gamma} = \sum_i \frac{\partial L}{\partial y_i} \hat{x}_i$$

$$\frac{\partial L}{\partial \beta} = \sum_i \frac{\partial L}{\partial y_i} \frac{\partial y_i}{\partial \beta}$$

$$\frac{\partial L}{\partial \beta} = \sum_i \frac{\partial L}{\partial y_i}$$

$$\frac{\partial L}{\partial \hat{x}} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial \hat{x}}$$

$$\frac{\partial L}{\partial \hat{x}} = \frac{\partial L}{\partial y} \gamma$$

$$\frac{\partial L}{\partial \sigma_B} = \sum_i \frac{\partial L}{\partial y_i} \frac{\partial y_i}{\partial \hat{x}_i} \frac{\partial \hat{x}_i}{\partial \sigma_B}$$

$$\frac{\partial L}{\partial \sigma} = \sum_i \frac{\partial L}{\partial \hat{x}_i} (x_i - \mu)\left(-\frac{(\sigma_B + \varepsilon)^{-3/2}}{2}\right)$$

$$\frac{\partial L}{\partial \mu_B} = \frac{\partial L}{\partial \hat{x}} \frac{\partial \hat{x}}{\partial \mu_B} + \frac{\partial L}{\partial \sigma_B} \frac{\partial \sigma_B}{\partial \mu_B}$$

$$\frac{\partial L}{\partial \mu_B} = \sum_i \frac{\partial L}{\partial \hat{x}_i} \frac{-1}{\sqrt{\sigma_B + \varepsilon}} + \frac{\partial L}{\partial \sigma_B} \frac{-2\sum_i (x_i - \mu)}{N}$$

$$\Rightarrow \frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial \hat{x}_i}\frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial L}{\partial \sigma_B}\frac{\partial \sigma_B}{\partial x_i} + \frac{\partial L}{\partial \mu_B}\frac{\partial \mu_B}{\partial x_i}$$

$$\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial \hat{x}_i}\frac{1}{\sqrt{\sigma_B+\varepsilon}} + \frac{\partial L}{\partial \sigma_B}\frac{\partial(x_i-\mu)}{N} + \frac{\partial L}{\partial \mu_B}\frac{1}{N}$$

when $\gamma_k = \sqrt{Var(x_k)}$ and $\beta_k = E(x_k)$,
we can produce the identity result,
making batch normalization able to have
the ability of identity transform.

Q3. Softmax & cross entropy

Define the following:
partial derivatives of the ith output versus
the jth input of softmax(), is:

$$\frac{\partial \hat{y}_i}{\partial z_j} = \frac{\partial \frac{e^{z_i}}{\sum_{k=1}^{N} e^{z_k}}}{\partial z_j}$$

By the quotient differentiation rule,
when $f(x) = \frac{g(x)}{h(x)}$, $f'(x) = \frac{g'(x)h(x) - h'(x)g(x)}{(h(x))^2}$

Here, $g_i = e^{z_i}$, $h_i = \sum_{k=1}^{N} e^{z_k}$

Hence,

$$\frac{\partial g_i}{\partial z_j} = \frac{\partial e^{z_i}}{\partial z_j} = \begin{cases} 0, & i \neq j \\ e^{z_i}, & i = j \end{cases}$$

$$\frac{\partial h_i}{\partial z_j} = \frac{\partial \sum_{k=1}^{N} e^{z_k}}{\partial z_j} = e^{z_j}$$

When $i \neq j$,

$$\frac{\partial \hat{y}_i}{\partial z_j} = \frac{0 \times \sum_{k=1}^{N} e^{z_k} - e^{z_i} e^{z_j}}{\left(\sum_{k=1}^{N} e^{z_k}\right)^2}$$

$$= - \frac{e^{z_j}}{\sum_{k=1}^{N} e^{z_k}} \times \frac{e^{z_i}}{\sum_{k=1}^{N} e^{z_k}}$$

$$= - \hat{y}_i \hat{y}_j$$

When $i = j$, $\frac{\partial \hat{y}_i}{\partial z_j} = \cancel{\frac{e^{z_i} \sum_{k=1}^{N} e^{z_k} - e^{z_i} e^{z_j}}{\left(\sum_{k=1}^{N} e^{z_k}\right)^2}}$

$$= \frac{e^{z_i}}{\sum_{k=1}^{N} e^{z_k}} \times \frac{\sum_{k=1}^{N} e^{z_k} - e^{z_j}}{\sum_{k=1}^{N} e^{z_k}}$$

$$= \hat{y}_i (1 - \hat{y}_j)$$

$$\Rightarrow \quad \frac{\partial \hat{y}_i}{\partial z_j} = \begin{cases} -\hat{y}_i \hat{y}_j, & i \neq j \\ \hat{y}_i (1-\hat{y}_j), & i = j \end{cases}$$

For cross entropy:

Denote as: $L(y, \hat{y}) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$

$$\frac{\partial L}{\partial z_j} = -\sum_{i=1}^{N} y_i \frac{1}{\hat{y}_i} \frac{\partial \hat{y}_i}{\partial z_j}$$

$$= -y_j \frac{1}{\hat{y}_i} (\hat{y}_j (1-\hat{y}_j)) - \sum_{i \neq j} y_i \frac{1}{\hat{y}_i} (-\hat{y}_j \hat{y}_i)$$

$$= -y_j (1-\hat{y}_j) + \sum_{i \neq j}^{N} y_i \hat{y}_j$$

$$= y_j \hat{y}_j + \sum_{i \neq j}^{N} y_i \hat{y}_j - y_j$$

$$= \sum_{i=1}^{N} y_i \hat{y}_j - y_j$$

$$= \hat{y}_j - y_j \quad \left(\because \sum_{i=1}^{N} y_i = 1 \right)$$

Hence proved.

Q4.

$$v^t = \beta_2 \cdot v^{t-1} + (1-\beta_2) \cdot (g^t)^2$$

$$v^1 = \beta_2 \cdot v^0 + (1-\beta_2)(g^0)^2$$

$$v^1 = (1-\beta_2)(g^0)^2$$

$$v^2 = \beta_2 v^1 + (1-\beta_2) \cdot (g^1)^2$$

$$v^2 = \beta_2(1-\beta_2)(g^0)^2 + (1-\beta_2)(g^1)^2$$

$$v^2 = (1-\beta_2)\left(\beta_2 (g^0)^2 + 1 \cdot (g^1)^2\right)$$

$$\vdots$$

$$V_t = (1-\beta_2) \sum_{i=1}^{t} \beta_2^{t-i} (g^i)^2$$

similarily,

$$m_t = (1-\beta_1) \sum_{i=1}^{t} \beta_1^{t-i} g^i$$

4b). When $\eta = \eta_0 \cdot t^{-\frac{1}{2}}$,

In Adam, $w^t \approx w^{t-1} - \frac{\eta_0 \cdot t^{-\frac{1}{2}}}{\sqrt{\hat{v}_t}} \hat{m}_t$

If $\beta_1 = 0$,

Using the result of (a),

$$m_t = 0 \cdot m^{t-1} + (1-0) \cdot g^t$$

$$\hat{m}_t \approx g^t$$

$$m_t = g^t,$$

$$\hat{v}^t \approx \frac{\beta_2}{1-\beta_2^t} v^{t-1} + \frac{1-\beta_2}{1-\beta_2^t} (g^t)^2$$

if $\beta_2 = 1 - 1/t$,

We can rewrite $\hat{v}_t$ as:

$$\hat{v}^t = \frac{1}{t} \sum_{i=1}^{t} g_i^2$$

Then, if $\eta = \eta_0 \cdot t^{-1/2}$,

$$w^t = w^{t-1} - \frac{\eta_0}{\sqrt{t}\left(\sqrt{\frac{1}{t}\sum_{i=1}^{t} g_i^2}\right)} g^t$$

$$w^t = w^{t-1} - \frac{\eta_0}{\sqrt{\sum_{i=1}^{t} g_i^2}} g^t$$

which is AdaGrad. Hence proved.