1 Batch Normalization

1. For each batch B of size m, and each dimension in $x = (x^1, x^2, \cdots, x^d)$,
   $\mu_{B,d} = \frac{1}{m} \sum_{i=1}^{m} x_i^d$
   $\sigma_{B,d}^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i^d - \mu_{B,d})^2$
   Normalization:
   $\hat{x}_i^d = \frac{x_i^d - \mu_{B,d}}{\sqrt{\sigma_{B,d}^2 + \epsilon}}$
   where ϵ is a constant to avoid dividing by 0.
2. $y_i^d = \gamma^k \hat{x}_i^d + \beta^k$
   $\frac{\partial E}{\partial \gamma^k} = \frac{\partial E}{\partial y_i^k} * \frac{\partial y_i^k}{\partial \gamma^k} = \frac{\partial E}{\partial y_i^k} * \hat{x}_i^k$
   $\frac{\partial E}{\partial \beta^k} = \frac{\partial E}{\partial y_i^k} * \frac{\partial y_i^k}{\partial \beta^k} = \frac{\partial E}{\partial y_i^k}$

2 Convolution

1. $(5-2) \times (5-2) = 3 \times 3 = 9$
2. Values after forward propagate

239	194	238
201	232	260
154	172	213

3. The gradient backpropagated out of this layer.

2	7	2
1	6	8
4	-1	1