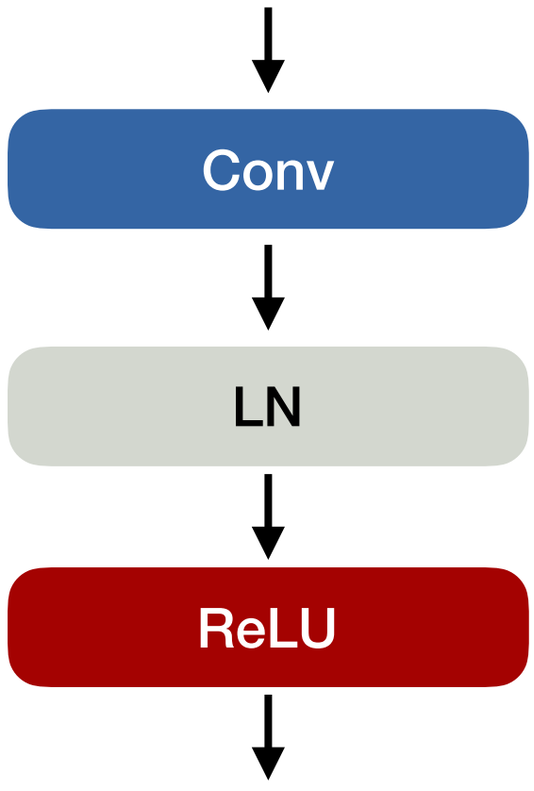
Layer normalization

Layer normalization

- Make activations zero mean and unit variance without collecting statistics across batches



Layer Normalization, Ba, J., Kiros, J. R. and Hinton, G., arXiv preprint arXiv: 1607.06450, 2016

Layer normalization

- Normalize by image-wise mean μ_k and standard deviation σ_k

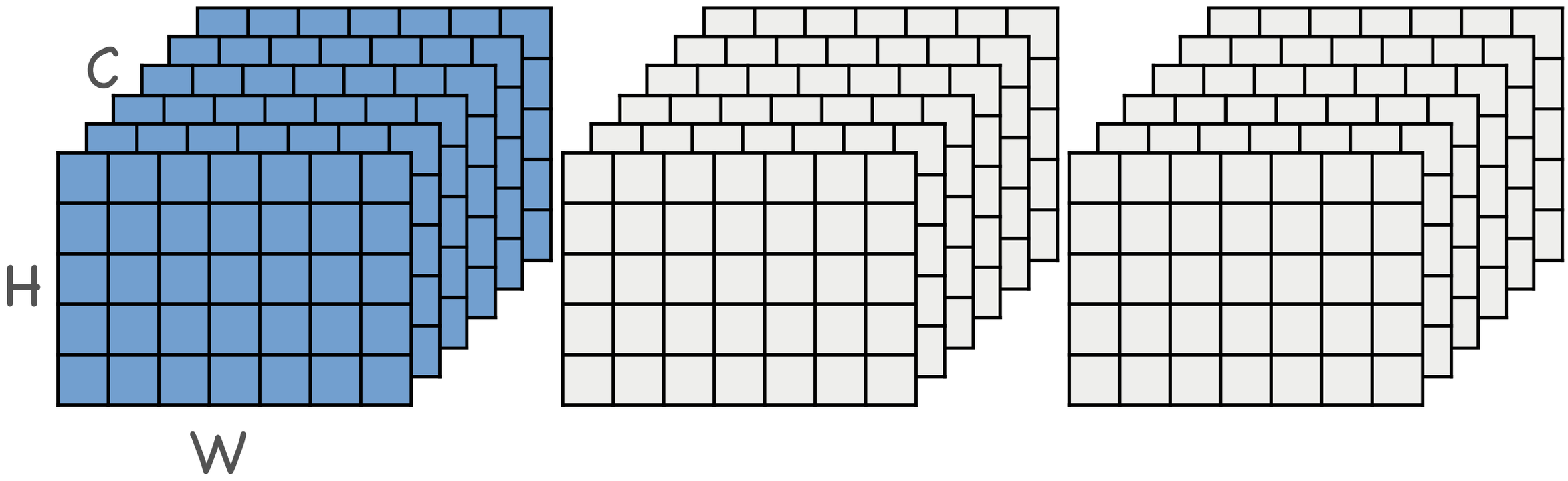
$$\mathbf{Z} \in \mathbb{R}^{B \times W \times H \times C}$$

$$\frac{\mathbf{Z}_{k,x,y,c} - \mu_k}{\sigma_k}$$

$$\mu_k = \frac{1}{WHC} \sum_{x,y,c} \mathbf{Z}_{k,x,y,c}$$

$$\sigma_k^2 = \frac{1}{WHC} \sum_{x,y,c} (\mathbf{Z}_{k,x,y,c} - \mu_k)^2$$

What does layer normalization do?



Comparison to batch norm

- No summary statistics
  - Training and testing are the same
- Works well for sequence models
- Does not scale activations individually