Where to add normalization?

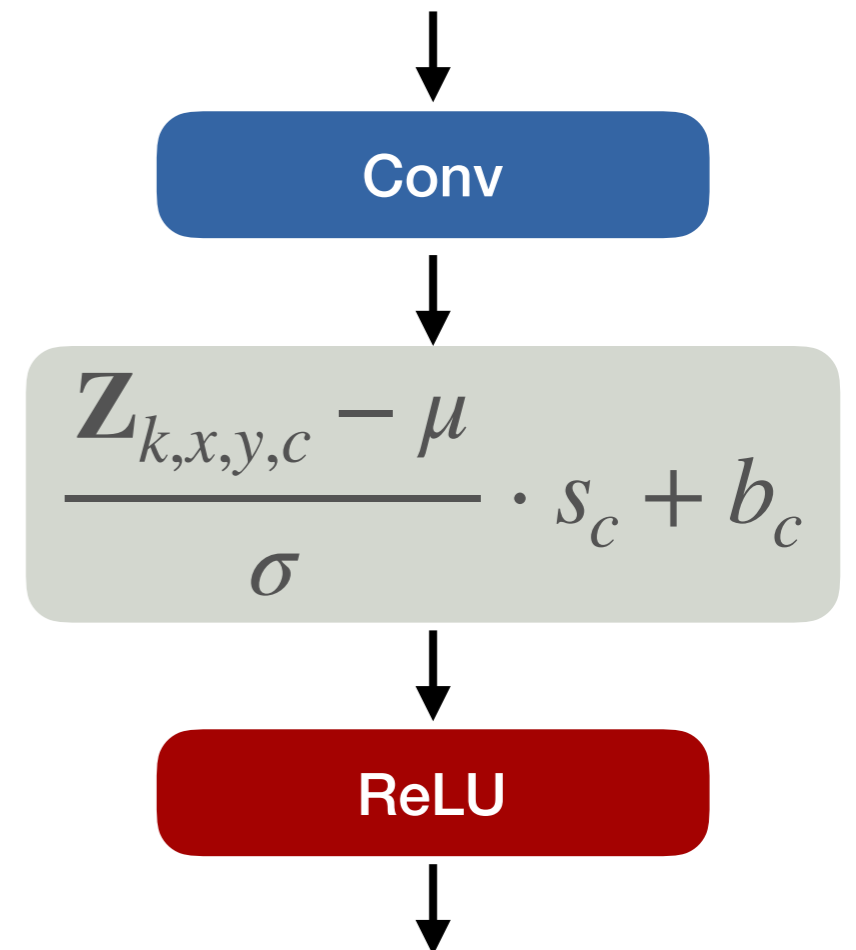# Where to add normalization?

- Option A

  - After convolution

- Option B

  - After ReLU (non-linearity)

Option A

Conv

↓

BN

↓

ReLU

Option B
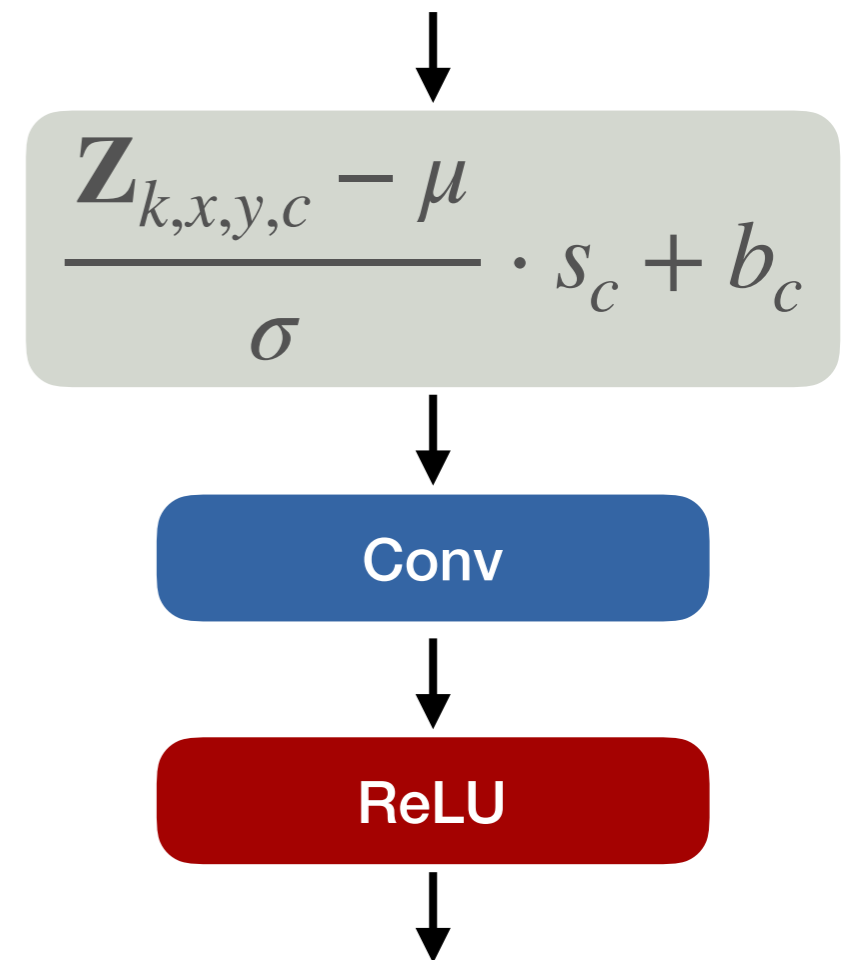
ReLU

↓

BN

↓
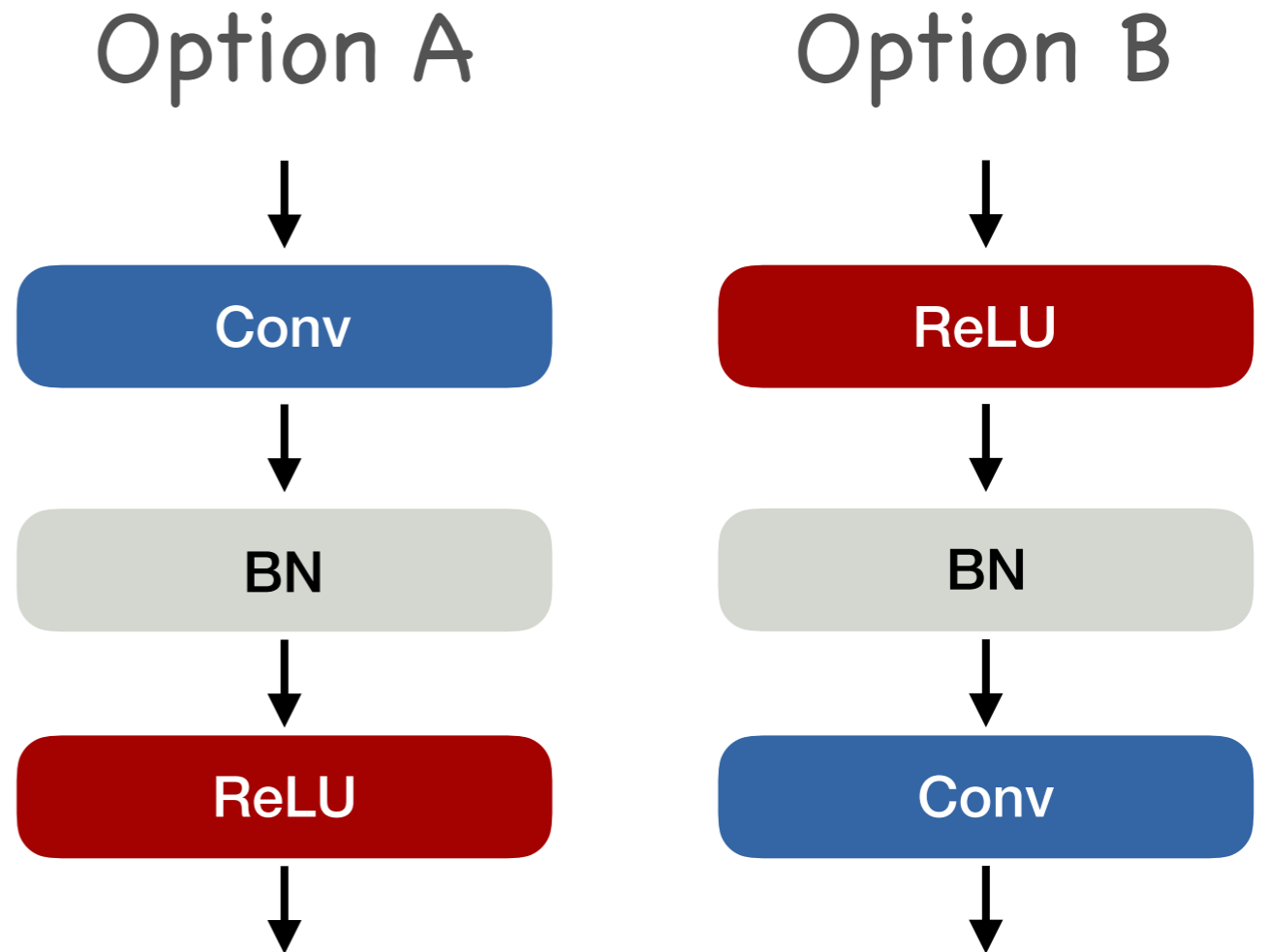
Conv

Option A

- No bias in conv

- Activations are zero mean

  - Half of activations zeroed out in ReLU

- Solution:

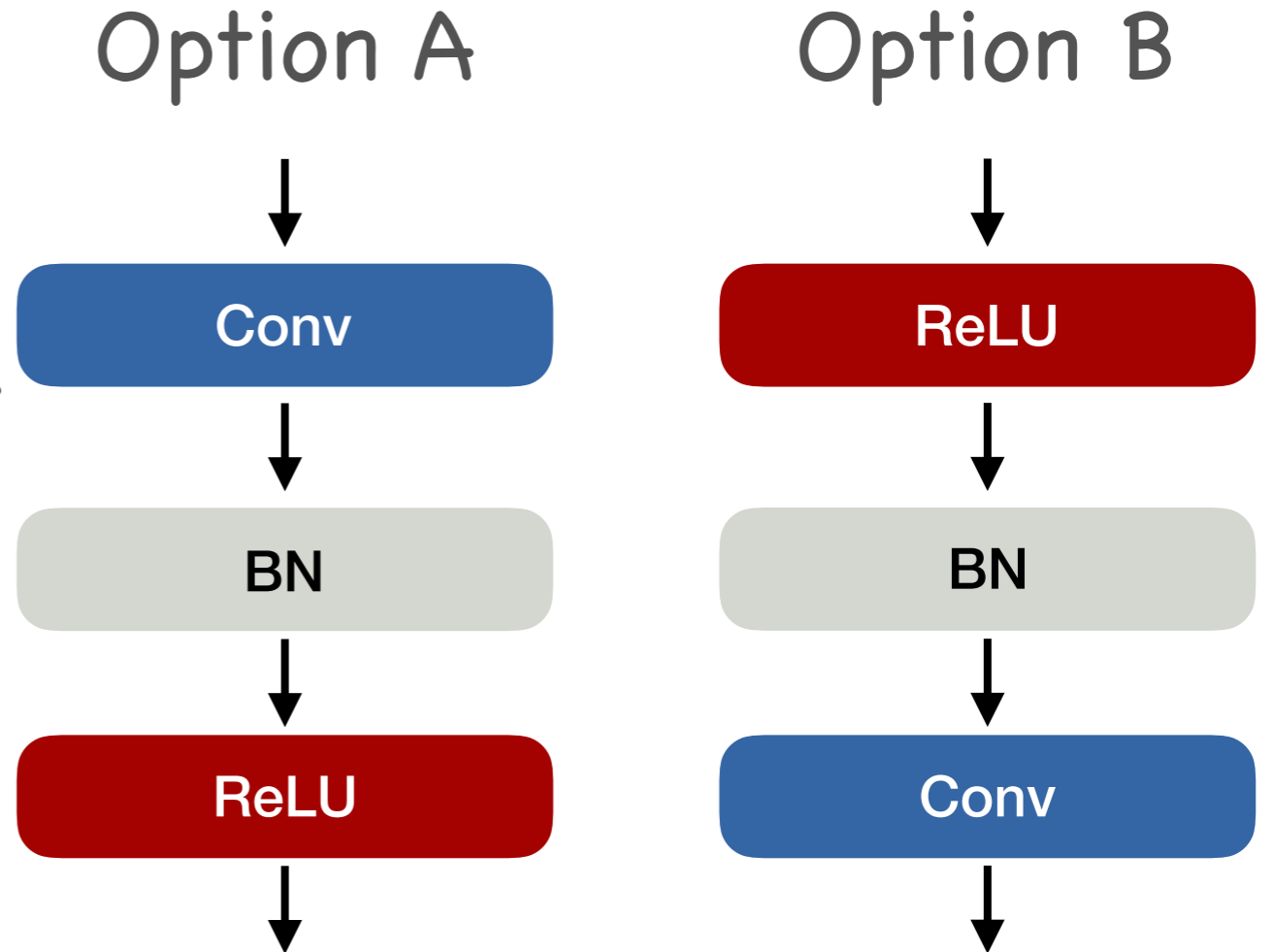 - Learn a scale s_c and bias b_c after norm

Conv

$$\frac{\mathbf{Z}_{k,x,y,c} - \mu}{\sigma} \cdot s_c + b_c$$

ReLU

Option B

$$\frac{\mathbf{Z}_{k,x,y,c} - \mu}{\sigma} \cdot s_c + b_c$$

Conv

ReLU

- Scale s_c and bias b_c optional

Where to add normalization?

- Both work

- Option A is more popular

- Option B is easier

  - Scale and bias optional

  - Conv unchanged

Option A

Conv
↓
BN
↓
ReLU

Option B

ReLU
↓
BN
↓
Conv

Where not to add batch norm?

- After fully connected layers

  - Mean and standard deviation estimates too unstable

Why does normalization work?

- Regularizes the network

- Handles badly scaled weights

- Single parameter to learn scale s_c

Conv

$$\frac{\mathbf{Z}_{k,x,y,c} - \mu}{\sigma} \cdot s_c + b_c$$

ReLU