Method 2

Compressing the Convolution

$$\mathcal{K}(i,j,l,s,t) = \sum_{r_4=1}^{R_4} \sum_{r_5=1}^{R_5} \mathcal{C}(i,j,l,r_4,r_5)\, \boldsymbol{U}^{(4)}(s,r_4)\, \boldsymbol{U}^{(5)}(t,r_5)$$

Using this in the convolution:

$$\mathcal{Y}(f',h',w',t) = \sum_{i=1}^{D_F} \sum_{j=1}^{D_H} \sum_{l=1}^{D_W} \sum_{s=1}^{S} \sum_{r_4=1}^{R_4} \sum_{r_5=1}^{R_5} \mathcal{C}(i,j,l,r_4,r_5)\, \boldsymbol{U}^{(4)}(s,r_4)\, \boldsymbol{U}^{(5)}(t,r_5)\, \mathcal{X}(f_i,h_j,w_l,s)$$

Rearranging:

$$\mathcal{Y}(f',h',w',t) = \sum_{r_5=1}^{R_5} \boldsymbol{U}^{(5)}(t,r_5) \underbrace{\sum_{i=1}^{D_F} \sum_{j=1}^{D_H} \sum_{l=1}^{D_W} \sum_{r_4=1}^{R_4} \mathcal{C}(i,j,l,r_4,r_5) \underbrace{\sum_{s=1}^{S} \boldsymbol{U}^{(4)}(s,r_4)\, \mathcal{X}(f_i,h_j,w_l,s)}_{\mathcal{Q}(f_i,h_j,w_l,r_4)}}_{\mathcal{Q}'(f',h',w',r_5)}$$

Method 2

Compressing the Convolution

$$\mathcal{Q}(f,h,w,r_4) = \sum_{s=1}^{S} \boldsymbol{U}^{(4)}(s,r_4)\ \mathcal{X}(f,h,w,s)$$

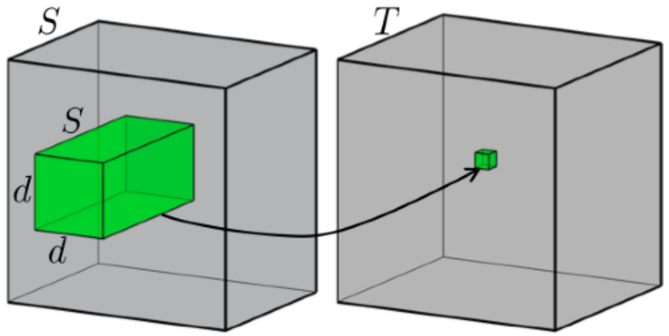
→ $1 \times 1 \times 1$ convolution with S input channels and R_4 output channels

$$\mathcal{Q}'(f',h',w',r_5) = \sum_{i=1}^{D_F}\sum_{j=1}^{D_H}\sum_{l=1}^{D_W}\sum_{r_4=1}^{R_4} \mathcal{C}(i,j,l,r_4,r_5)\mathcal{Q}(f_i,h_j,w_l,r_4)$$

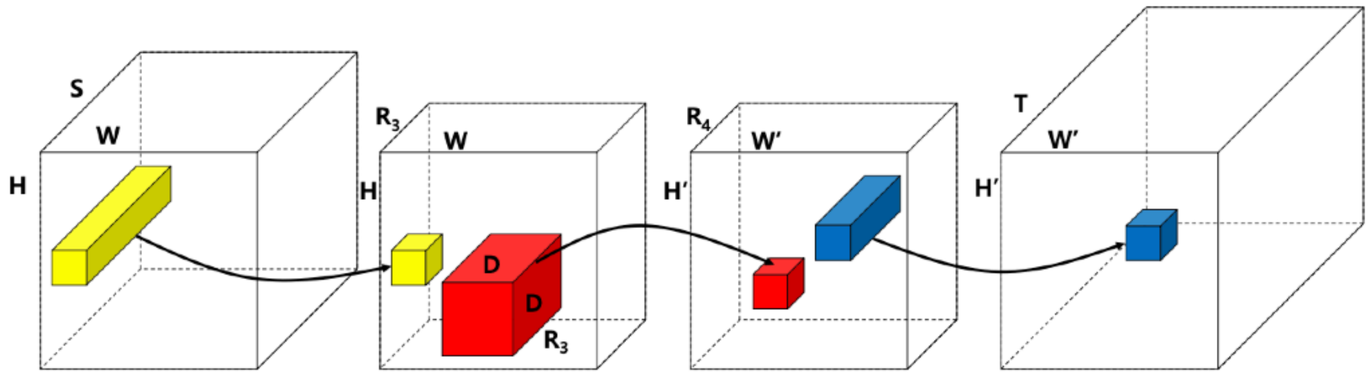
→ $D_F \times D_H \times D_W$ convolution with R_4 input channels and R_5 output channels

$$\mathcal{Y}(f',h',w',t) = \sum_{r_5=1}^{R_5} \boldsymbol{U}^{(5)}(t,r_5)\ \mathcal{Q}'(f',h',w',r_5)$$

→ $1 \times 1 \times 1$ convolution with R_5 input channels and T output channels



Original convolution



Sequence of convolutions of the compressed version

Illustration of the 2D case with different Rs due to lower dimensions