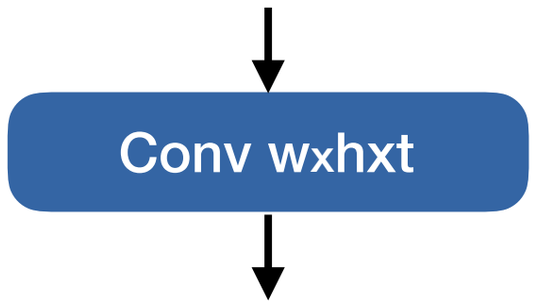
3D convolutions

3D convolutions

- Convolution across space and time
  - Input video is a 4D tensor
    - time
    - width, height
    - color channels
- 3D kernel
  - time, width, height

Formal definition

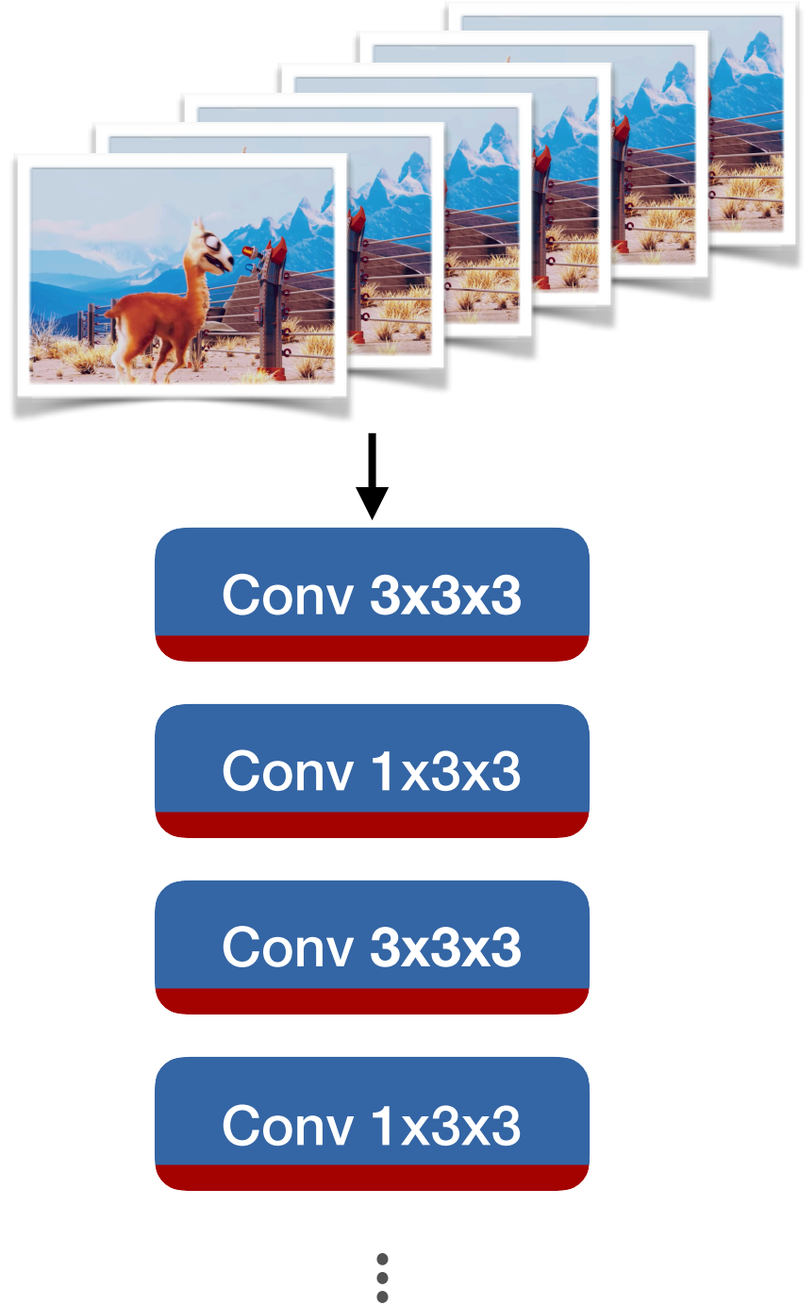
- Input: $\mathbf{X} \in \mathbb{R}^{T \times H \times W \times C_1}$
- Kernel: $\mathbf{w} \in \mathbb{R}^{t \times h \times w \times C_1 \times C_2}$
- Bias: $\mathbf{b} \in \mathbb{R}^{C_2}$
- Output: $\mathbf{Z} \in \mathbb{R}^{\left(\frac{T-t+2p_t}{s_t}+1\right) \times \left(\frac{H-h+2p_h}{s_h}+1\right) \times \left(\frac{W-w+2p_w}{s_w}+1\right) \times C_2}$



$$\mathbf{Z}_{d,a,b,c} = \mathbf{b}_c + \sum_{l=0}^{t} \sum_{i=0}^{h} \sum_{j=0}^{w} \sum_{k=0}^{C_1} \mathbf{X}_{d+l,a+i,b+j,c+k} \mathbf{w}_{l,i,j,k}$$

3D CNNs

- Take a image CNN
  - Replace some (not all) layers with 3D conv



3D CNNs - Issues

- 3D convolutions are too large
- Slow
- Too many parameters