Introduction to Deep Learning

CNN BackPropagation Fall 2021

Delivered by: Rukayat Sadiq

Backpropagation in CNNs

- In the backward pass, we get the loss gradient with respect to the next layer
- In CNNs the loss gradient is computed w.r.t the input and also w.r.t the filter.

Convolution Backprop with single Stride

- To understand the computation of loss gradient w.r.t input, let us use the following example:

- Horizontal and vertical stride = 1

X_{11}	X_{12}	X_{13}
X_{21}	X_{22}	X_{23}
X_{31}	X_{32}	X_{33}

Input **X**

F_{11}	F_{12}
F_{21}	F_{22}

Filter **F**

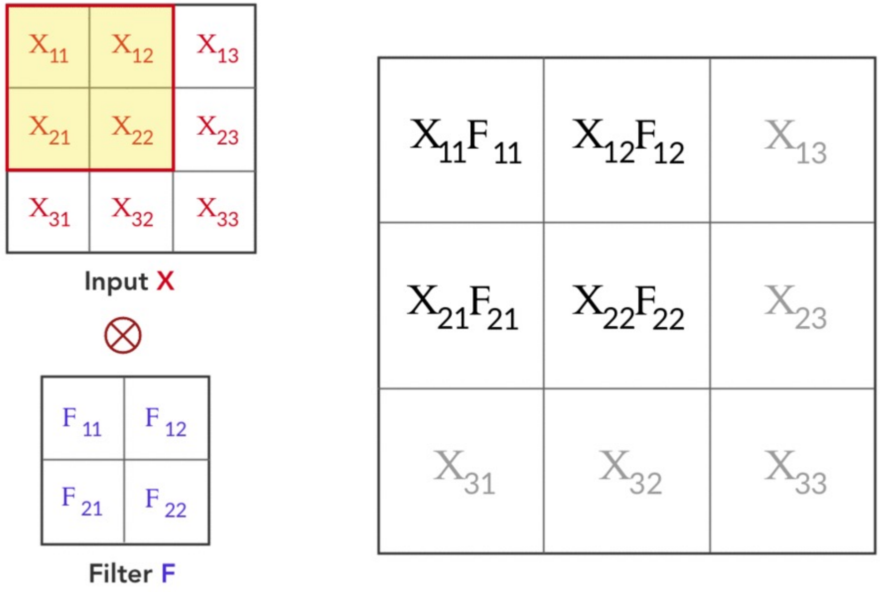
Convolution Forward Pass

- Convolution between Input X and Filter F, gives us an output O. This can be represented as:

$$\underbrace{\begin{bmatrix} O_{11} & O_{12} \\ O_{21} & O_{22} \end{bmatrix}}_{\text{Output } O} = \text{Convolution}\left(\underbrace{\begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix}}_{\text{Input } X}, \underbrace{\begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}}_{\text{Filter } F} \right)$$

Convolution Forward Pass

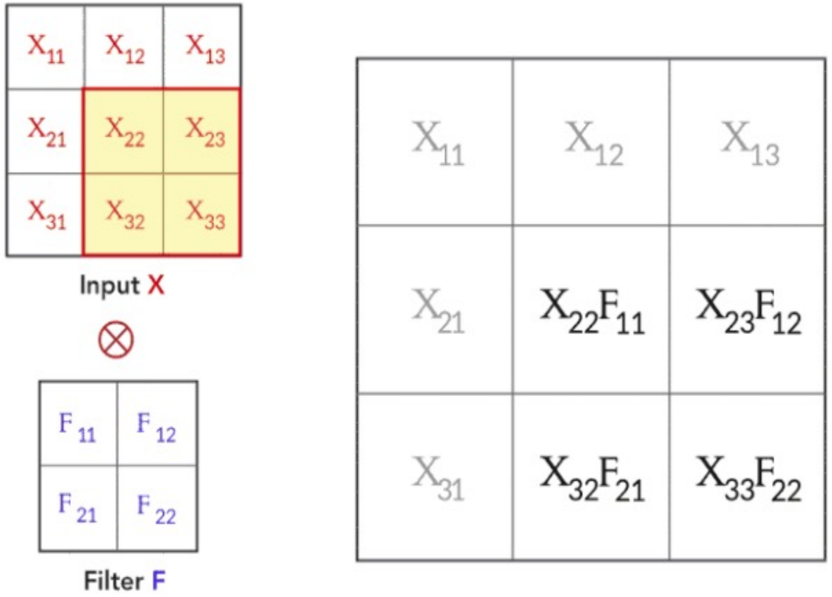
- Convolution between Input X and Filter F, gives us an output O. This can be represented as:



$$O_{11} = X_{11}F_{11} + X_{12}F_{12} + X_{21}F_{21} + X_{22}F_{22}$$

Convolution Forward Pass

- Convolution between Input X and Filter F, gives us an output O. This can be represented as:



$$O_{11} = X_{11}F_{11} + X_{12}F_{12} + X_{21}F_{21} + X_{22}F_{22}$$

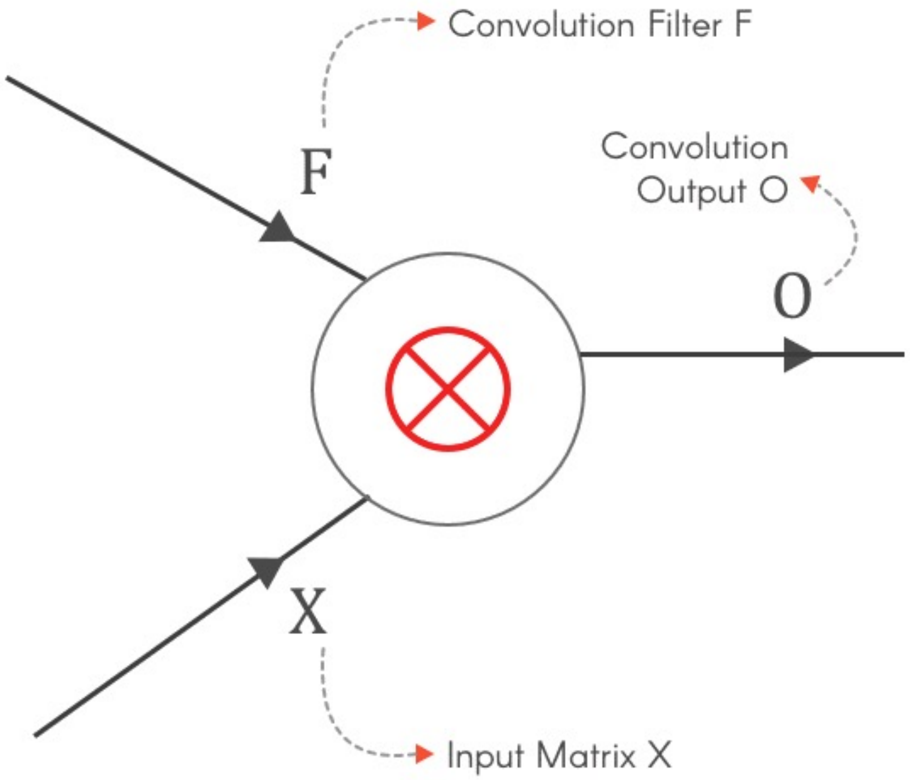
$$O_{12} = X_{12}F_{11} + X_{13}F_{12} + X_{22}F_{21} + X_{23}F_{22}$$

$$O_{21} = X_{21}F_{11} + X_{22}F_{12} + X_{31}F_{21} + X_{32}F_{22}$$

$$O_{22} = X_{22}F_{11} + X_{23}F_{12} + X_{32}F_{21} + X_{33}F_{22}$$

Loss gradient

- We want to calculate the gradients wrt to input ‘X’ and filter ‘F’



Loss gradient w.r.t the filter

We can use the chain rule to obtain the gradient wrt the filter as shown in the equation.

$$\underbrace{\frac{\partial L}{\partial F}}_{\text{Gradient to update Filter F}} = \underbrace{\frac{\partial L}{\partial O}}_{\text{Loss Gradient from previous layer}} * \underbrace{\frac{\partial O}{\partial F}}_{\text{Local Gradients}}$$

For every element of F

$$\frac{\partial L}{\partial F_i} = \sum_{k=1}^{M} \frac{\partial L}{\partial O_k} * \frac{\partial O_k}{\partial F_i}$$

Loss gradient w.r.t the filter

We can expand the chain rule summation as:

For every element of F

$$\frac{\partial L}{\partial F_i} = \sum_{k=1}^{M} \frac{\partial L}{\partial O_k} * \frac{\partial O_k}{\partial F_i}$$

$$\frac{\partial L}{\partial F_{11}} = \frac{\partial L}{\partial O_{11}} * \frac{\partial O_{11}}{\partial F_{11}} + \frac{\partial L}{\partial O_{12}} * \frac{\partial O_{12}}{\partial F_{11}} + \frac{\partial L}{\partial O_{21}} * \frac{\partial O_{21}}{\partial F_{11}} + \frac{\partial L}{\partial O_{22}} * \frac{\partial O_{22}}{\partial F_{11}}$$

$$\frac{\partial L}{\partial F_{12}} = \frac{\partial L}{\partial O_{11}} * \frac{\partial O_{11}}{\partial F_{12}} + \frac{\partial L}{\partial O_{12}} * \frac{\partial O_{12}}{\partial F_{12}} + \frac{\partial L}{\partial O_{21}} * \frac{\partial O_{21}}{\partial F_{12}} + \frac{\partial L}{\partial O_{22}} * \frac{\partial O_{22}}{\partial F_{12}}$$

$$\frac{\partial L}{\partial F_{21}} = \frac{\partial L}{\partial O_{11}} * \frac{\partial O_{11}}{\partial F_{21}} + \frac{\partial L}{\partial O_{12}} * \frac{\partial O_{12}}{\partial F_{21}} + \frac{\partial L}{\partial O_{21}} * \frac{\partial O_{21}}{\partial F_{21}} + \frac{\partial L}{\partial O_{22}} * \frac{\partial O_{22}}{\partial F_{21}}$$

$$\frac{\partial L}{\partial F_{22}} = \frac{\partial L}{\partial O_{11}} * \frac{\partial O_{11}}{\partial F_{22}} + \frac{\partial L}{\partial O_{12}} * \frac{\partial O_{12}}{\partial F_{22}} + \frac{\partial L}{\partial O_{21}} * \frac{\partial O_{21}}{\partial F_{22}} + \frac{\partial L}{\partial O_{22}} * \frac{\partial O_{22}}{\partial F_{22}}$$

Loss gradient w.r.t the filter

- Replacing the local gradients of the filter i.e, $\frac{\partial O_i}{\partial F_i}$, we get this:

$$\begin{bmatrix} \frac{\partial L}{\partial F_{11}} & \frac{\partial L}{\partial F_{12}} \\ \frac{\partial L}{\partial F_{21}} & \frac{\partial L}{\partial F_{22}} \end{bmatrix} = \text{Convolution}\left(\begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix}, \begin{bmatrix} \frac{\partial L}{\partial O_{11}} & \frac{\partial L}{\partial O_{12}} \\ \frac{\partial L}{\partial O_{21}} & \frac{\partial L}{\partial O_{22}} \end{bmatrix} \right)$$

where

$$\begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix} = \text{Input X} \qquad \begin{bmatrix} \frac{\partial L}{\partial O_{11}} & \frac{\partial L}{\partial O_{12}} \\ \frac{\partial L}{\partial O_{21}} & \frac{\partial L}{\partial O_{22}} \end{bmatrix} = \frac{\partial L}{\partial O}$$

Loss gradient from previous layer

$$\frac{\partial L}{\partial F_{11}} = \frac{\partial L}{\partial O_{11}} * X_{11} + \frac{\partial L}{\partial O_{12}} * X_{12} + \frac{\partial L}{\partial O_{21}} * X_{21} + \frac{\partial L}{\partial O_{22}} * X_{22}$$

$$\frac{\partial L}{\partial F_{12}} = \frac{\partial L}{\partial O_{11}} * X_{12} + \frac{\partial L}{\partial O_{12}} * X_{13} + \frac{\partial L}{\partial O_{21}} * X_{22} + \frac{\partial L}{\partial O_{22}} * X_{23}$$

$$\frac{\partial L}{\partial F_{21}} = \frac{\partial L}{\partial O_{11}} * X_{21} + \frac{\partial L}{\partial O_{12}} * X_{22} + \frac{\partial L}{\partial O_{21}} * X_{31} + \frac{\partial L}{\partial O_{22}} * X_{32}$$

$$\frac{\partial L}{\partial F_{22}} = \frac{\partial L}{\partial O_{11}} * X_{22} + \frac{\partial L}{\partial O_{12}} * X_{23} + \frac{\partial L}{\partial O_{21}} * X_{32} + \frac{\partial L}{\partial O_{22}} * X_{33}$$

Loss gradient w.r.t the filter

- If you closely look at it, this represents an operation we are quite familiar with. We can represent it as a **convolution operation between input X** and loss gradient **∂L/∂O as shown below:**

$$\begin{bmatrix} \frac{\partial L}{\partial F_{11}} & \frac{\partial L}{\partial F_{12}} \\ \frac{\partial L}{\partial F_{21}} & \frac{\partial L}{\partial F_{22}} \end{bmatrix} = \text{Convolution}\left(\begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix}, \begin{bmatrix} \frac{\partial L}{\partial O_{11}} & \frac{\partial L}{\partial O_{12}} \\ \frac{\partial L}{\partial O_{21}} & \frac{\partial L}{\partial O_{22}} \end{bmatrix} \right)$$

where

$$\begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix} = \text{Input X}$$

$$\begin{bmatrix} \frac{\partial L}{\partial O_{11}} & \frac{\partial L}{\partial O_{12}} \\ \frac{\partial L}{\partial O_{21}} & \frac{\partial L}{\partial O_{22}} \end{bmatrix} = \frac{\partial L}{\partial O}$$

Loss gradient from previous layer

$$\frac{\partial L}{\partial F_{11}} = \frac{\partial L}{\partial O_{11}} * X_{11} + \frac{\partial L}{\partial O_{12}} * X_{12} + \frac{\partial L}{\partial O_{21}} * X_{21} + \frac{\partial L}{\partial O_{22}} * X_{22}$$

$$\frac{\partial L}{\partial F_{12}} = \frac{\partial L}{\partial O_{11}} * X_{12} + \frac{\partial L}{\partial O_{12}} * X_{13} + \frac{\partial L}{\partial O_{21}} * X_{22} + \frac{\partial L}{\partial O_{22}} * X_{23}$$

$$\frac{\partial L}{\partial F_{21}} = \frac{\partial L}{\partial O_{11}} * X_{21} + \frac{\partial L}{\partial O_{12}} * X_{22} + \frac{\partial L}{\partial O_{21}} * X_{31} + \frac{\partial L}{\partial O_{22}} * X_{32}$$

$$\frac{\partial L}{\partial F_{22}} = \frac{\partial L}{\partial O_{11}} * X_{22} + \frac{\partial L}{\partial O_{12}} * X_{23} + \frac{\partial L}{\partial O_{21}} * X_{32} + \frac{\partial L}{\partial O_{22}} * X_{33}$$

Loss gradient w.r.t the input

- If you closely look at it, this represents an operation we are quite familiar with. We can represent it as a **convolution operation between input X** and loss gradient **∂L/∂O as shown below:**

For every element of X_i

$$\frac{\partial L}{\partial X_i} = \sum_{k=1}^{M} \frac{\partial L}{\partial O_k} * \frac{\partial O_k}{\partial X_i}$$

Loss gradient w.r.t the input

- Similarly, we can expand the chain rule summation for the gradient with respect to the input. After substituting the local gradients i.e $\frac{\partial O_i}{\partial X_i}$, we have:

$$\frac{\partial L}{\partial X_{11}} = \frac{\partial L}{\partial O_{11}} * F_{11}$$

$$\frac{\partial L}{\partial X_{12}} = \frac{\partial L}{\partial O_{11}} * F_{12} + \frac{\partial L}{\partial O_{12}} * F_{11}$$

$$\frac{\partial L}{\partial X_{13}} = \frac{\partial L}{\partial O_{12}} * F_{12}$$

$$\frac{\partial L}{\partial X_{21}} = \frac{\partial L}{\partial O_{11}} * F_{21} + \frac{\partial L}{\partial O_{21}} * F_{11}$$

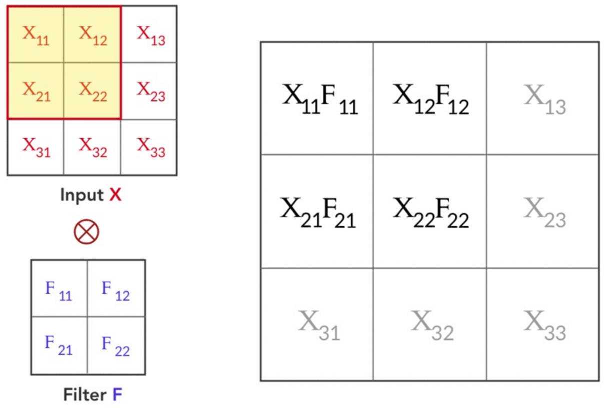
$$\frac{\partial L}{\partial X_{22}} = \frac{\partial L}{\partial O_{11}} * F_{22} + \frac{\partial L}{\partial O_{12}} * F_{21} + \frac{\partial L}{\partial O_{21}} * F_{12} + \frac{\partial L}{\partial O_{22}} * F_{11}$$

$$\frac{\partial L}{\partial X_{23}} = \frac{\partial L}{\partial O_{12}} * F_{22} + \frac{\partial L}{\partial O_{22}} * F_{12}$$

$$\frac{\partial L}{\partial X_{31}} = \frac{\partial L}{\partial O_{21}} * F_{21}$$

$$\frac{\partial L}{\partial X_{32}} = \frac{\partial L}{\partial O_{21}} * F_{22} + \frac{\partial L}{\partial O_{22}} * F_{21}$$

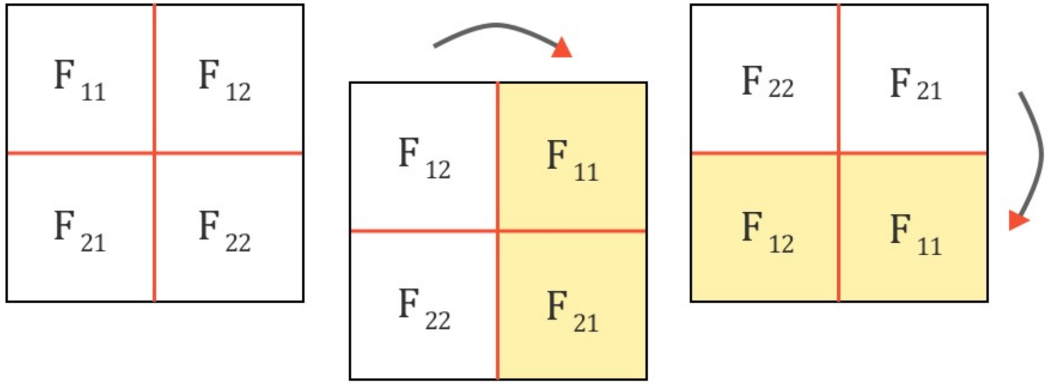
$$\frac{\partial L}{\partial X_{33}} = \frac{\partial L}{\partial O_{22}} * F_{22}$$



$$O_{11} = X_{11}F_{11} + X_{12}F_{12} + X_{21}F_{21} + X_{22}F_{22}$$

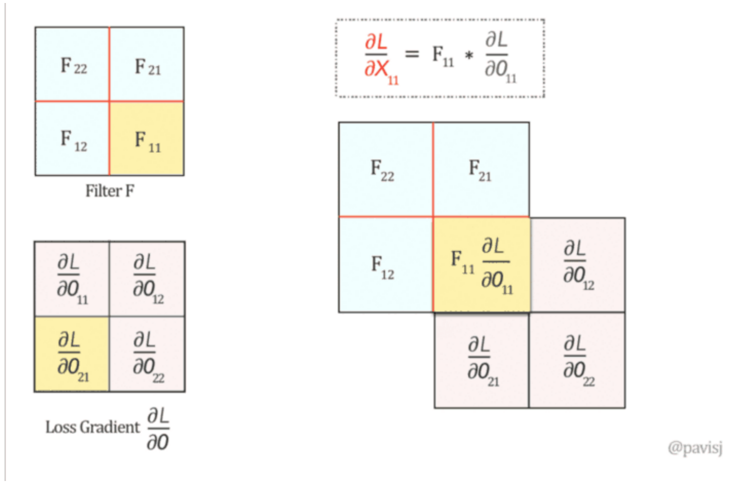
Loss gradient w.r.t the input

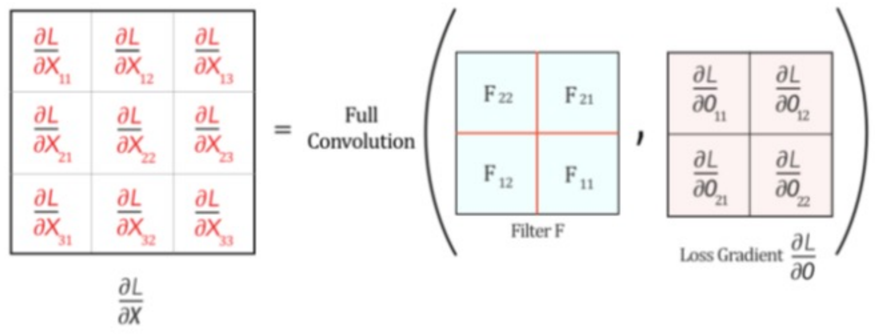
- First, let us rotate the Filter F by 180 degrees. This is done by flipping it first vertically and then horizontally.



Loss gradient w.r.t the input

- We see that the loss gradient wrt the input $\frac{\partial L}{\partial X}$ is given as a full convolution between the filter and Loss gradient $\frac{\partial L}{\partial O}$.





Takeaway

- Both the Forward pass and the Backpropagation of a Convolutional layer are Convolutions

$$\frac{\partial L}{\partial F} = \text{Convolution}\left(\text{Input } X,\ \text{Loss gradient } \frac{\partial L}{\partial O}\right)$$

$$\frac{\partial L}{\partial X} = \begin{array}{c}\text{Full}\\ \text{Convolution}\end{array}\left(\begin{array}{c}180^{\circ}\text{rotated}\\ \text{Filter } F\end{array},\ \begin{array}{c}\text{Loss}\\ \text{Gradient}\end{array} \frac{\partial L}{\partial O}\right)$$

Loss gradient w.r.t the input

- To understand the computation of loss gradient w.r.t input, let us use the following example:

- > Horizontal and vertical stride = 2

W

H					
	x_{00}	x_{01}	x_{02}	x_{03}	x_{04}
	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
	x_{20}	x_{21}	x_{22}	x_{23}	x_{24}
	x_{30}	x_{31}	x_{32}	x_{33}	x_{34}
	x_{40}	x_{41}	x_{42}	x_{43}	x_{44}

Input activations
Input channels C = 1, number of images N = 1,
Image height H = 5, width = 5

S

R			
	f_{00}	f_{01}	f_{02}
	f_{10}	f_{11}	f_{12}
	f_{20}	f_{21}	f_{22}

Filter (aka kernel)
Input channels C = 1, number of filters K = 1,
Filter height R = 3, width S = 3,
stride_R = stride_S = 2

Q

P		
	y_{00}	y_{01}
	y_{10}	y_{11}

Output
Output channels K = 1, number of outputs N = 1,
Output height P = 2, width Q = 2

Recap: Forward pass

- This is how the forward pass looks like for the example:

$x_{00}f_{00}$	$x_{01}f_{01}$	$x_{02}f_{02}$	x_{03}	x_{04}
$x_{10}f_{10}$	$x_{11}f_{11}$	$x_{12}f_{12}$	x_{13}	x_{14}
$x_{20}f_{20}$	$x_{21}f_{21}$	$x_{22}f_{22}$	x_{23}	x_{24}
x_{30}	x_{31}	x_{32}	x_{33}	x_{34}
x_{40}	x_{41}	x_{42}	x_{43}	x_{44}

y_{00}	y_{01}
y_{10}	y_{11}

$$y_{00} = x_{00}f_{00} + x_{01}f_{01} + x_{02}f_{02} + x_{10}f_{10} + x_{11}f_{11} + x_{12}f_{12} + x_{20}f_{20} + x_{21}f_{21} + x_{22}f_{22}$$

Backward Pass:

- **Assumption**: we have the loss gradient w.r.t the output pixels.
- **Requirement**: calculate the loss gradient w.r.t the input activations

Loss gradients w.r.t input

$\frac{\partial L}{\partial x_{00}}$	$\frac{\partial L}{\partial x_{01}}$	$\frac{\partial L}{\partial x_{02}}$	$\frac{\partial L}{\partial x_{03}}$	$\frac{\partial L}{\partial x_{04}}$
$\frac{\partial L}{\partial x_{10}}$	$\frac{\partial L}{\partial x_{11}}$	$\frac{\partial L}{\partial x_{12}}$	$\frac{\partial L}{\partial x_{13}}$	$\frac{\partial L}{\partial x_{14}}$
$\frac{\partial L}{\partial x_{20}}$	$\frac{\partial L}{\partial x_{21}}$	$\frac{\partial L}{\partial x_{22}}$	$\frac{\partial L}{\partial x_{23}}$	$\frac{\partial L}{\partial x_{24}}$
$\frac{\partial L}{\partial x_{30}}$	$\frac{\partial L}{\partial x_{31}}$	$\frac{\partial L}{\partial x_{32}}$	$\frac{\partial L}{\partial x_{33}}$	$\frac{\partial L}{\partial x_{34}}$
$\frac{\partial L}{\partial x_{40}}$	$\frac{\partial L}{\partial x_{41}}$	$\frac{\partial L}{\partial x_{42}}$	$\frac{\partial L}{\partial x_{43}}$	$\frac{\partial L}{\partial x_{44}}$

Loss gradients w.r.t output

$\frac{\partial L}{\partial y_{00}}$	$\frac{\partial L}{\partial y_{01}}$
$\frac{\partial L}{\partial y_{10}}$	$\frac{\partial L}{\partial y_{11}}$

Backward pass:

- Each input contributes to one or more outputs. The total gradient of the loss wrt to each input pixel is computed using the formula shown

- The gradient computation is done using chain rule and partial differentiation

- i and j represent the position of a single output pixel

$$\frac{\partial \mathrm{L}}{\partial x_{mn}} = \sum_{ij} \frac{\partial L}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial x_{mn}}$$

Backward Pass example:

- Consider input x_{00} in the input shown. It contributed to the output y_{00}

x_{00}	x_{01}	x_{02}	x_{03}	x_{04}
x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
x_{20}	x_{21}	x_{22}	x_{23}	x_{24}
x_{30}	x_{31}	x_{32}	x_{33}	x_{34}
x_{40}	x_{41}	x_{42}	x_{43}	x_{44}

y_{00}	y_{01}
y_{10}	y_{11}

$$\frac{\partial L}{\partial x_{mn}} = \sum_{ij} \frac{\partial L}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial x_{mn}}$$

Consider x_{00}. What output pixels y_{ij} does it contribute to?

$$y_{00} = x_{00}f_{00} + x_{01}f_{01} + x_{02}f_{02} + x_{10}f_{10} + x_{11}f_{11} + x_{12}f_{12} + x_{20}f_{20} + x_{21}f_{20} + x_{22}f_{22}$$

We see that x_{00} only contributes to y_{00}. Also, $\frac{\partial y_{00}}{\partial x_{00}} = f_{00}$. Thus, $\frac{\partial L}{\partial x_{00}} = \frac{\partial L}{\partial y_{00}} f_{00}$

Backward Pass example:

- Input x_{01} also contributed to the output y_{00} so the loss gradient w.r.t x_{01} is computed as shown:

x_{00}	x_{01}	x_{02}	x_{03}	x_{04}
x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
x_{20}	x_{21}	x_{22}	x_{23}	x_{24}
x_{30}	x_{31}	x_{32}	x_{33}	x_{34}
x_{40}	x_{41}	x_{42}	x_{43}	x_{44}

y_{00}	y_{01}
y_{10}	y_{11}

$$\frac{\partial \mathrm{L}}{\partial x_{mn}} = \sum_{ij} \frac{\partial L}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial x_{mn}}$$

Next, consider x_{01}. What output pixels y_{ij} does it contribute to?

$$y_{00} = x_{00}f_{00} + x_{01}f_{01} + x_{02}f_{02} + x_{10}f_{10} + x_{11}f_{11} + x_{12}f_{12} + x_{20}f_{20} + x_{21}f_{20} + x_{22}f_{22}$$

Again, x_{01} only contributes to y_{00}. Also, $\frac{\partial y_{00}}{\partial x_{01}} = f_{01}$. Thus, $\frac{\partial \mathrm{L}}{\partial x_{01}} = \frac{\partial L}{\partial y_{00}} f_{01}$

Backward Pass example:

- Input x_{02} contributed to the output y_{00} and y_{01} so the loss gradient w.r.t x_{02} is computed as shown:

x_{00}	x_{01}	x_{02}	x_{03}	x_{04}
x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
x_{20}	x_{21}	x_{22}	x_{23}	x_{24}
x_{30}	x_{31}	x_{32}	x_{33}	x_{34}
x_{40}	x_{41}	x_{42}	x_{43}	x_{44}

y_{00}	y_{01}
y_{10}	y_{11}

$$\frac{\partial L}{\partial x_{mn}} = \sum_{ij} \frac{\partial L}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial x_{mn}}$$

Next, consider x_{02}. It contributes to y_{00} and y_{01}.

$$y_{00} = x_{00}f_{00} + x_{01}f_{01} + x_{02}f_{02} + x_{10}f_{10} + x_{11}f_{11} + x_{12}f_{12} + x_{20}f_{20} + x_{21}f_{20} + x_{22}f_{22}$$

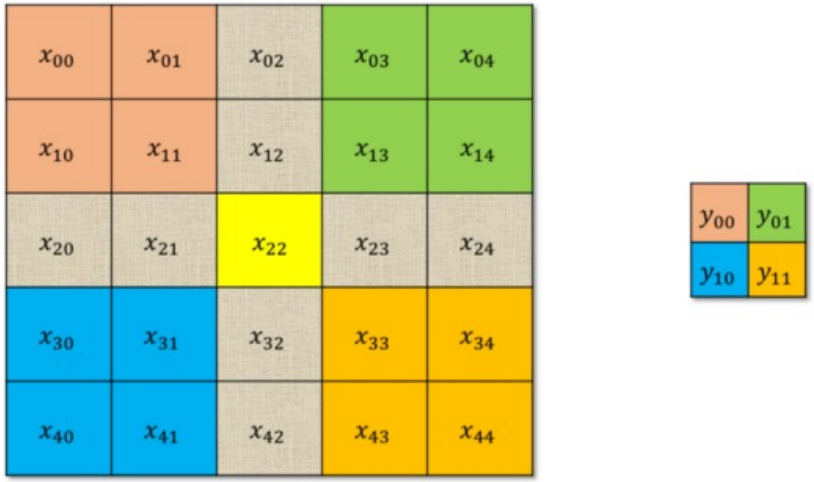
$$y_{01} = x_{02}f_{00} + x_{03}f_{01} + x_{04}f_{02} + x_{12}f_{10} + x_{13}f_{11} + x_{14}f_{12} + x_{22}f_{20} + x_{23}f_{21} + x_{24}f_{22}$$

$$\text{Thus, } \frac{\partial L}{\partial x_{02}} = \frac{\partial L}{\partial y_{00}} f_{02} + \frac{\partial L}{\partial y_{01}} f_{00}$$

Backward Pass example:

- Input x_{22} contributed to the output y_{00} , y_{01} , y_{10}, and y_{11} so the loss gradient w.r.t x_{22} is computed as shown:

$$\frac{\partial L}{\partial x_{mn}} = \sum_{ij} \frac{\partial L}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial x_{mn}}$$



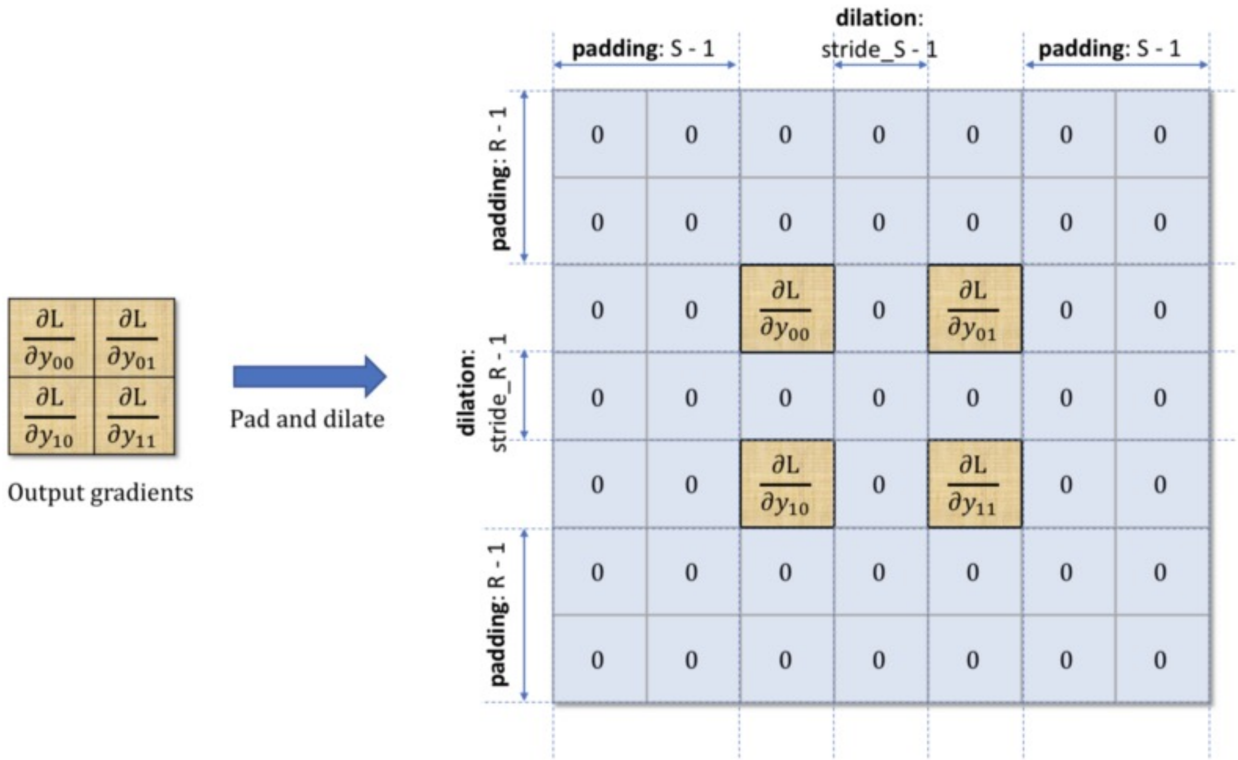
Finally, consider x_{22}. It contributes to all outputs: y_{00} , y_{01}, y_{10}, and y_{11}

$$y_{00} = x_{00}f_{00} + x_{01}f_{01} + x_{02}f_{02} + x_{10}f_{10} + x_{11}f_{11} + x_{12}f_{12} + x_{20}f_{20} + x_{21}f_{20} + x_{22}f_{22}$$
$$y_{01} = x_{02}f_{00} + x_{03}f_{01} + x_{04}f_{02} + x_{12}f_{10} + x_{13}f_{11} + x_{14}f_{12} + x_{22}f_{20} + x_{23}f_{21} + x_{24}f_{22}$$
$$y_{10} = x_{20}f_{00} + x_{21}f_{01} + x_{22}f_{02} + x_{30}f_{10} + x_{31}f_{11} + x_{32}f_{12} + x_{40}f_{20} + x_{41}f_{20} + x_{42}f_{22}$$
$$y_{11} = x_{22}f_{00} + x_{23}f_{01} + x_{24}f_{02} + x_{32}f_{10} + x_{33}f_{11} + x_{34}f_{12} + x_{42}f_{20} + x_{43}f_{20} + x_{44}f_{22}$$

Thus, $\frac{\partial L}{\partial x_{22}} = \frac{\partial L}{\partial y_{00}} f_{22} + \frac{\partial L}{\partial y_{01}} f_{20} + \frac{\partial L}{\partial y_{10}} f_{20} + \frac{\partial L}{\partial y_{11}} f_{00}$

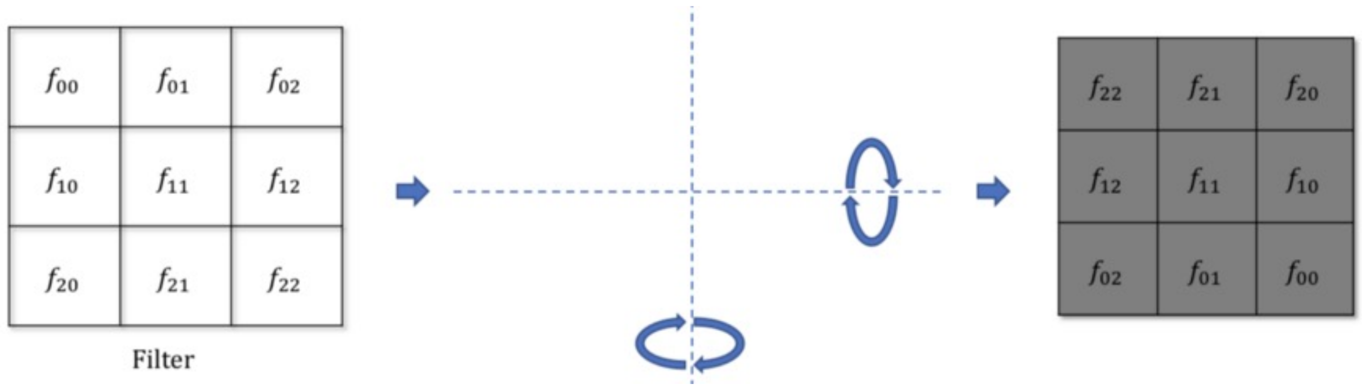
Backward Pass example:

- To visualize the pattern more clearly, we pad the gradient tensor with zeros at the top and bottom as well as to the left and right.
- The number of zeros padded on either side is equal to the stride (horizontal and vertical)
- We also dilate the output gradient pixels with the stride – vertically and horizontally



Backward Pass example:

- We also rotate the filter vertically and horizontally as shown:



Backward Pass example:

- After these modifications, we can now see the calculation of the gradient tensor as follows:

$$\frac{\partial L}{\partial x_{00}} = \frac{\partial L}{\partial y_{00}} f_{00}$$

$$\begin{bmatrix} \frac{\partial L}{\partial x_{00}} & \frac{\partial L}{\partial x_{01}} & \frac{\partial L}{\partial x_{02}} & \frac{\partial L}{\partial x_{03}} & \frac{\partial L}{\partial x_{04}} \\ \frac{\partial L}{\partial x_{10}} & \frac{\partial L}{\partial x_{11}} & \frac{\partial L}{\partial x_{12}} & \frac{\partial L}{\partial x_{13}} & \frac{\partial L}{\partial x_{14}} \\ \frac{\partial L}{\partial x_{20}} & \frac{\partial L}{\partial x_{21}} & \frac{\partial L}{\partial x_{22}} & \frac{\partial L}{\partial x_{23}} & \frac{\partial L}{\partial x_{24}} \\ \frac{\partial L}{\partial x_{30}} & \frac{\partial L}{\partial x_{31}} & \frac{\partial L}{\partial x_{32}} & \frac{\partial L}{\partial x_{33}} & \frac{\partial L}{\partial x_{34}} \\ \frac{\partial L}{\partial x_{40}} & \frac{\partial L}{\partial x_{41}} & \frac{\partial L}{\partial x_{42}} & \frac{\partial L}{\partial x_{43}} & \frac{\partial L}{\partial x_{44}} \end{bmatrix} = \begin{bmatrix} 0 * f_{22} & 0 * f_{21} & 0 * f_{20} & 0 & 0 & 0 & 0 \\ 0 * f_{12} & 0 * f_{11} & 0 * f_{10} & 0 & 0 & 0 & 0 \\ 0 * f_{02} & 0 * f_{01} & \frac{\partial L}{\partial y_{00}} f_{00} & 0 & \frac{\partial L}{\partial y_{01}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\partial L}{\partial y_{10}} & 0 & \frac{\partial L}{\partial y_{11}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Takeaway:

- Convolving with a stride greater than 1 is the same as convolving with stride 1 and “dropping” out of every rows, and of every columns
- Padding the gradient of the output $\frac{\partial L}{\partial y}$ after dilation helps recover the size of the input feature map

Loss gradient w.r.t the Filter

- To understand the computation of loss gradient w.r.t filter, we will use the same example:

- > Horizontal and vertical stride = 2

W

H

x_{00}	x_{01}	x_{02}	x_{03}	x_{04}
x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
x_{20}	x_{21}	x_{22}	x_{23}	x_{24}
x_{30}	x_{31}	x_{32}	x_{33}	x_{34}
x_{40}	x_{41}	x_{42}	x_{43}	x_{44}

Input activations
Input channels C = 1, number of images N = 1,
Image height H = 5, width = 5

S

R

f_{00}	f_{01}	f_{02}
f_{10}	f_{11}	f_{12}
f_{20}	f_{21}	f_{22}

Filter (aka kernel)
Input channels C = 1, number of filters K = 1,
Filter height R = 3, width S = 3,
stride_R = stride_S = 2

Q

P

y_{00}	y_{01}
y_{10}	y_{11}

Output
Output channels K = 1, number of outputs N = 1,
Output height P = 2, width Q = 2

Backward Pass:

Assumption: we have the loss gradient w.r.t the output pixels.

Requirement: calculate the loss gradient w.r.t the filter

Loss gradients w.r.t filter

$\frac{\partial L}{\partial f_{00}}$	$\frac{\partial L}{\partial f_{01}}$	$\frac{\partial L}{\partial f_{02}}$
$\frac{\partial L}{\partial f_{10}}$	$\frac{\partial L}{\partial f_{11}}$	$\frac{\partial L}{\partial f_{12}}$
$\frac{\partial L}{\partial f_{20}}$	$\frac{\partial L}{\partial f_{21}}$	$\frac{\partial L}{\partial f_{22}}$

Loss gradients w.r.t output

$\frac{\partial L}{\partial y_{00}}$	$\frac{\partial L}{\partial y_{01}}$
$\frac{\partial L}{\partial y_{10}}$	$\frac{\partial L}{\partial y_{11}}$

Backward pass:

- Unlike the inputs which contribute to some outputs, each filter contributes to all outputs

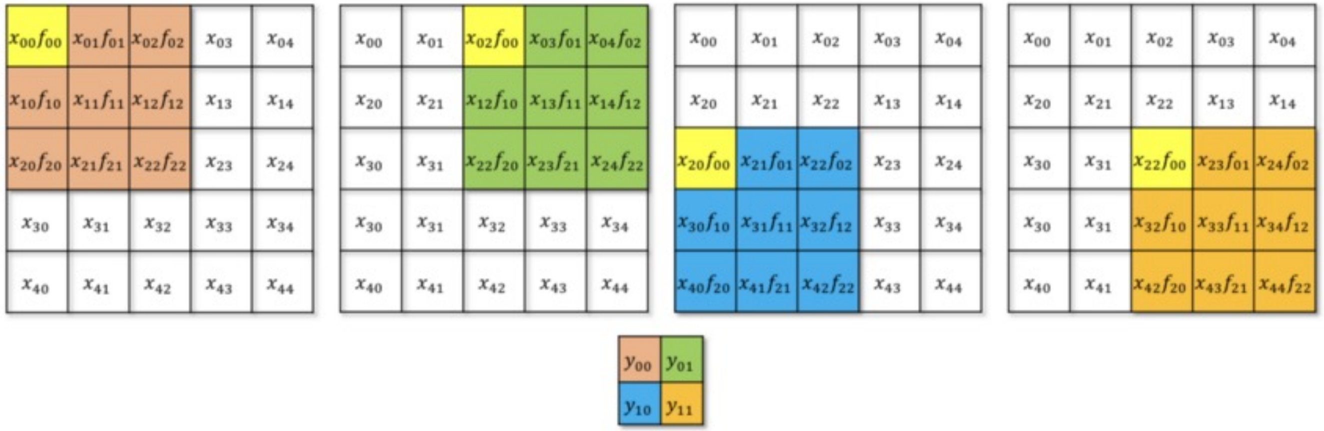
- The gradient computation is done using chain rule and partial differentiation

- i and j represent the position of a single output pixel

$$\frac{\partial \mathrm{L}}{\partial f_{mn}} = \sum_{ij} \frac{\partial L}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial f_{mn}}$$

Backward Pass example:

- Considering the filter f_{00} , the loss gradient is computed as shown:
- Notice the inputs involved in the computation



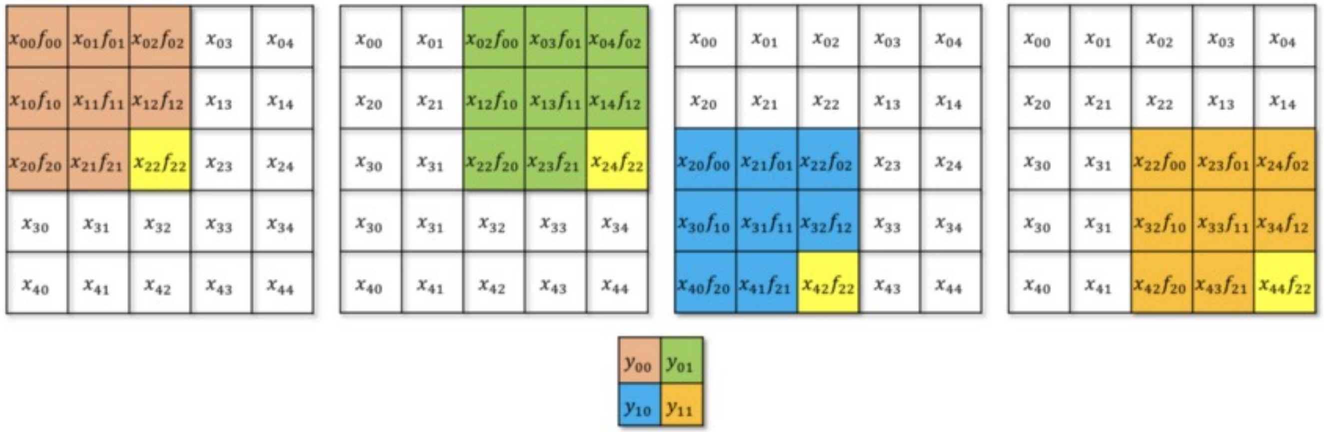
First, consider f_{00}. It contributes to all outputs: y_{00} , y_{01}, y_{10}, and y_{11}

$$
\begin{aligned}
y_{00} &= x_{00}f_{00} + x_{01}f_{01} + x_{02}f_{02} + x_{10}f_{10} + x_{11}f_{11} + x_{12}f_{12} + x_{20}f_{20} + x_{21}f_{20} + x_{22}f_{22} \\
y_{01} &= x_{02}f_{00} + x_{03}f_{01} + x_{04}f_{02} + x_{12}f_{10} + x_{13}f_{11} + x_{14}f_{12} + x_{22}f_{20} + x_{23}f_{21} + x_{24}f_{22} \\
y_{10} &= x_{20}f_{00} + x_{21}f_{01} + x_{22}f_{02} + x_{30}f_{10} + x_{31}f_{11} + x_{32}f_{12} + x_{40}f_{20} + x_{41}f_{20} + x_{42}f_{22} \\
y_{11} &= x_{22}f_{00} + x_{23}f_{01} + x_{24}f_{02} + x_{32}f_{10} + x_{33}f_{11} + x_{34}f_{12} + x_{42}f_{20} + x_{43}f_{20} + x_{44}f_{22}
\end{aligned}
$$

$$\frac{\partial L}{\partial f_{mn}} = \sum_{ij} \frac{\partial L}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial f_{mn}} \text{. Thus, } \frac{\partial L}{\partial f_{00}} = \frac{\partial L}{\partial y_{00}} x_{00} + \frac{\partial L}{\partial y_{01}} x_{02} + \frac{\partial L}{\partial y_{10}} x_{20} + \frac{\partial L}{\partial y_{11}} x_{22}$$

Backward Pass example:

- Considering the filter f_{22}, the loss gradient is computed as shown:
- Notice the inputs involved in the computation



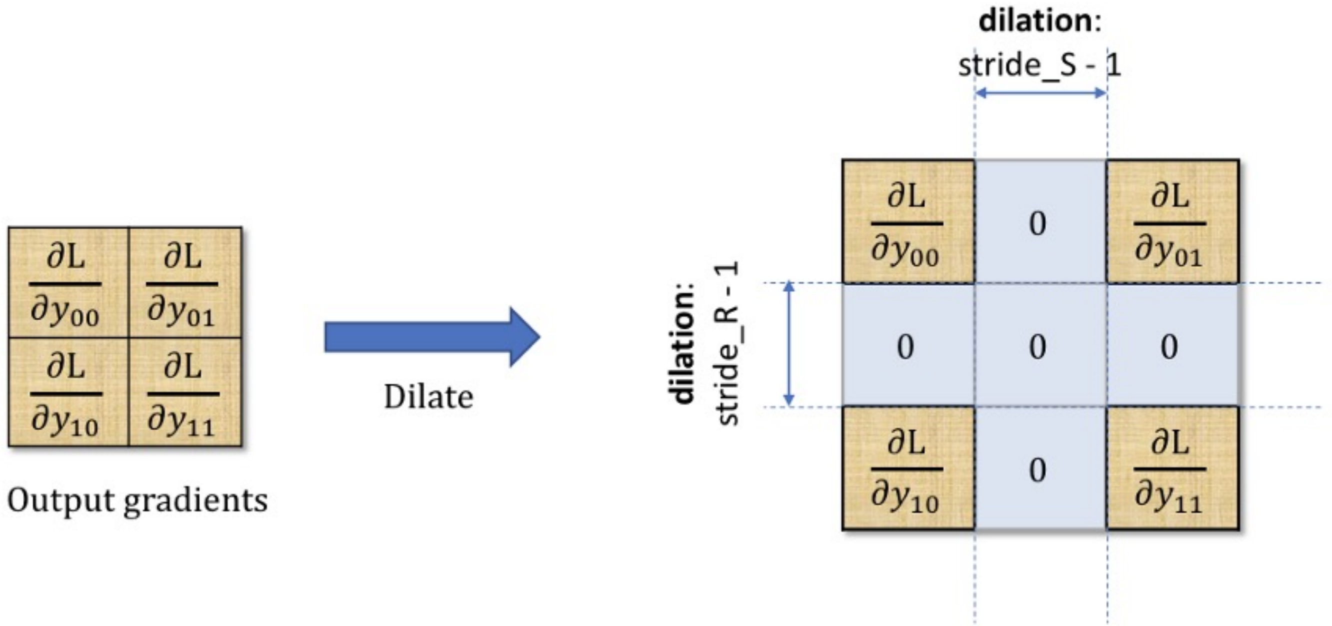
Finally, consider f_{22}. It contributes to all outputs: y_{00} , y_{01}, y_{10}, and y_{11}

$$
\begin{aligned}
y_{00} &= x_{00}f_{00} + x_{01}f_{01} + x_{02}f_{02} + x_{10}f_{10} + x_{11}f_{11} + x_{12}f_{12} + x_{20}f_{20} + x_{21}f_{20} + x_{22}f_{22} \\
y_{01} &= x_{02}f_{00} + x_{03}f_{01} + x_{04}f_{02} + x_{12}f_{10} + x_{13}f_{11} + x_{14}f_{12} + x_{22}f_{20} + x_{23}f_{21} + x_{24}f_{22} \\
y_{10} &= x_{20}f_{00} + x_{21}f_{01} + x_{22}f_{02} + x_{30}f_{10} + x_{31}f_{11} + x_{32}f_{12} + x_{40}f_{20} + x_{41}f_{20} + x_{42}f_{22} \\
y_{11} &= x_{22}f_{00} + x_{23}f_{01} + x_{24}f_{02} + x_{32}f_{10} + x_{33}f_{11} + x_{34}f_{12} + x_{42}f_{20} + x_{43}f_{20} + x_{44}f_{22}
\end{aligned}
$$

$$
\frac{\partial L}{\partial f_{mn}} = \sum_{ij} \frac{\partial L}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial f_{mn}} \text{. Thus, } \frac{\partial L}{\partial f_{22}} = \frac{\partial L}{\partial y_{00}} x_{22} + \frac{\partial L}{\partial y_{01}} x_{24} + \frac{\partial L}{\partial y_{10}} x_{42} + \frac{\partial L}{\partial y_{11}} x_{44}
$$

Backward Pass example:

- To visualize the underlying pattern, we will modify the output gradient tensor by dilating the pixels with the stride vertically and horizontally:



Backward Pass example:

- After these modifications, we can now see the calculation of the filter gradient tensor as follows :

$$\frac{\partial L}{\partial f_{00}} = \frac{\partial L}{\partial y_{00}} x_{00} + \frac{\partial L}{\partial y_{01}} x_{02} + \frac{\partial L}{\partial y_{10}} x_{20} + \frac{\partial L}{\partial y_{11}} x_{22}$$

$$\begin{bmatrix} \frac{\partial L}{\partial f_{00}} & \frac{\partial L}{\partial f_{01}} & \frac{\partial L}{\partial f_{02}} \\ \frac{\partial L}{\partial f_{10}} & \frac{\partial L}{\partial f_{11}} & \frac{\partial L}{\partial f_{12}} \\ \frac{\partial L}{\partial f_{20}} & \frac{\partial L}{\partial f_{21}} & \frac{\partial L}{\partial f_{22}} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial y_{00}} x_{00} & 0 * x_{01} & \frac{\partial L}{\partial y_{01}} x_{02} & x_{03} & x_{04} \\ 0 * x_{10} & 0 * x_{11} & 0 * x_{12} & x_{13} & x_{14} \\ \frac{\partial L}{\partial y_{10}} x_{20} & 0 * x_{21} & \frac{\partial L}{\partial y_{11}} x_{22} & x_{23} & x_{24} \\ x_{30} & x_{31} & x_{32} & x_{33} & x_{34} \\ x_{40} & x_{41} & x_{42} & x_{43} & x_{44} \end{bmatrix}$$

Takeaway:

- **The CNN Backpropagation operation with stride>1 is identical to a stride = 1 Convolution operation of the input gradient tensor with a dilated version of the output gradient tensor!**

References:

https://medium.com/@mayank.utexas/backpropagation-for-convolution-with-strides-8137e4fc2710

https://medium.com/@mayank.utexas/backpropagation-for-convolution-with-strides-fb2f2efc4faa

https://medium.com/@pavisj/convolutions-and-backpropagations-46026a8f5d2c

https://towardsdatascience.com/backpropagation-in-a-convolutional-layer-24c8d64d8509