Proof of gradient sizes

Notation

X: the 3D input of size $n \times n \times k$.
$X^{(i)}$: the ith channel of the 2D input.
p: the number of filters.
F_j: the jth filter of size $m \times m \times k$ where $j \in 1 \ldots p$.
$F_j^{(i)}$: the ith channel of the jth filter.
Y: the output of size $(n-m+1) \times (n-m+1) \times p$ where p is the number of filters.
$Y^{(j)}$: the jth channel of the output.
F': the filter F rotated 180°.

Loss-to-filter gradient

First we consider the side of the ith channel of the loss-to-filter gradient,

$$\frac{\partial L}{\partial F_j^{(i)}} = X^{(i)} \circledast \frac{\partial L}{\partial Y^{(j)}}$$

$X^{(i)}$ has size $n \times n$ and $\frac{\partial L}{\partial Y^{(j)}}$ has size $(n-m+1) \times (n-m+1)$, so the result of the convolution will be of length/width $n-(n-m+1)+1 = m$, which is exactly the length/width of the filter F_j. Each channel i of the input $X^{(i)}$ corresponds to one channel of the filter gradient $\frac{\partial L}{\partial F_j^{(i)}}$ and there are k input channels by definition, so the complete filter gradient,

$$\frac{\partial L}{\partial F_j} = \left[\frac{\partial L}{\partial F_j^{(1)}}, \ldots, \frac{\partial L}{\partial F_j^{(k)}} \right]$$

will have depth k. Thus the overall dimension of the filter gradient is $m \times m \times k$, which is exactly the size of the filter.

Loss-to-input gradient

Now we examine the ith channel of the input gradient,

$$\frac{\partial L}{\partial X^{(i)}} = \sum_j \left(F_j^{(i)} \right)' \circledast \frac{\partial L}{\partial Y^{(j)}}$$

With a stride of one, as we will use for all convolution layers, the output $Y^{(j)}$ will have size $(n-m+1) \times (n-m+1)$.

$\frac{\partial L}{\partial Y^{(j)}}$ is the same size as the output $Y^{(j)}$, eg $(n-m+1) \times (n-m+1)$. We pad this gradient by $m-1$ on all 4 sides, so that the length/width is $(n-m+1)+2(m-1) = n+m-1$. Then the convolution $\left(F_j^{(i)} \right)' \circledast \frac{\partial L}{\partial Y^{(j)}}$ will produce an output of length/width $(n+m-1)-m+1 = n$. Thus the loss to input channel i, $\frac{\partial L}{\partial X^{(i)}}$ will have size $n \times n$. As seen above, the contributions for the various filters F_j are added to this channel of the gradient and do not affect the dimension.

The depth of the gradient $\frac{\partial L}{\partial X}$ comes from the number of channels in the filter, k. The filter is defined to have the same depth as the input X, thus the gradient $\frac{\partial L}{\partial X}$ will have depth k for an overall dimension of $n \times n \times k$, as seen below.

$$\frac{\partial L}{\partial X} = \left[\frac{\partial L}{\partial X^{(1)}} \quad \ldots \quad \frac{\partial L}{\partial X^{(k)}} \right]$$