DeepGrid

Organic Deep Learning.

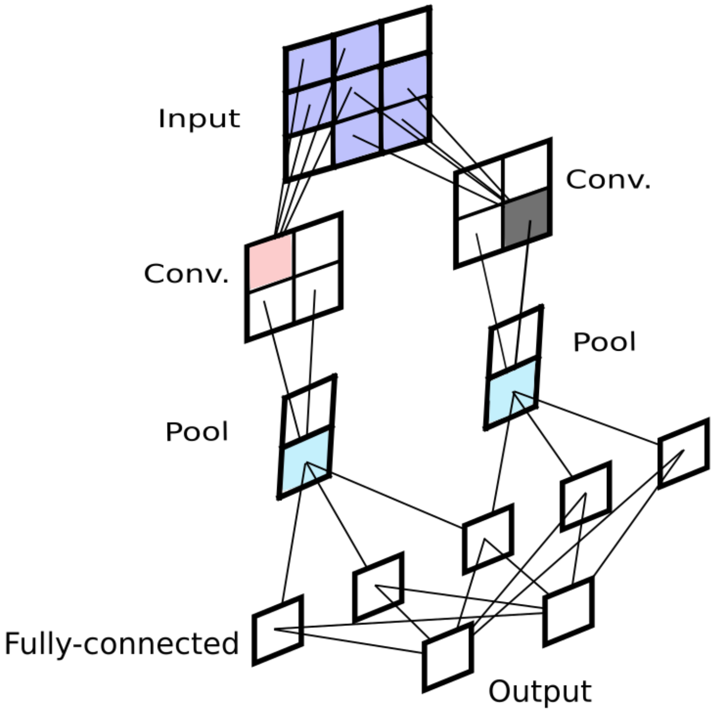
Latest Article:
Backpropagation In Convolutional Neural Networks
5 September 2016

Backpropagation In Convolutional Neural Networks

Jefkine, 5 September 2016

Introduction

Convolutional neural networks (CNNs) are a biologically-inspired variation of the multilayer perceptrons (MLPs). CNNs emulate the basic mechanics of the animal visual cortex. Neurons in CNNs share weights unlike in MLPs where each neuron has a separate weight vector. This sharing of weights ends up reducing the overall number of trainable weights hence introducing sparsity.



Utilizing the weights sharing strategy, neurons are able to perform convolutions on the data with the convolution filter being formed by the weights. This is then followed by a pooling operation which is a form of non-linear down-sampling that progressively reduces the spatial size of the representation hence reducing the amount of parameters and computation in the network. An illustration can be seen in the diagram above.

After several convolutional and max pooling layers, the image size (fearture map size) is reduced and more complex features are extracted. Eventually with a small enough feature map, the contents are squashed into a one dimension vector and fed into a fully-connected MLP for processing.

Existing between the convolution and the pooling layer is a ReLU layer in which a non-saturating activation function is applied element-wise, i.e. $f(x) = max(0, x)$ thresholding at zero.

The last layer of the fully-connected MLP seen as the output, is a loss layer which is used to specify how the network training penalizes the deviation between the predicted and true labels.

Cross-correlation

Given an input image I and a kernel or filter F of dimensions $k \times k$, a cross-correlation operation leading to an output image C is given by:

$$\begin{aligned} C &= I \otimes F \\ C(x,y) &= \sum_{a=0}^{k-1} \sum_{b=0}^{k-1} I(x+a, y+b) F(a,b) \end{aligned} \tag{1}$$

Convolution

Given an input image I and a kernel or filter F of dimensions $k \times k$, a convolution operation leading to an output image C is given by:

$$\begin{aligned} C &= I * F \\ C(x,y) &= \sum_{a=0}^{k-1} \sum_{b=0}^{k-1} I(x-a, y-b) F(a,b) \end{aligned} \tag{2}$$

Convolution is the same as cross-correlation, except that the kernel is “flipped” (horizontally and vertically).

In the two operations above, the region of support for $C(x,y)$ is given by ranges $0 \leq x \leq k-1$ and $0 \leq y \leq k-1$. These are the ranges for which the pixels are defined. In the case of undefined pixels, the input image could be zero padded to result in an output of a size similar to the input image.

Notation

1. l is the l^{th} layer where $l = 1$ is the first layer and $l = L$ is the last layer.
2. $w^l_{x,y}$ is the weight vector connecting neurons of layer l with neurons of layer $l+1$.
3. $o^l_{x,y}$ is the output vector at layer l

$$\begin{aligned} o^l_{x,y} &= w^l_{x,y} * a_{x,y} \\ &= \sum_{x'} \sum_{y'} w^l_{x',y'} a^l_{x-x',y-y'} \end{aligned}$$

1. a^l is the activated output vector for a hidden layer l.

$$a^l_{x,y} = f(o^{l-1}_{x,y})$$

2. $f(x)$ is the activation function. A ReLU function $f(x) = max(0,x)$ is used as the activation function.
3. A^L is the matrix representing all entries of the last output layer L given by

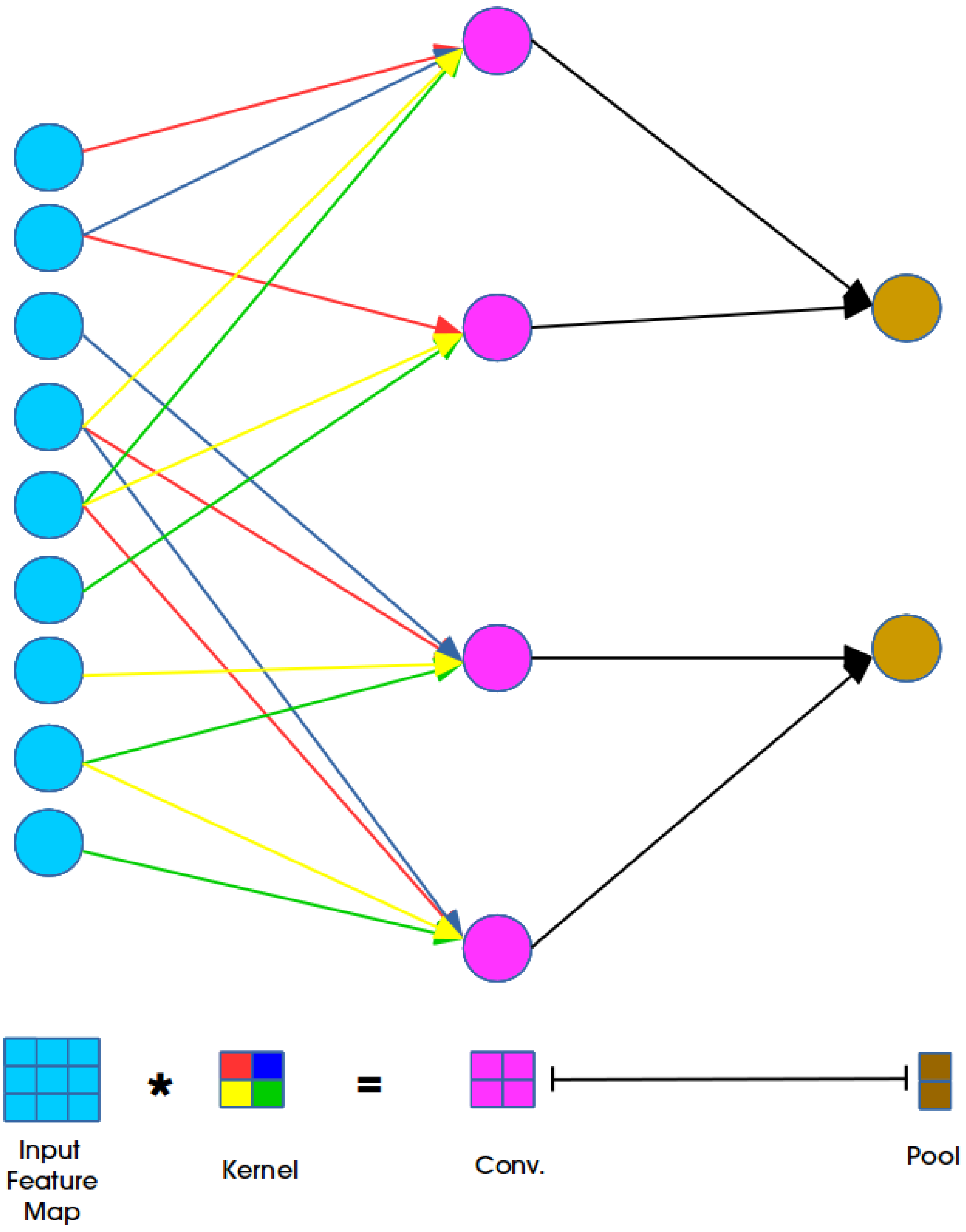
$$a^L_{x,y} = f(o^{L-1}_{x,y})$$

4. P is the matrix representing all the training patterns for network training
5. T is the matrix of all the targets.

Foward Propagarion

To perform a convolution operation, the kernel is flipped 180° and slid across the input feature map in equal and finite strides. At each location, the product between each element of the kernel and the input element it overlaps is computed and the results summed up to obtain the output at that current location.

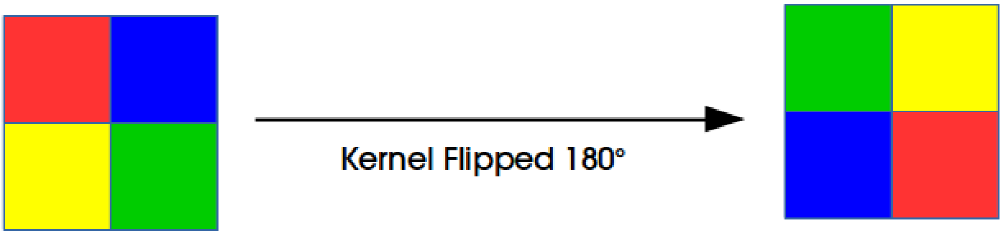
This procedure is repeated using different kernels to form as many output feature maps as desired. The concept of weight sharing is used as demonstrated in the diagram below:



Units in convolution layer have receptive fields of width 3 in the input feature map and are thus only connected to 3 adjacent neurons in the input layer. This is the idea of **sparse connectivity** in CNNs where there exists local connectivity pattern between neurons in adjacent layers.

The color codes of the weights joining the input layer to the convolution layer show how the kernel weights are distributed (shared) amongst neurons in the adjacent layers. Weights of the same color are constrained to be identical.

For the convolution operation to be performed, the kernel is flipped as shown in the diagram below before:



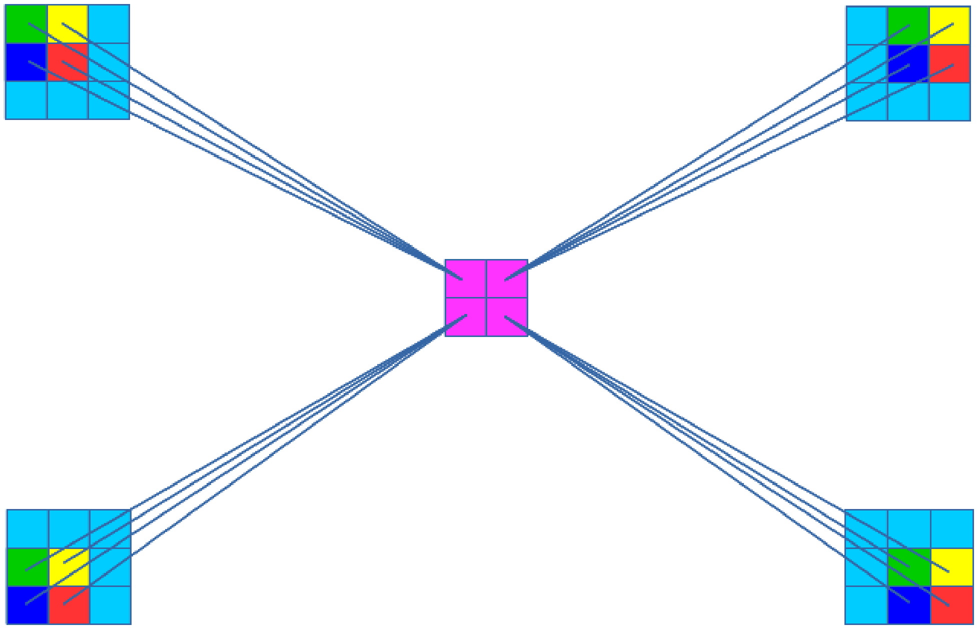
The convolution equation is given by:

$$O(x,y) = w^l_{x,y} * a^l_{x,y} + b^l_{x,y} \tag{3}$$

$$O(x,y) = \sum_{x'}\sum_{y'} w^l_{x',y'} a^l_{x-x',y-y'} + b^l_{x,y} \tag{4}$$

$$= \sum_{x'}\sum_{y'} w^l_{x',y'} f(o^{l-1}_{x-x',y-y'}) + b^l_{x,y} \tag{5}$$

This is illustrated below:



Error

For a total of P predictions, the predicted network outputs a_p^L and their corresponding targeted values t_p the the mean squared error is given by:

$$E = \frac{1}{2} \sum_{p=1}^{P} \left(t_p - a_p^L\right)^2 \tag{6}$$

Learning will be achieved by adjusting the weights such that A^L is as close as possible or equals to T. In the classical back-propagation algorithm, the weights are changed according to the gradient descent direction of an error surface E.

Backpropagation

For back-propagation there are two updates performed, for the weights and the deltas. Lets begin with the weight update. The gradient component is for each weight can be obtained by applying the chain rule.

$$\begin{aligned} \frac{\partial E}{\partial w_{x,y}^l} &= \sum_{x'} \sum_{y'} \frac{\partial E}{\partial o_{x',y'}^l} \frac{\partial o_{x',y'}^l}{\partial w_{x,y}^l} \\ &= \sum_{x'} \sum_{y'} \delta_{x',y'}^l \frac{\partial o_{x',y'}^l}{\partial w_{x,y}^l} \\ &= \sum_{x'} \sum_{y'} \delta_{x',y'}^l \frac{\partial o_{x',y'}^l}{\partial w_{x,y}^l} \end{aligned} \tag{7}$$

In Eqn (7) , $o^l_{x',y'}$ is equivalent to $w^l_{x',y'}a^l_{x',y'} + b^l$ and so applying the convolution operation here gives us an equation of the form:

$$\begin{aligned}
\frac{\partial o^l_{x',y'}}{\partial w^l_{x,y}} &= \frac{\partial}{\partial w^l_{x,y}}\left(\sum_{x''}\sum_{y''} w^l_{x'',y''} a^l_{x'-x'',y'-y''} + b^l\right) \\
&= \frac{\partial}{\partial w^l_{x,y}}\left(\sum_{x''}\sum_{y''} w^l_{x'',y''} f\left(o^{l-1}_{x'-x'',y'-y''}\right) + b^l\right) \qquad (8)
\end{aligned}$$

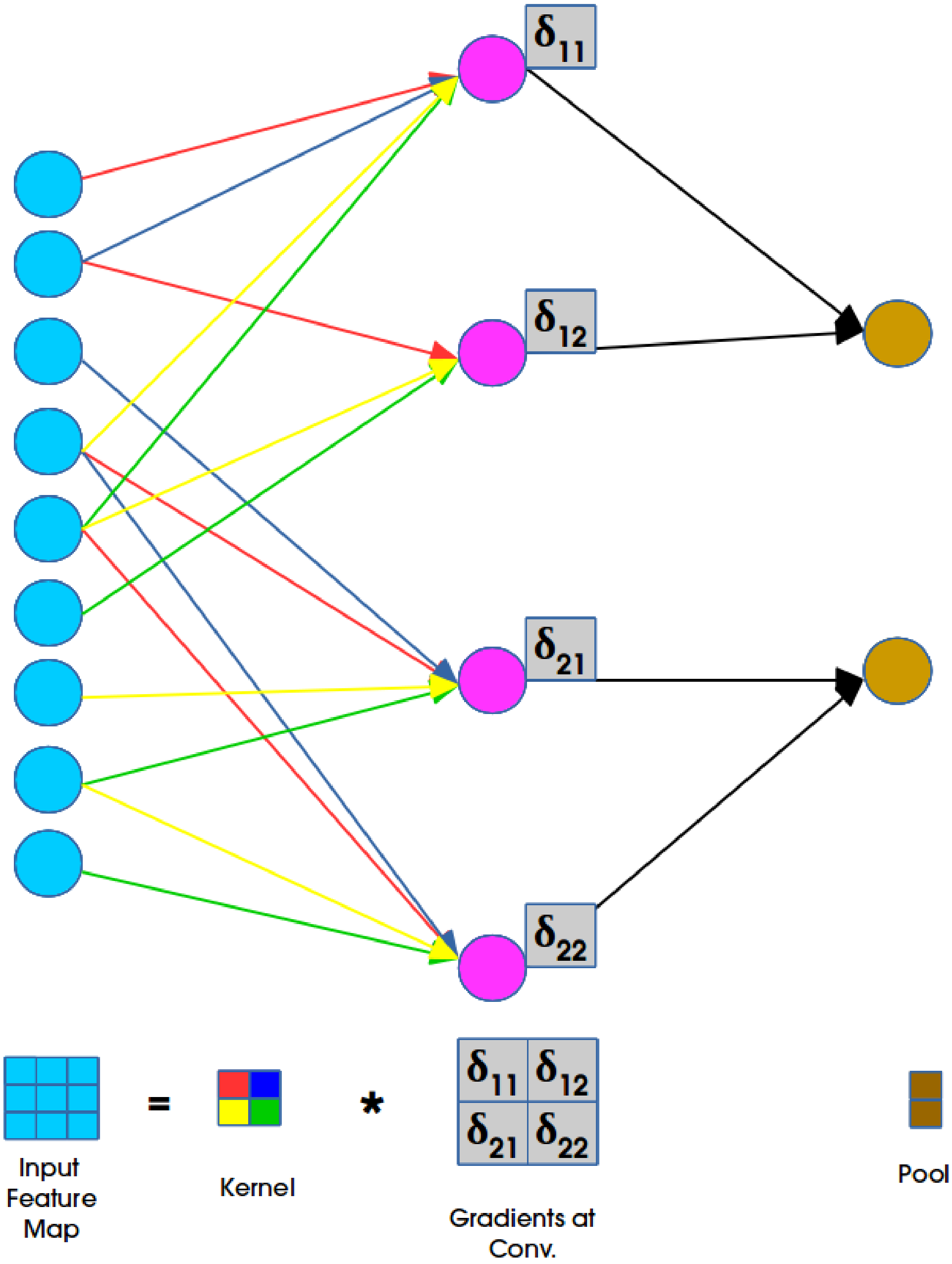
Expanding the summation in Eqn (8) and taking the partial derivatives for all the components results in zero values for all except the components where $x = x''$ and $y = y''$ in $w^l_{x'',y''}$ which implies $x' - x'' \mapsto x' - x$ and $y' - y'' \mapsto y' - y$ in $f\left(o^{l-1}_{x'-x'',y'-y''}\right)$ as follows:

$$\begin{aligned}
\frac{\partial o^l_{x',y'}}{\partial w^l_{x,y}} &= \frac{\partial}{\partial w^l_{x,y}}\left(w^l_{0,0} f\left(o^{l-1}_{x'-0,y'-0}\right) + \cdots + w^l_{x,y} f\left(o^{l-1}_{x'-x,y'-y}\right) + \cdots + b^l\right) \\
&= \frac{\partial}{\partial w^l_{x,y}}\left(w^l_{x,y} f\left(o^{l-1}_{x'-x,y'-y}\right)\right) \\
&= f\left(o^{l-1}_{x'-x,y'-y}\right) \qquad (9)
\end{aligned}$$

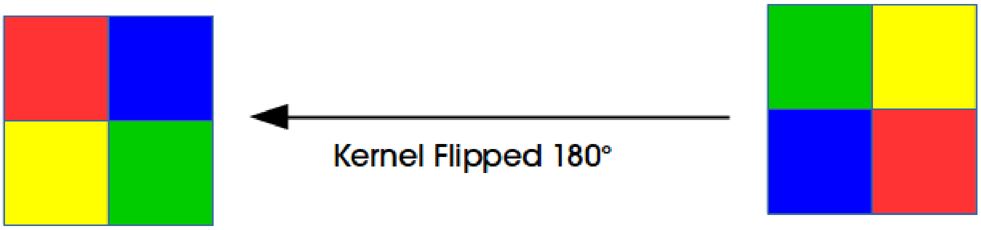
Substituting Eqn (9) in Eqn (7) gives us the following results:

$$\begin{aligned}
\frac{\partial E}{\partial w^l_{x,y}} &= \sum_{x'}\sum_{y'} \delta^l_{x',y'} f\left(o^{l-1}_{x'-x,y'-y}\right) \qquad (10) \\
&= \delta^l_{x,y} * f\left(o^{l-1}_{-x,-y}\right) \qquad (11) \\
&= \delta^l_{x,y} * f\left(\mathrm{rot}_{180^\circ}\left(o^{l-1}_{x,y}\right)\right) \qquad (12)
\end{aligned}$$

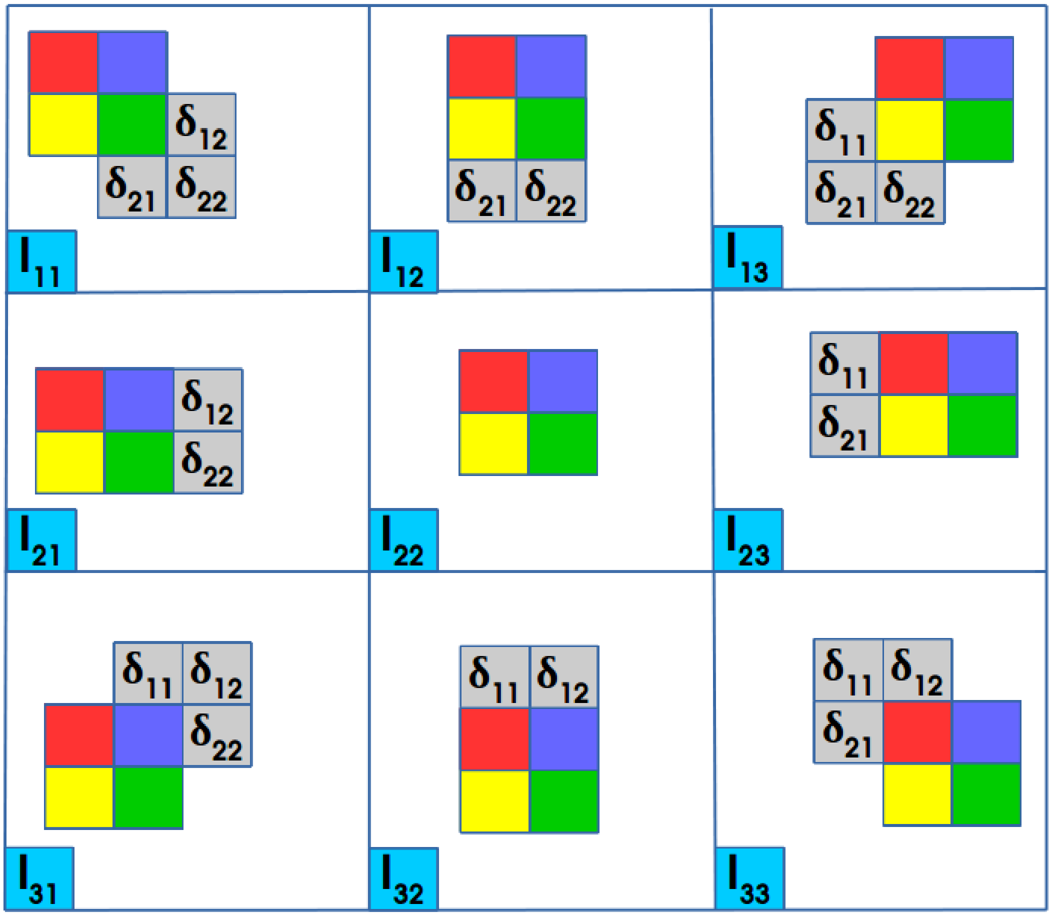
In Eqn (12) above, rotation of the kernel, causes the shift from $f\left(o^{l-1}_{-x,-y}\right)$ to $f\left(\mathrm{rot}_{180^\circ}\left(o^{l-1}_{x,y}\right)\right)$. The diagram below shows gradients $(\delta_{11}, \delta_{12}, \delta_{21}, \delta_{22})$ generated during back-propagation:



For back-propagation routine, the kernel is flipped 180° yet again before the convolution operation is done on the gradients to reconstruct the input feature map.



The convolution operation used to reconstruct the input feature map is shown below:



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

During the reconstruction process, the deltas $(\delta_{11}, \delta_{12}, \delta_{21}, \delta_{22})$ are used. The deltas are provided by an equation of the form:

$$\delta^l_{x,y} = \frac{\partial E}{\partial o^l_{x,y}} \tag{13}$$

Using chain rule and introducing sums give us the following results:

$$\begin{aligned} \frac{\partial E}{\partial o^l_{x,y}} &= \sum_{x'} \sum_{y'} \frac{\partial E}{\partial o^{l+1}_{x',y'}} \frac{\partial o^{l+1}_{x',y'}}{\partial o^l_{x,y}} \\ &= \sum_{x'} \sum_{y'} \delta^{l+1}_{x',y'} \frac{\partial o^{l+1}_{x',y'}}{\partial o^l_{x,y}} \end{aligned} \tag{14}$$

In Eqn (14), $o^{l+1}_{x',y'}$ is equivalent to $w^{l+1}_{x',y'} a^{l+1}_{x',y'} + b^{l+1}$ and so applying the convolution operation here gives us an equation of the form:

$$\begin{aligned}
\frac{\partial o^{l+1}_{x',y'}}{\partial o^{l}_{x,y}} &= \frac{\partial}{\partial o^{l}_{x,y}} \left(\sum_{x''} \sum_{y''} w^{l+1}_{x'',y''} a^{l+1}_{x'-x'',y'-y''} + b^{l+1} \right) \\
&= \frac{\partial}{\partial o^{l}_{x,y}} \left(\sum_{x''} \sum_{y''} w^{l+1}_{x'',y''} f\left(o^{l}_{x'-x'',y'-y''} \right) + b^{l+1} \right) \qquad (15)
\end{aligned}$$

Expanding the summation in Eqn (15) and taking the partial derivatives for all the components results in zero values for all except the components where $x = x' - x''$ and $y = y' - y''$ in $f\left(o^{l}_{x'-x'',y'-y''} \right)$ which implies $x'' = x + x'$ and $y'' = y + y'$ in $w^{l+1}_{x'',y''}$ as follows:

$$\begin{aligned}
\frac{\partial o^{l+1}_{x',y'}}{\partial o^{l}_{x,y}} &= \frac{\partial}{\partial o^{l}_{x,y}} \left(w^{l+1}_{0,0} f\left(o^{l}_{x'-0,y'-0} \right) + \cdots + w^{l+1}_{x+x',y+y'} f\left(o^{l}_{x,y} \right) + \cdots + b^{l+1} \right) \\
&= \frac{\partial}{\partial o^{l}_{x,y}} \left(w^{l+1}_{x+x',y+y'} f\left(o^{l}_{x,y} \right) \right) \\
&= w^{l+1}_{x+x',y+y'} \frac{\partial}{\partial o^{l}_{x,y}} \left(f\left(o^{l}_{x,y} \right) \right) \\
&= w^{l+1}_{x+x',y+y'} f'\left(o^{l}_{x,y} \right) \qquad (16
\end{aligned}$$

Substituting Eqn (16) in Eqn (14) gives us the following results:

$$\frac{\partial E}{\partial o^{l}_{x,y}} = \sum_{x'} \sum_{y'} \delta^{l+1}_{x',y'} w^{l+1}_{x+x',y+y'} f'\left(o^{l}_{x,y} \right) \qquad (17)$$

In Eqn (17), we now have a cross-correlation which is transformed to a convolution by “flipping” the kernel (horizontally and vertically) as follows:

$$\begin{aligned}
\frac{\partial E}{\partial o^{l}_{x,y}} &= \sum_{x'} \sum_{y'} \delta^{l+1}_{x',y'} \mathrm{rot}_{180^\circ} \left(w^{l+1}_{x+x',y+y'} \right) f'\left(o^{l}_{x,y} \right) \\
&= \sum_{x'} \sum_{y'} \delta^{l+1}_{x',y'} w^{l+1}_{x-x',y-y'} f'\left(o^{l}_{x,y} \right) \\
&= \delta^{l+1}_{x,y} * \mathrm{rot}_{180^\circ} \left(w^{l+1}_{x,y} \right) f'\left(o^{l}_{x,y} \right) \qquad (18)
\end{aligned}$$

Pooling Layer

The function of the pooling layer is to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network, and hence to also control overfitting. No learning takes place on the pooling layers [2].

Pooling units are obtained using functions like max-pooling, average pooling and even L2-norm pooling. At the pooling layer, forward propagation results in an $N \times N$ pooling block being reduced to a single value - value of the "winning unit". Back-propagation of the pooling layer then computes the error which is acquired by this single value "winning unit".

To keep track of the "winning unit" its index noted during the forward pass and used for gradient routing during back-propagation. Gradient routing is done in the following ways:

- **Max-pooling** - the error is just assigned to where it comes from - the "winning unit" because other units in the previous layer's pooling blocks did not contribute to it hence all the other assigned values of zero
- **Average pooling** - the error is multiplied by $\frac{1}{N \times N}$ and assigned to the whole pooling block (all units get this same value).

Conclusion

Convolutional neural networks employ a weight sharing strategy that leads to a significant reduction in the number of parameters that have to be learned. The presence of larger receptive field sizes of neurons in successive convolution layers coupled with the presence of pooling layers also lead to translation invariance. As we have observed the derivations of forward and backward propagations will differ depending on what layer we are propagating through.

References

1. Dumoulin, Vincent, and Francesco Visin. "A guide to convolution arithmetic for deep learning." stat 1050 (2016): 23. [pdf]
2. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.,Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation 1(4), 541–551 (1989)
3. Wikipedia page on Convolutional neural network
4. Convolutional Neural Networks (LeNet) deeplearning.net

Related Posts

Formulating The ReLU 24 Aug 2016

0 Comments Deep Grid 1 Login

Recommend Share Sort by Best

Start the discussion…

Be the first to comment.

Subscribe Add Disqus to your site Add Disqus Add Privacy