Backprop Algorithm in CNN

I. 前向传播.

① 经过卷积层. 水平滑动的步长等于垂直滑动的步长.

假设前层的激活输出为 $a_{i,j}^{l-1}$，本层的加权输入为 $z_{i,j}^{l}$，激活输出为 $a_{i,j}^{l}$. 有

$$a_{i,j}^{l} = f(z_{i,j}^{l}) = f\left(\sum_{p=1}^{s}\sum_{q=1}^{s} a^{l-1}_{\substack{i+p-1+(stride-1)\times(i-1)\\ j+q-1+(stride-1)\times(j-1)}} \times k_{pq}^{l} + b^{l}\right)$$

注：一个kernel可能具有多个channel，取决于输入的channel；而 b^l 只和channel数相关；即便有多个channels，还是等同于一个kernel.

* 对于神经网络中的一般意义的卷积而言：

① 不是数学意义上的卷积，stride可以不为1，且kernel不必要翻转.
应当是一种意义上的相关。

② 经过测试，当stride = 1时，Conv2d 在 TensorFlow 中具有和 matlab 中不翻转的Conv2有相同的效果.（在 padding = 'same' / 'valid' 的情况下.

理论上，要达到最好的效果所填充的0的个数介于 padding = 'same' 和 padding = 'valid' 之间.
（不填充0元素）

② 经过池化层

需要注意的是，池化层是不具有paras的；这里将卷积层和池化层分开了；当然，也可以将它们合并在一起，统称为卷积层.

Max-pooling：可以对卷积后的特征进行下采样，减小特征的size；

pooling的操作可以类似于卷积（stride = pooling_size），但是 kernel 根据池化类型和输入数据而定.

Max-pooling：$\begin{bmatrix}0.1 & 0.2\\ 0.3 & 0.6\end{bmatrix} \xrightarrow{\text{Max-pooling}} [\,0.6\,]$ （能保留纹理上的特征）

average-pooling：$\begin{bmatrix}0.1 & 0.2\\ 0.3 & 0.6\end{bmatrix} \xrightarrow[\text{pooling}]{\text{average}} \begin{bmatrix}0.3 & 0.3\\ 0.3 & 0.3\end{bmatrix}$ （能保留整体的特征）

L_2-norm-pooling：

pool一定意义上具有平移不变性（对特征出现的位置不敏感）

通过对pool_size进行选取，可以把不同大小的图像转化为相同维度的特征.

③. 经过 fc layer（全连接层）：仍然可以看成卷积的形式，对于不同的feature map
⇓
用size为相同的 Conv kernel进行卷积

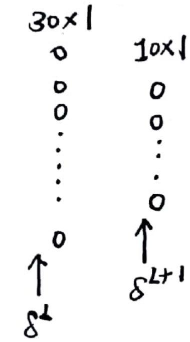
一般来说，需要先进行 flatten，再进行前向传播.

$$a^{l} = f(W^{l}\cdot a^{l-1} + b^{l})$$

W^l：flatten后的权重矩阵　　b^l：Bias 偏置

feature map 的 channels 和 Conv-kernels 相同.

2. 反向传播



①经过全连接层，和全连接网络中的BP算法相同：

假设第L+1层的误差测度为：$\delta^{L+1} = \frac{\partial C}{\partial z^{l+1}}$

而该层的权值和偏置则以如下方式更新：

$$\Delta w_{m,n}^{L+1} = \sum_k \frac{\partial C}{\partial z_k^{L+1}} \cdot \frac{\partial z_k^{L+1}}{\partial w_{m,n}^{L+1}} = \sum_k \frac{\partial C}{\partial z_k^{L+1}} \cdot \frac{\partial \sum_a w_{ka}^{l+1} \cdot a_a^{l} + b_k^{l+1}}{\partial w_{m,n}^{L+1}}$$

$$= \delta_m^{L+1} \cdot a_n^{l}$$

$$\Delta b_j^{l+1} = \sum_k \frac{\partial C}{\partial z_k^{L+1}} \cdot \frac{\partial z_k^{L+1}}{\partial b_j^{L+1}} = \sum_k \frac{\partial C}{\partial z_k^{L+1}} \cdot \frac{\partial \sum_a w_{ka}^{L+1} \cdot a_a^{l} + b_k^{l+1}}{\partial b_j^{l+1}}$$

$$= \delta_j^{L+1} \cdot 1$$

$$\delta_j^L = \sum_k \frac{\partial C}{\partial z_k^{l+1}} \cdot \frac{\partial z_k^{l+1}}{\partial z_j^{l}} = \sum_k \delta_k^{l+1} \cdot \frac{\partial \sum_m w_{km}^{l+1} \cdot a_m^{l} + b_k^{l+1}}{\partial z_j^{l}} = \sum_k \delta_k^{l+1} \cdot w_{kj}^{l+1} \cdot \sigma'(z_j^{l})$$

Matrix Form.

交换顺序　　上一层激活函数的导数.

$$\Delta w^{L+1} = \delta^{L+1} \cdot (a^{l})^T. \quad \Delta b^{l+1} = \delta^{L+1} \qquad \delta^L = (\delta^{L+1} \cdot (w^{L+1})^T) \odot \sigma'(z^{l})$$

② 经过池化层（没有参数）。可以重叠. <降维>

假设经过池化之后的误差测度矩阵为 δ^l，在进行反向传播之前，需要进行相应的操作.

i) 还原成矩阵形式.

ii) 进行 upsample.（上采样）

考虑池化层前的矩阵维度为：k × width × height（kernel × 图像的大小）

*对于 max-pooling：

正向传播 <记录最大值的位置>

$$\begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{bmatrix} \xrightarrow[\text{stride}=2]{\text{假定 pool\_size}=3\times3} \begin{bmatrix} \circ & \circ \\ \circ & \circ \end{bmatrix}$$

反向传播

$$\begin{bmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{bmatrix} \xrightarrow{\text{只与最大值有关}} \underbrace{\begin{bmatrix} 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \end{bmatrix}}_{\text{补0}} \xrightarrow[\text{滑动窗口为 }3\times3\text{ stride}=2]{\text{赋值}} \begin{bmatrix} 0&0&0&\delta_{12}&0 \\ 0&\delta_{11}&0&0&0 \\ 0&0&0&0&0 \\ \delta_{21}&0&0&0&0 \\ 0&0&0&0&\delta_{22} \end{bmatrix}$$

<with no activation>（放到相应的位置）

如果~~池化~~卷积层也添加了激活函数，那么，同样需要加入. $\sigma'(z^{l-1})$

$$\delta^{L-1} = \text{upsample}(\delta^L) \odot \sigma'(z^{L-1})$$ → pool之前相应的像素值.

注：如果上一层结果未被完全进行pool，赋值时可不动. <最好避免这种情况>

丢失信息.

由 扫描全能王 扫描创建

- Mean-pooling.

对于平均池化而言，可仿照 Max-pooling 进行运作。

$$\begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{bmatrix} \xrightarrow[\text{stride}=2]{\text{pool}:3\times 3} \begin{bmatrix} \circ & \circ \\ \circ & \circ \end{bmatrix}$$ <正向传播>

前层输出

反向传播.

$$\begin{bmatrix} \circ & \circ \\ \circ & \circ \end{bmatrix} \xrightarrow{\text{先补}0} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \longrightarrow \begin{bmatrix} \frac{1}{9}\delta_{11}\cdot\sigma'(z_{11}^{l}) & & \\ & \ddots & \\ & & \ddots \end{bmatrix}$$

$$\text{unsample}(\delta^{l}) \odot \sigma'(z^{l-1})$$

前层激活函数.

注：i) 在pool的过程中，最好不要缺少特征.（选择适当的poolsize和stride）

ii) 注意前层的激活函数在反向传播中的作用.

1) 对于 max-pooling，可直接相加。

2) 对于 mean-pooling：作用到相应元素求和.

1)
$$\begin{bmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{bmatrix} \longrightarrow \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ & \delta_{11}\cdot\sigma'(z_{22}^{l}) & & & \\ & & & & \\ & & & & \end{bmatrix}$$

如果出现在重复位置.

$$\delta_{11}\cdot\sigma'(z_{11}^{l}) + \delta_{12}\cdot\sigma'(z_{11}^{l})$$

2) 对于类似情况：

$$\frac{1}{\text{poolSize}}\delta_{11}\cdot\sigma'(z_{11}^{l}) + \frac{\delta_{12}}{\text{poolSize}}\cdot\sigma'(z_{11}^{l})$$

③ 经过卷积层

CNN中的卷积层不是标准数学意义上的卷积，更多以一种相关体现。

1) 反向传播。

一般性假设：

① 输入有K个channel，shape为：$W \times H \times K$

② 卷积核的 shape为：$M \times N \times K$（通道数要保持一致）$\times$ NUM个

③ 输出的 shape 为：$A \times B \times NUM$

输出的误差测度的反向传播：

$$\delta^{L}_{w,h,k} = \sum_{num=1}^{NUM} \sum_{a,b} \frac{\partial C}{\partial z^{L+1}_{a,b,num}} \cdot \frac{\partial \left(\sum_{m=1}^{M} \sum_{n=1}^{N} w_{m,n,k} \cdot \sigma(z^{L}_{index}) + b \right)}{\partial z^{L}_{w,h,k}}$$

$$index = a + m - 1 + (a-1) \times (stride - 1),$$
$$b + n - 1 + (b-1) \times (stride - 1),$$
$$k$$

所有滤波器组的第k个通道上的误差的叠加：

*考虑单个滤波器组。

$$\delta^{L}_{w,h} = \sum_{a,b} \frac{\partial C}{\partial z^{L+1}_{a,b,\underset{1}{num}}} \cdot \frac{\partial \sum_{m=1}^{M} \sum_{n=1}^{N} W_{m,n} \cdot \sigma(z^{L}_{index})}{\partial z^{L}_{w,h}}$$

$$\delta^{L}_{w,h} = \sum_{a,b} \delta^{L+1}_{a,b} \cdot W_{w-(a-1)\times stride+1,\ h-(b-1)\times stride+1} \cdot \sigma'(z^{L}_{w,h})$$

退化为类似于卷积的形式，含有stride

关于经过卷积层，误差测度的反向传播公式修正：

$$\delta^{l}_{w,h,i} = \sum_{k}\sum_{a,b} \delta^{l+1}_{a,b} \cdot W_{w-(a-1)\times stride,\ h-(b-1)\times stride,\ k,\ i} \cdot \sigma'(z^{l}_{w,h})$$

其中，i表示第i个滤波器组，k表示通道的索引，$\delta^{l}_{w,h,i}$中的下标i表示输入的第i个通道。

① 误差测度的反向传播.

1) padding = valid，stride = 1

2) padding = valid，stride 不为1，有多余元素

3) padding = Same 的情形.

之前，依据个例提出的公式不具有大普适性，下面重新进行讨论.

① 需要考虑特殊情形和正向传播中是否有元素未参与运算.

总体的思路不变：<对于 padding = valid>

① 插值

② 边缘填充

③ 作相关运算.

$$\left\lfloor \frac{5-2}{2} + 1 \right\rfloor = 2$$

例：考虑输入为：

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & \cdot & \cdot & \cdot & a_{25} \\ \vdots & & & & \vdots \\ a_{51} & \cdot & \cdot & \cdot & a_{55} \end{bmatrix}$$

卷积核为：

$$\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}$$

stride = 2

输出的误差测度为：

$$\begin{bmatrix} \delta^{l+1}_{11} & \delta^{l+1}_{12} \\ \delta^{l+1}_{21} & \delta^{l+1}_{22} \end{bmatrix}$$

$$\delta^{l}_{11} = w_{11} \cdot \delta^{l+1}_{11}$$

$$\delta^{l}_{12} = w_{12} \cdot \delta^{l+1}_{11}$$

$$\delta^{l}_{13} = w_{11} \cdot \delta^{l+1}_{12}$$

$$\delta^{l}_{14} = w_{12} \cdot \delta^{l+1}_{12}$$

$$\delta^{l}_{15} = 0$$

$$\vdots$$

计算过程.

① 插0值

$$\begin{bmatrix} \delta^{l+1}_{11} & 0 & \delta^{l+1}_{12} \\ 0 & 0 & 0 \\ \delta^{l+1}_{21} & 0 & \delta^{l+1}_{22} \end{bmatrix}$$

③ 结果填充

$$\begin{bmatrix} [\delta^{l}]_{4\times4} & \begin{matrix}0\\0\\0\\0\end{matrix} \\ 0\ 0\ 0\ 0 & 0 \end{bmatrix}$$

$(2-1)\cdot 2 = 4$

② 填充 <加入有效值测度>

$$\left\lfloor \frac{x-2}{stride=1} + 1 \right\rfloor = 4. \quad (x-3)/2 = 1$$

$x = 5$.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \delta^{l+1}_{11} & 0 & \delta^{l+1}_{12} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \delta^{l+1}_{21} & 0 & \delta^{l+1}_{22} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} * rot180^{\circ} \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}$$

② padding = Same 的情形.

假定输入为:

$$\begin{bmatrix} a^{11} & a^{12} & a^{13} & a^{14} & a^{15} \\ a^{21} & a^{22} & a^{23} & a^{24} & a^{25} \\ a^{31} & a^{32} & a^{33} & a^{34} & a^{35} \\ a^{41} & a^{42} & a^{43} & a^{44} & a^{45} \\ a^{51} & a^{52} & a^{53} & a^{54} & a^{55} \end{bmatrix} * \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix} \longrightarrow \begin{bmatrix} \delta^{l+1}_{11} & \delta^{l+1}_{12} & \delta^{l+1}_{13} & \delta^{l+1}_{14} & \delta^{l+1}_{15} \\ \delta^{l+1}_{21} & \delta^{l+1}_{22} & \delta^{l+1}_{23} & \delta^{l+1}_{24} & \delta^{l+1}_{25} \\ \vdots & & & \vdots & \vdots \\ \delta^{l+1}_{51} & \cdots & & & \delta^{l+1}_{55} \end{bmatrix}$$

stride = 2

要得到同样大小的输出.

原始输入加入 padding_zero: $\lfloor \frac{x-3}{stride} + 1 \rfloor = 5 \quad x = 11$

$(x-5)/2 = 3$. 两边各加3个0.

$$\begin{bmatrix} 0\,0\,0 & & & & & & 0\,0\,0 \\ 0\,0\,0 & & & & & & 0\,0\,0 \\ 0\,0\,0 & & & & & & 0\,0\,0 \\ & a^{11} & a^{12} & a^{13} & a^{14} & a^{15} & \\ & \vdots & & & & \vdots & \\ & a^{51} & \cdots & & & a^{55} & \\ 0\,0\,0 & & & & & & 0\,0\,0 \\ 0\,0\,0 & & & & & & 0\,0\,0 \\ 0\,0\,0 & & & & & & 0\,0\,0 \end{bmatrix}$$

1) 先对 δ^{l+1} 作插值. 后的维度为 9×9

2) 补0: $\lfloor \frac{x-3}{1} \rfloor + 1 = 11 \quad x = 13 \quad (13-9)/2 = 2$个0

3)
$$\begin{bmatrix} 0\,0 & & & & 0\,0 \\ 0\,0 & & & & 0\,0 \\ & \delta^{l+1}_{11} & \cdots & \delta^{l+1}_{15} & \\ & \vdots & & \vdots & \\ & \delta^{l+1}_{51} & \cdots & \delta^{l+1}_{55} & \\ 0\,0 & & & & 0\,0 \\ 0\,0 & & & & 0\,0 \end{bmatrix} * rot180^\circ\left(\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix}\right)$$

4) 取有效值. 未 padding 前的维度.

注: 因为 padding = same 的情形下, 不会产生元素未计入卷积的情形, 因此和 padding = valid 情形有所区别.

卷积核的参数更新：

之前有：$\frac{\partial C}{\partial w_{m,n,k}} = \sum_{a,b} \delta^{l+1}_{a,b,i} \cdot a^{l}_{m+(a-1)\cdot stride,\ n+(b-1)\cdot stride,\ k}$

对应于第i个滤波器组的第k个通道的参数更新：当选用的输入和滤波器都为单核单通道时，退化为：

$$\frac{\partial C}{\partial w_{m,n}} = \sum_{a,b} \delta^{l+1}_{a,b} \cdot a^{l}_{m+(a-1)\times stride,\ n+(b-1)\times stride}$$ → 卷积层正向传播的stride.

① padding = Valid 举例：

假设输入为 4×4 size

kernel stride=2　　δ^{l+1}的size为 2×2

$$\begin{bmatrix} a^l_{11} & a^l_{12} & a^l_{13} & a^l_{14} \\ a^l_{21} & a^l_{22} & a^l_{23} & a^l_{24} \\ a^l_{31} & a^l_{32} & a^l_{33} & a^l_{34} \\ a^l_{41} & a^l_{42} & a^l_{43} & a^l_{44} \end{bmatrix} * \begin{bmatrix} w^{11} & w^{12} \\ w^{21} & w^{22} \end{bmatrix} \Rightarrow \begin{bmatrix} \delta^{l+1}_{11} & \delta^{l+1}_{12} \\ \delta^{l+1}_{21} & \delta^{l+1}_{22} \end{bmatrix}$$

① 先插值0　stride-1=1

$$\begin{bmatrix} \delta^{l+1}_{11} & 0 & \delta^{l+1}_{12} \\ 0 & 0 & 0 \\ \delta^{l+1}_{21} & 0 & \delta^{l+1}_{22} \end{bmatrix} \rightarrow \text{不需对 } a^l \text{ 作 padding} \rightarrow \begin{bmatrix} a^l_{11} & a^l_{12} & \cdots & a^l_{14} \\ \vdots & & & \vdots \\ a^l_{41} & \cdots & \cdots & a^l_{44} \end{bmatrix} * \begin{bmatrix} \delta^{l+1}_{11} & 0 & \delta^{l+1}_{12} \\ 0 & 0 & 0 \\ \delta^{l+1}_{21} & 0 & \delta^{l+1}_{22} \end{bmatrix}$$

② 出现多余元素时.

stride = 2

$$\begin{bmatrix} w^{11} & w^{12} & w^{13} \\ w^{21} & w^{22} & w^{23} \\ w^{31} & w^{32} & w^{33} \end{bmatrix} \rightarrow \begin{bmatrix} \delta^{l+1}_{11} \end{bmatrix}$$

1)插值 → 2)检查有效输入为 $\begin{bmatrix} a^l_{11} & a^l_{12} & a^l_{13} \\ a^l_{21} & a^l_{22} & a^l_{23} \\ a^l_{31} & a^l_{32} & a^l_{33} \end{bmatrix}$　3)作相关　~~4)填充0~~

② padding = SAME 时.

假设 kernel为：$\lfloor \frac{x-2}{3} + 1 \rfloor = 4$　　$\frac{11-4}{2} = 3$

$$\begin{bmatrix} a^l_{11} & a^l_{12} & a^l_{13} & a^l_{14} \\ a^l_{21} & \cdots & \cdots & a^l_{24} \\ a^l_{31} & \cdots & \cdots & a^l_{34} \\ a^l_{41} & \cdots & \cdots & a^l_{44} \end{bmatrix} \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \rightarrow \delta^{l+1}_{4\times 4} \rightarrow \text{插值后0加} \rightarrow \text{不需作padding}$$

$\lfloor \frac{x-10}{1} + 1 \rfloor = 2$

Stride = 3　→ 结果输出.