Feedforward NN 与 Convolutional NN 中的 Back Propagation

1 引入

经过 forward propagation 之后，我们可以得到当前神经网络的 cost function $J(\boldsymbol{W}, \boldsymbol{b})$，为了找到 cost function 的最小值，我们需要沿其梯度方向更新每层网络的权重矩阵 $\boldsymbol{W}$ 和偏置列向量 $\boldsymbol{b}$ 的值：

$$\boldsymbol{W}^{[l]} := \boldsymbol{W}^{[l]} - \frac{\partial J}{\partial \boldsymbol{W}^{[l]}} \tag{1}$$

$$\boldsymbol{b}^{[l]} := \boldsymbol{b}^{[l]} - \frac{\partial J}{\partial \boldsymbol{b}^{[l]}} \tag{2}$$

所以我们需要计算 $\frac{\partial J}{\partial \boldsymbol{W}^{[l]}}$ 和 $\frac{\partial J}{\partial \boldsymbol{b}^{[l]}}$。

由于直接求数值积分的高计算消耗，我们利用 computation graph 的逻辑，利用链式法则逐层迭代地对 $\nabla_{\boldsymbol{W}} J$ 和 $\nabla_{\boldsymbol{b}} J$ 进行计算。

2 Feedforward NN 的 Back Propagation

由于

$$\frac{\partial J}{\partial \boldsymbol{W}^{[l]}} = \frac{\partial J}{\partial \boldsymbol{Z}^{[l]}} \frac{\partial \boldsymbol{Z}^{[l]}}{\partial \boldsymbol{W}^{[l]}} \tag{3}$$

所以我们想要先计算 $\frac{\partial J}{\partial \boldsymbol{Z}^{[l]}}$，再利用链式法则计算 $\frac{\partial J}{\partial \boldsymbol{W}^{[l]}}$。

又由于

$$\frac{\partial J}{\partial \boldsymbol{Z}^{[l]}} = \frac{\partial J}{\partial \boldsymbol{A}^{[l]}} \frac{\partial \boldsymbol{A}^{[l]}}{\partial \boldsymbol{Z}^{[l]}} \tag{4}$$

所以我们最终发现应该先计算 $\frac{\partial J}{\partial \boldsymbol{A}^{[l]}}$。

2.1 计算 $\partial J/\partial \boldsymbol{A}^{[l]}$

对于 J 的全微分，有：

$$\mathrm{d}J = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\partial J}{\partial Z_{ij}^{[l]}}\mathrm{d}Z_{ij}^{[l]} = \sum_{i,j}\left(\nabla_{Z^{[l]}}J\right)_{ij}\mathrm{d}Z_{ij}^{[l]} = \left\langle \nabla_{\boldsymbol{Z}^{[l]}}J, \mathrm{d}Z^{[l]}\right\rangle_F = \mathrm{tr}\left(\left(\nabla_{\boldsymbol{Z}^{[l]}}J\right)^\top \mathrm{d}\boldsymbol{Z}^{[l]}\right) \tag{5}$$

代入 $\mathrm{d}\boldsymbol{Z}^{[l]} = \boldsymbol{W}^{[l]}\mathrm{d}\boldsymbol{A}^{[l-1]}$，得：

$$\mathrm{d}J = \mathrm{tr}\left(\left(\nabla_{Z^{[l]}}J\right)^\top \boldsymbol{W}^{[l]}\mathrm{d}\boldsymbol{A}^{[l-1]}\right) \tag{6}$$

所以：

$$\nabla_{\boldsymbol{A}^{[l-1]}}J = \left(\left(\nabla_{Z^{[l]}}J\right)^\top \boldsymbol{W}^{[l]}\right)^\top \tag{7}$$

由于 $(\boldsymbol{AB})^\top = \boldsymbol{B}^\top\boldsymbol{A}^\top$，所以：

$$\frac{\partial J}{\partial \boldsymbol{A}^{[l-1]}} = \boldsymbol{W}^{[l]\top}\frac{\partial J}{\partial \boldsymbol{Z}^{[l]}} \tag{8}$$

从 $\boldsymbol{A}^{[l-1]}$ 推导 $\boldsymbol{A}^{[l]}$：

$$\frac{\partial J}{\partial \boldsymbol{A}^{[l]}} = \boldsymbol{W}^{[l+1]\top}\frac{\partial J}{\partial \boldsymbol{Z}^{[l+1]}} \tag{9}$$

2.2 计算 $\partial J/\partial \boldsymbol{Z}^{[l]}$

和 (4) 式相同地，有：

$$\frac{\partial J}{\partial \boldsymbol{Z}^{[l]}} = \frac{\partial J}{\partial \boldsymbol{A}^{[l]}}\frac{\partial \boldsymbol{A}^{[l]}}{\partial \boldsymbol{Z}^{[l]}}$$

我们刚才已经求得了 $\dfrac{\partial J}{\partial \boldsymbol{A}^{[l]}}$。这里我们分析一下 $\dfrac{\partial \boldsymbol{A}^{[l]}}{\partial \boldsymbol{Z}^{[l]}}$。由于 activation function element-wise 地作用在 $\boldsymbol{Z}$，所以 $\dfrac{\partial \boldsymbol{A}}{\partial \boldsymbol{Z}}$ 得到的 Jacobian 为对角矩阵。因此可以用 Hadamard 乘（也就是 element-wise 乘法）实现同样效果，即：

$$\begin{aligned}\frac{\partial J}{\partial \boldsymbol{Z}^{[l]}} &= \frac{\partial J}{\partial \boldsymbol{A}^{[l]}}\mathbf{J}_g = \frac{\partial J}{\partial \boldsymbol{A}^{[l]}}\mathrm{diag}\left(g^{[l]\prime}\left(\boldsymbol{Z}^{[l]}\right)\right)\\ &= \frac{\partial J}{\partial \boldsymbol{A}^{[l]}} \odot g^{[l]\prime}\left(\boldsymbol{Z}^{[l]}\right) \qquad (10)\\ &\left(= \boldsymbol{W}^{[l+1]\top}\frac{\partial J}{\partial \boldsymbol{Z}^{[l+1]}} \odot g^{[l]\prime}\left(\boldsymbol{Z}^{[l]}\right)\right) \qquad (11)\end{aligned}$$

2.3 计算 $\partial J/\partial \boldsymbol{W}^{[l]}$

和 (3) 式相同地，有：

$$\frac{\partial J}{\partial \boldsymbol{W}^{[l]}} = \frac{\partial J}{\partial \boldsymbol{Z}^{[l]}}\frac{\partial \boldsymbol{Z}^{[l]}}{\partial \boldsymbol{W}^{[l]}}$$

我们刚才已经求得了 $\dfrac{\partial J}{\partial \boldsymbol{Z}^{[l]}}$。这里我们可以接着用 vec-Kronecker 恒等式求 $\dfrac{\partial \boldsymbol{Z}^{[l]}}{\partial \boldsymbol{W}^{[l]}}$，或者模仿 $\dfrac{\partial J}{\partial \boldsymbol{A}^{[l]}}$ 的推导，从全微分和迹运算进行推导（相对简单）。

2.3.1 硬求 $\partial \boldsymbol{Z}^{[l]}/\partial \boldsymbol{W}^{[l]}$

对式 $\mathrm{d}\boldsymbol{Z}^{[l]} = \boldsymbol{W}^{[l]}\mathrm{d}\boldsymbol{A}^{[l-1]}$ 的两侧同时做 vec 运算并插入单位矩阵，根据恒等式 $\operatorname{vec}(\boldsymbol{ABC}) = \left(\boldsymbol{C}^{\top} \otimes \boldsymbol{A}\right)\operatorname{vec}(\boldsymbol{B})$ 得到：

$$\operatorname{vec}\left(\mathrm{d}\boldsymbol{Z}^{[l]}\right) = \left(\boldsymbol{A}^{[l-1]\top} \otimes I_{n_l}\right)\operatorname{vec}\left(\mathrm{d}\boldsymbol{W}^{[l]}\right) \tag{12}$$

其中“$\otimes$”表示 Kronecker 积。所以 $\dfrac{\partial \boldsymbol{Z}^{[l]}}{\partial \boldsymbol{W}^{[l]}}$ 得到的 Jacobian（vec 形式）为：

$$\frac{\partial \operatorname{vec}\left(\boldsymbol{Z}^{[l]}\right)}{\partial \operatorname{vec}\left(\boldsymbol{W}^{[l]}\right)} = \boldsymbol{A}^{[l-1]\top} \otimes I_{n_l} \tag{13}$$

同样改写 $\dfrac{\partial J}{\partial \boldsymbol{W}^{[l]}}$ 为 vec 形式，有：

$$\frac{\partial J}{\partial \operatorname{vec}\left(\boldsymbol{W}^{[l]}\right)} = \left(\frac{\partial \operatorname{vec}\left(\boldsymbol{Z}^{[l]}\right)}{\partial \operatorname{vec}\left(\boldsymbol{W}^{[l]}\right)}\right)^{\top} \frac{\partial J}{\partial \operatorname{vec}\left(\boldsymbol{Z}^{[l]}\right)} \tag{14}$$

将式 (12) 代入式 (13)，有：

$$\operatorname{vec}\left(\frac{\partial J}{\partial \boldsymbol{W}^{[l]}}\right) = \left(\boldsymbol{A}^{[l-1]\top} \otimes I_{n_l}\right)^{\top} \operatorname{vec}\left(\frac{\partial J}{\partial \boldsymbol{Z}^{[l]}}\right) \tag{15}$$

由于 $(\boldsymbol{A} \otimes \boldsymbol{B})^{\top} = \boldsymbol{A}^{\top} \otimes \boldsymbol{B}^{\top}$，所以：

$$\operatorname{vec}\left(\frac{\partial J}{\partial \boldsymbol{W}^{[l]}}\right) = \left(\boldsymbol{A}^{[l-1]} \otimes I_{n_l}\right)\operatorname{vec}\left(\frac{\partial J}{\partial \boldsymbol{Z}^{[l]}}\right) \tag{16}$$

根据恒等式 $\operatorname{vec}(\boldsymbol{ABC}) = \left(\boldsymbol{C}^{\top} \otimes \boldsymbol{A}\right)\operatorname{vec}(\boldsymbol{B})$ 得到：

$$\operatorname{vec}\left(\frac{\partial J}{\partial \boldsymbol{W}^{[l]}}\right) = I_{n_l}\frac{\partial J}{\partial \boldsymbol{Z}^{[l]}}\boldsymbol{A}^{[l-1]\top} \tag{17}$$

因为 I_{n_l} 左乘并不改变数值，所以：

$$\operatorname{vec}\left(\frac{\partial J}{\partial \boldsymbol{W}^{[l]}}\right) = \frac{\partial J}{\partial \boldsymbol{Z}^{[l]}}\boldsymbol{A}^{[l-1]\top} \Rightarrow \frac{\partial J}{\partial \boldsymbol{W}^{[l]}} = \frac{\partial J}{\partial \boldsymbol{Z}^{[l]}}\boldsymbol{A}^{[l-1]\top} \tag{18}$$

2.3.2 J 的全微分

由 J 的全积分

$$\mathrm{d}J = \operatorname{tr}\left(\left(\nabla_{\boldsymbol{Z}^{[l]}}J\right)^{\top}\mathrm{d}\boldsymbol{Z}^{[l]}\right)$$

代入 $\mathrm{d}\boldsymbol{Z}^{[l]} = \mathrm{d}\boldsymbol{W}^{[l]}\boldsymbol{A}^{[l-1]}$，得：

$$\mathrm{d}J = \operatorname{tr}\left(\left(\nabla_{\boldsymbol{Z}^{[l]}}J\right)^{\top}\mathrm{d}\boldsymbol{W}^{[l]}\boldsymbol{A}^{[l-1]}\right)$$

由迹运算的循环不变性即 $\operatorname{tr}(\boldsymbol{ABC}) = \operatorname{tr}(\boldsymbol{CAB})$ 得到：

$$\operatorname{tr}\left(\left(\nabla_{\boldsymbol{Z}^{[l]}}J\right)^{\top}\mathrm{d}\boldsymbol{W}^{[l]}\boldsymbol{A}^{[l-1]}\right) = \operatorname{tr}\left(\boldsymbol{A}^{[l-1]}\left(\nabla_{\boldsymbol{Z}^{[l]}}J\right)^{\top}\mathrm{d}\boldsymbol{W}^{[l]}\right) = \operatorname{tr}\left(\left(\nabla_{\boldsymbol{Z}^{[l]}}J\boldsymbol{A}^{[l-1]\top}\right)^{\top}\mathrm{d}\boldsymbol{W}^{[l]}\right) \tag{19}$$

所以：

$$\frac{\partial J}{\partial \boldsymbol{W}^{[l]}} = \nabla_{\boldsymbol{Z}^{[l]}}J\boldsymbol{A}^{[l-1]\top} = \frac{\partial J}{\partial \boldsymbol{Z}^{[l]}}\boldsymbol{A}^{[l-1]\top} \tag{20}$$

2.4 计算 $\partial J/\partial \boldsymbol{b}^{[l]}$

和计算 $\dfrac{\partial J}{\partial \boldsymbol{W}^{[l]}}$ 的过程类似地，我们也有两种方法计算 $\dfrac{\partial J}{\partial \boldsymbol{b}^{[l]}}$。这里我们只介绍第二种迹运算的方法。第一种利用 vec-Kronecker 恒等式的方法留给读者作为练习。

由 forward propagation 的公式：

$$\boldsymbol{Z}^{[l]} = \boldsymbol{W}^{[l]}\boldsymbol{A}^{[l-1]} + \boldsymbol{b}^{[l]}\mathbf{1}_{1\times m},\ \boldsymbol{b} \in \mathbb{R}^{n_l\times 1}$$

可以得到：

$$\mathrm{d}\boldsymbol{Z}^{[l]} = \mathrm{d}\left(\boldsymbol{b}^{[l]}\mathbf{1}_{1\times m}\right) = \left(\mathrm{d}\boldsymbol{b}^{[l]}\right)\mathbf{1}_{1\times m} + \boldsymbol{b}^{[l]}\underbrace{(\mathrm{d}\mathbf{1}_{1\times m})}_{=0} = \mathrm{d}\boldsymbol{b}^{[l]}\mathbf{1}_{1\times m} \tag{21}$$

由 J 的全积分

$$\mathrm{d}J = \mathrm{tr}\left(\left(\nabla_{\boldsymbol{Z}^{[l]}}J\right)^{\top}\mathrm{d}\boldsymbol{Z}^{[l]}\right)$$

代入 $\mathrm{d}\boldsymbol{Z}^{[l]} = \mathrm{d}\boldsymbol{b}^{[l]}\mathbf{1}_{1\times m}$（其中 m 为样本数），得：

$$\mathrm{tr}\left(\left(\nabla_{\boldsymbol{Z}^{[l]}}J\right)^{\top}\mathrm{d}\boldsymbol{Z}^{[l]}\right) = \mathrm{tr}\left(\left(\nabla_{\boldsymbol{Z}^{[l]}}J\right)^{\top}\mathrm{d}\boldsymbol{b}^{[l]}\mathbf{1}_{1\times m}\right) \tag{22}$$

由迹运算的循环不变性，得：

$$\mathrm{tr}\left(\left(\nabla_{\boldsymbol{Z}^{[l]}}J\right)^{\top}\mathrm{d}\boldsymbol{b}^{[l]}\mathbf{1}_{1\times m}\right) = \mathrm{tr}\left(\left(\nabla_{\boldsymbol{Z}^{[l]}}J\left(\mathbf{1}_{1\times m}\right)^{\top}\right)^{\top}\mathrm{d}\boldsymbol{b}^{[l]}\right) = \mathrm{tr}\left(\left(\nabla_{\boldsymbol{Z}^{[l]}}J\mathbf{1}_{m\times 1}\right)^{\top}\mathrm{d}\boldsymbol{b}^{[l]}\right) \tag{23}$$

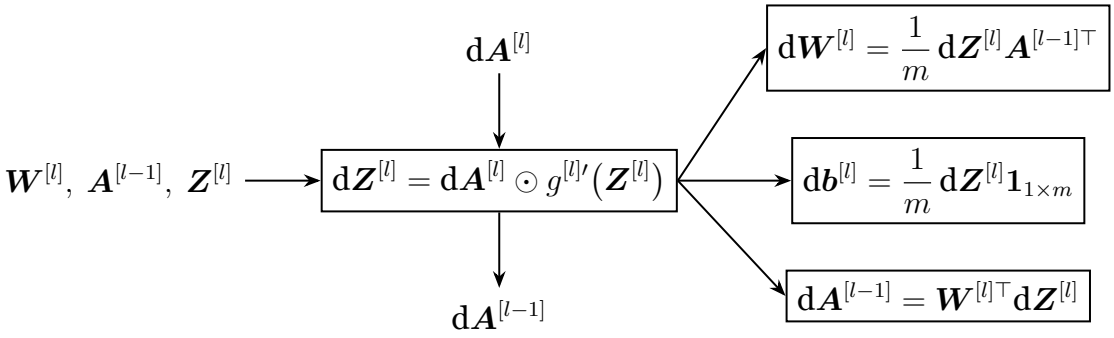
所以：

$$\frac{\partial J}{\partial \boldsymbol{b}^{[l]}} = \frac{\partial J}{\partial \boldsymbol{Z}^{[l]}}\mathbf{1}_{m\times 1} \tag{24}$$

这即等价为对 $\dfrac{\partial J}{\partial \boldsymbol{Z}^{[l]}}$ 逐行求和。故在程序中写作 `np.sum(dZ, axis=1, keepdims=True)`。

2.5 总结：Feedforward NN 的 Back Propagation

整合之前的推导，我们可以得到：



3 Convolutional Layer 的 Back Propagation

注意：

1. 我们用 “$*$” 表示 convolution，用 “$\star$” 表示 cross-correlation。
2. 本节不会在推导中大量使用矩阵运算，因为卷积操作可以用行列索引方便表示。
3. 下标 $[a,b]$ 表示第 a 行第 b 列的元素，索引从 $[0,0]$ 开始。$f^{[l]}$ 表示 l 层的 kernel/filter 的行/列数。l 层的激活张量 $\mathbf{A}^{[l]}$ 的 `shape` 为 $[n_H^{[l-1]}, n_W^{[l-1]}, n_c^{[l-1]}, m]$。其中 m 为样本数，$n_c^{[l-1]}$ 为 $l-1$ 层的 channel 数。
4. 阅读的时候可以认为每层神经网络的 kernel/filter 的数量为 1，且该 kernel/filter 的层数也为 1。

3.1 从 $\partial J/\partial \boldsymbol{Z}^{[l]}$ 推导 $\partial J/\partial \boldsymbol{W}^{[l]}$

由链式法则：

$$\frac{\partial J}{\partial w_{[a,b]}^{[l]}} = \sum_{x=0}^{n_H^{[l]}} \sum_{y=0}^{n_W^{[l]}} \frac{\partial J}{\partial z_{[x,y]}^{[l]}} \frac{\partial z_{[x,y]}^{[l]}}{\partial w_{[a,b]}^{[l]}} \tag{25}$$

其中，$z_{[x,y]}^{[l]}$ 可以表示为：

$$z_{[x,y]}^{[l]} = a^{[l-1]} \star w^{[l]} + b^{[l]} = \sum_{a=0}^{f^{[l]}} \sum_{b=0}^{f^{[l]}} a_{[x+a,y+b]}^{[l-1]} w_{[a,b]}^{[l]} + b^{[l]} \tag{26}$$

即对应范围的卷积再加上偏置项（在 implement forward propagation from scratch 的时候其实也是将对应的区域 slice 出来再进行卷积操作）。

将 (26) 式代入 (25) 式得到：

$$\frac{\partial J}{\partial w_{[a,b]}^{[l]}} = \sum_{x=0}^{n_H^{[l]}} \sum_{y=0}^{n_W^{[l]}} \frac{\partial J}{\partial z_{[x,y]}^{[l]}} \frac{\partial \left(\sum_{a'=0}^{f^{[l]}} \sum_{b'=0}^{f^{[l]}} a_{[x+a',y+b']}^{[l-1]} w_{[a',b']}^{[l]} + b^{[l]} \right)}{\partial w_{[a,b]}^{[l]}} \tag{27}$$

注意！当且仅当 $a' = a$ 且 $b' = b$ 的时候第二项偏导不为 0。所以：

$$\frac{\partial J}{\partial w_{[a,b]}^{[l]}} = \sum_{x=0}^{n_H^{[l]}} \sum_{y=0}^{n_W^{[l]}} \frac{\partial J}{\partial z_{[x,y]}^{[l]}} a_{[x+a,y+b]}^{[l-1]} = \sum_{x=0}^{n_H^{[l]}} \sum_{y=0}^{n_W^{[l]}} \frac{\partial J}{\partial z_{[x,y]}^{[l]}} g\left(z_{[x+a,y+b]}^{[l-1]} \right) \tag{28}$$

所以：

$$\begin{aligned} \frac{\partial J}{\partial \boldsymbol{W}^{[l]}} &= \frac{\partial J}{\partial \boldsymbol{Z}^{[l]}} \star g\left(\boldsymbol{Z}^{[l-1]} \right) \\ &\left(= \frac{\partial J}{\partial \boldsymbol{Z}^{[l]}} * g\left(\text{ROT180}\left(\boldsymbol{Z}^{[l-1]} \right) \right) \right) \end{aligned} \tag{29}$$

其中 ROT180 表示上下翻转一次，接着左右翻转一次。

3.2 从 $\partial J/\partial \boldsymbol{Z}^{[l]}$ 推导 $\partial J/\partial \boldsymbol{b}^{[l]}$

与式 (27) 类似地：

$$\frac{\partial J}{\partial b^{[l]}} = \sum_{x=0}^{n_H^{[l]}} \sum_{y=0}^{n_W^{[l]}} \frac{\partial J}{\partial z_{[x,y]}^{[l]}} \frac{\partial \left(\sum_{a'=0}^{f^{[l]}} \sum_{b'=0}^{f^{[l]}} a_{[x+a',y+b']}^{[l-1]} w_{[a',b']}^{[l]} + b^{[l]} \right)}{\partial b^{[l]}} \tag{30}$$

显然第二项偏导等于 1。所以：

$$\frac{\partial J}{\partial b^{[l]}} = \sum_{x=0}^{n_H^{[l]}} \sum_{y=0}^{n_W^{[l]}} \frac{\partial J}{\partial z_{[x,y]}^{[l]}} = \mathbf{1}_{n_H^{[l]} \times n_W^{[l]}} \star \frac{\partial J}{\partial \boldsymbol{Z}^{[l]}} \tag{31}$$

3.3 推导 $\partial J/\partial \boldsymbol{Z}^{[l]}$

综合式 (29) 与式 (31)，可以得到：

$$\frac{\partial J}{\partial w^{[l]}} = g\left(\boldsymbol{Z}^{[l-1]}\right) \star \frac{\partial J}{\partial \boldsymbol{Z}^{[l]}}$$
$$\frac{\partial J}{\partial b^{[l]}} = \mathbf{1}_{n_H^{[l]} \times n_W^{[l]}} \star \frac{\partial J}{\partial \boldsymbol{Z}^{[l]}}$$

所以我们要求得 $\dfrac{\partial J}{\partial \boldsymbol{Z}^{[l]}}$。

显然地，我们有：

$$z_{[x',y']}^{[l+1]} = \sum_{a=0}^{f^{[l+1]}} \sum_{b=0}^{f^{[l+1]}} \left(g\left(z_{[x'+a,y'+b]}^{[l]} \right) \right) w_{[a,b]}^{[l+1]} + b^{[l+1]} \tag{32}$$

$$\frac{\partial J}{\partial z_{[x,y]}^{[l]}} = \sum_{x'=0}^{n_H^{[l+1]}} \sum_{y'=0}^{n_W^{[l+1]}} \frac{\partial J}{\partial z_{[x',y']}^{[l+1]}} \frac{\partial z_{[x',y']}^{[l+1]}}{\partial z_{[x,y]}^{[l]}} \tag{33}$$

将式 (32) 代入式 (33)，得到：

$$\frac{\partial J}{\partial z_{[x,y]}^{[l]}} = \sum_{x'=0}^{n_H^{[l+1]}} \sum_{y'=0}^{n_W^{[l+1]}} \frac{\partial J}{\partial z_{[x',y']}^{[l+1]}} \frac{\partial \sum_{a=0}^{f^{[l+1]}} \sum_{b=0}^{f^{[l+1]}} \left(g\left(z_{[x'+a,y'+b]}^{[l]} \right) \right) w_{[a,b]}^{[l+1]} + b^{[l+1]}}{\partial z_{[x,y]}^{[l]}} \tag{34}$$

同样地，当且仅当 $x' + a = x$ 且 $y' + b = y$ 时第二项偏导不为 0。所以：

$$\frac{\partial J}{\partial z_{[x,y]}^{[l]}} = \sum_{x'=0}^{n_H^{[l+1]}} \sum_{y'=0}^{n_W^{[l+1]}} \frac{\partial J}{\partial z_{[x',y']}^{[l+1]}} g'\left(z_{[x,y]}^{[l]} \right) w_{[a,b]}^{[l+1]} \tag{35}$$

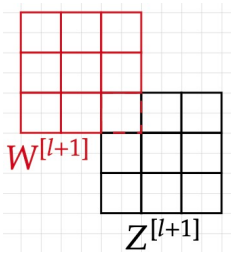
代入 $a = x - x'$, $b = y - y'$ 得到：

$$\frac{\partial J}{\partial z_{[x,y]}^{[l]}} = \sum_{x'=0}^{n_H^{[l+1]}} \sum_{y'=0}^{n_W^{[l+1]}} \frac{\partial J}{\partial z_{[x',y']}^{[l+1]}} g'\left(z_{[x,y]}^{[l]} \right) w_{[x-x',y-y']}^{[l+1]} \tag{36}$$

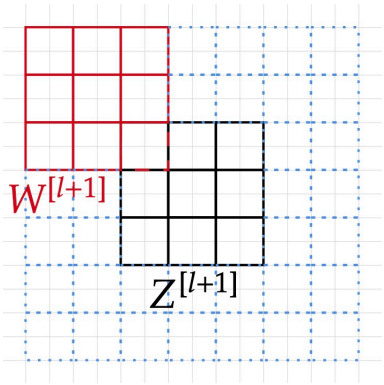
因为是标量乘法所以可以移项得到：

$$\begin{aligned}\frac{\partial J}{\partial z_{[x,y]}^{[l]}} &= \sum_{x'=0}^{n_H^{[l+1]}} \sum_{y'=0}^{n_W^{[l+1]}} \frac{\partial J}{\partial z_{[x',y']}^{[l+1]}} \cdot w_{[x-x',y-y']}^{[l+1]} \cdot g'\left(z_{[x,y]}^{[l]}\right) = \frac{\partial J}{\partial \boldsymbol{Z}^{[l+1]}} *_{full} \boldsymbol{W}^{[l+1]} \odot g'(\boldsymbol{Z}^{[l]}) \\ &= \frac{\partial J}{\partial \boldsymbol{Z}^{[l+1]}} \star_{full} \text{ROT180}\left(\boldsymbol{W}^{[l+1]}\right) \odot g'(\boldsymbol{Z}^{[l]}) \end{aligned} \tag{37}$$

注意其中 $*_{full}$ 提示这是一个 full convolution，为什么呢？例如假设 $[x', y'] = [0, 0]$，$[x, y] = [2, 2]$，则卷积运算要在 $\boldsymbol{Z}^{[l+1]}$ 左上角覆盖权重矩阵的 $[x - x', y - y'] = [2, 2]$ 索引。如图：

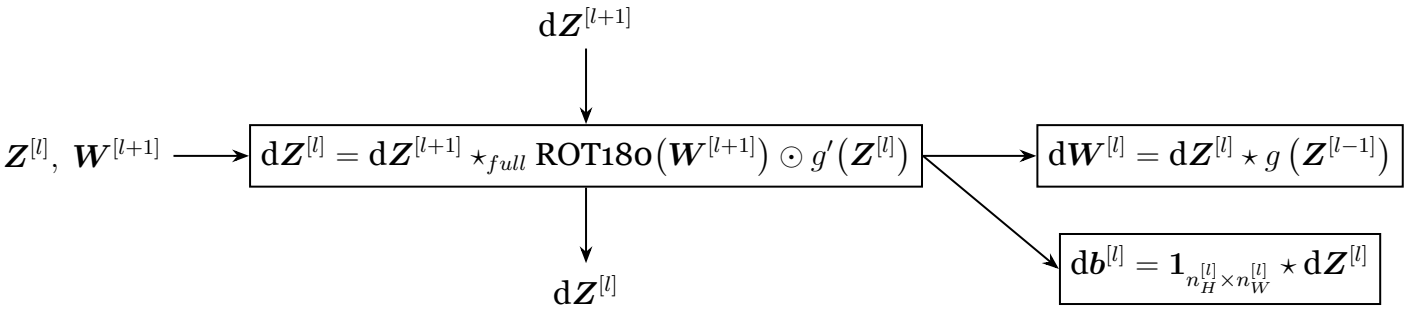


即在卷积运算中 $\boldsymbol{W}^{[l+1]}$ 不能局限在 $\boldsymbol{Z}^{[l+1]}$ 范围中，必须在 $\boldsymbol{Z}^{[l+1]}$ 四周额外用 0 垫衬 (pad) $n_H^{[l+1]} - n_H^{[l]} = f^{[l+1]} - 1$ 的大小：



3.4 总结：Convolutional Layer 的 Back Propagation

整合之前的推导，我们可以得到：



4 Pooling Layer 的 Back Propagation

对于 max pooling，在 back prop 时将某一池化像素的梯度都回传给上一层中池化范围内的最大值所在像素，其余像素梯度为 0。

对于 average pooling，则把梯度的各个池化区域的值取平均后放在上一层对应的池化位置。