FDU 神经网络 3. 卷积神经网络

本文参考以下教材:

- 神经网络与深度学习 (邱锡鹏) 第 5 章
- Deep Learning (I. Goodfellow, Y. Bengio, A. Courville) Chapter $7, 9$
- 深度学习 (I. Goodfellow, Y. Bengio, A. Courville, 赵申剑等译) 第 $7, 9$ 章

欢迎批评指正!

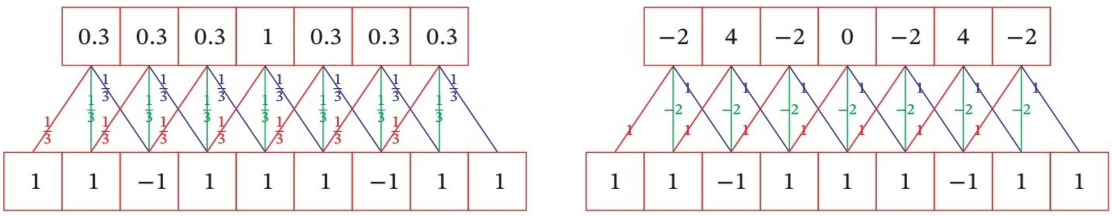
3.1 卷积

3.1.1 卷积

卷积 (convolution) 是分析数学中一种重要的运算.

- **一维卷积:**
  考虑长度为 K 的一维卷积核 $[w_1, \dots, w_K]$ 和一维信号序列 $x_1, x_2, \dots$ 的卷积:
  (简单起见，假设卷积结果 y 的首项元素的下标从 K 开始)

$$y := w \star x$$
$$y_t := \sum_{k=1}^{K} w_k x_{t-k+1} \ (\forall\, t \geq K)$$



(a) 滤波器 [1/3, 1/3, 1/3]　　(b) 滤波器 [1, −2, 1]

图 5.1　一维卷积示例

下层为输入信号序列，上层为卷积结果. 连接边上的数字为滤波器中的权重.
左图的卷积结果为近似值.

- **二维卷积:**
  考虑二维图像 $X \in \mathbb{R}^{M \times N}$ 和二维卷积核 $W \in \mathbb{R}^{(2U+1)\times(2V+1)}$ (其中 $U \ll M,\ V \ll N$) 的卷积:
  (简单起见，假设卷积结果 Y 的左上角元素的下标从 $(U+1, V+1)$ 开始，卷积核 W 中心点的下标为 $(0,0)$)

$$Y := W \star X \in \mathbb{R}^{(M-2U)\times(N-2V)}$$
$$y_{ij} := \sum_{u=-U}^{U} \sum_{v=-V}^{V} w_{uv} x_{i-u,j-v}$$
$$\text{where } U+1 \leq i \leq M-U \text{ and } V+1 \leq j \leq N-V$$

  上述卷积过程看起来就像这样:

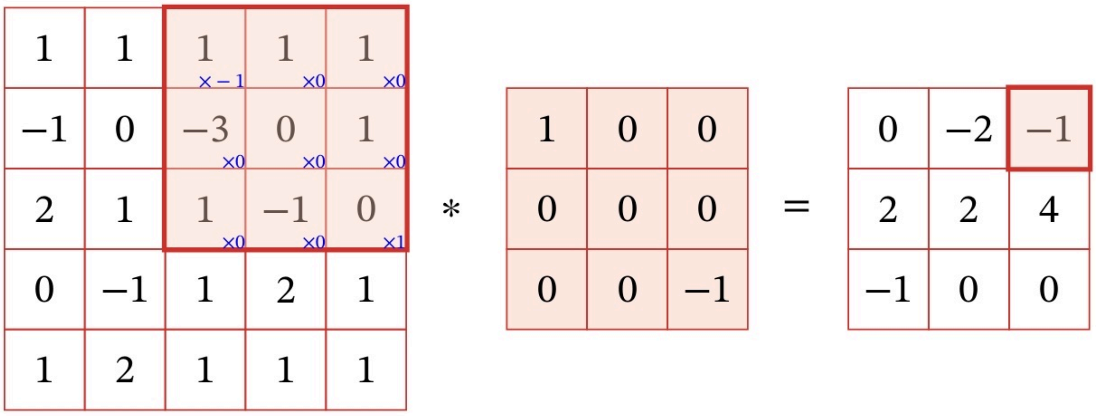


图 5.2 二维卷积示例

根据卷积定义，左图的计算需要进行卷积核翻转.

如果我们先对图像 $X \in \mathbb{R}^{M\times N}$ 进行零填充 (上下两端各补 U 个零，左右各补 V 个零)
得到 $X_{\text{padded}} \in \mathbb{R}^{(M+2U)\times(N+2V)}$ 再进行卷积运算，
那么这样得到的卷积结果 $Y = W \star X_{\text{padded}}$ 的维度和原始图像 X 一样，都是 $M \times N$ 维的:
(简单起见，假设卷积核 W 中心点的下标为 $(0,0)$)

$$
\begin{gathered}
Y := W \star X_{\text{padded}} \in \mathbb{R}^{M\times N} \\
y_{ij} := \sum_{u=-U}^{U}\sum_{v=-V}^{V} w_{uv} x_{i-u,j-v} \\
\text{where } 1 \leq i \leq M \text{ and } 1 \leq j \leq N
\end{gathered}
$$

其中 $x_{i-u,j-v}$ 在超出原始图像范围时取零 (隐式零填充)

> **写在前面:**
> 按最优化的记号，关于 z 的梯度应该用 ∇_z 符号表示 (标量对列向量求梯度得到的是同形的列向量)，
> 但为了方便我们使用 $\frac{\partial}{\partial z}$ 来代表梯度，尽管 $\frac{\partial}{\partial z}$ 符号在数学上代表导数 (标量对列向量求导数得到的是转置后同形的行向量)
> 机器学习的记号真是太混乱啦! 而且与数学的记号不兼容.
> 无论如何，现在我们认为 $\frac{\partial}{\partial z}$ 代表梯度.
> (邱锡鹏老师书上的记号也是这样默认的，不过着实让我困惑了一阵)

给定可微函数 $f: \mathbb{R}^{M\times N} \mapsto \mathbb{R}$，则我们有: **(存疑)**

$$
\begin{aligned}
\frac{\partial f(Y)}{\partial w_{u'v'}} &:= \sum_{i=1}^{M}\sum_{j=1}^{N} \frac{\partial y_{ij}}{\partial w_{u'v'}} \cdot \frac{\partial f(Y)}{\partial y_{ij}} \quad \left(\text{note that } y_{ij} := \sum_{u=-U}^{U}\sum_{v=-V}^{V} w_{uv} x_{i-u,j-v}\right) \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N} x_{i-u',j-v'} \cdot \frac{\partial f(Y)}{\partial y_{ij}} \quad (-U \leq u' \leq U, \; -V \leq v' \leq V) \\
\hline
\frac{\partial f(Y)}{\partial W} &= \frac{\partial f(Y)}{\partial Y} \star \text{rotate}(X_{\text{padded}}, 180^\circ) \\
&= \text{rotate}\left(\frac{\partial f(Y)}{\partial Y}, 180^\circ\right) \star X_{\text{padded}} \\
&= \frac{\partial f(Y)}{\partial Y} * X_{\text{padded}}
\end{aligned}
$$

其中 $*$ 是互相关运算，X_{padded} 是零填充的图像.

3.1.2 互相关

考虑二维图像 $X \in \mathbb{R}^{M\times N}$ 和二维卷积核 $W \in \mathbb{R}^{(2U+1)\times(2V+1)}$ (其中 $U \ll M,\ V \ll N$) 的互相关:
(简单起见，假设互相关结果 Y 的左上角元素的下标从 $(U+1, V+1)$ 开始，卷积核 W 中心点的下标为 $(0,0)$)

$$Y := W * X \in \mathbb{R}^{(M-2U)\times(N-2V)}$$
$$y_{ij} := \sum_{u=-U}^{U}\sum_{v=-V}^{V} w_{uv} x_{i+u,j+v}$$
$$\text{where } U+1 \le i \le M-U \text{ and } V+1 \le j \le N-V$$

互相关运算 $*$ 和卷积运算 $\star$ 的区别在于，互相关无需将卷积核旋转 180°，而卷积需要.
换言之，卷积运算相当于将卷积核旋转 180° 之后再做互相关运算:

$$W \star X = \text{rotate}(W, 180^\circ) * X$$

互相关和卷积的性质有所区别:

性质	互相关	卷积
交换律	(不满足)	$f \star g = g \star f$
结合律	(不满足)	$f \star (g \star h) = (f \star g) \star h$
分配律	$f * (g+h) = (f*g) + (f*h)$	$f \star (g+h) = (f \star g) + (f \star h)$

根据卷积的交换律可知旋转卷积核 W 还是旋转图像 X 是无关紧要的，但按照惯例我们旋转 W.
根据卷积的结合律可知多次卷积可以合并为一次卷积.

3.2 卷积神经网络

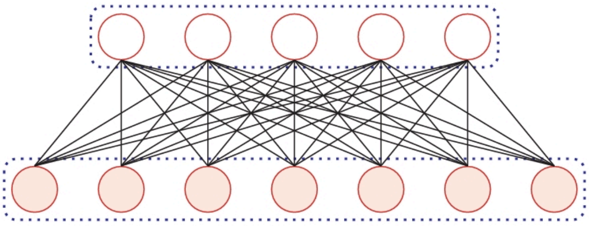
卷积神经网络一般是由卷积层、汇聚层和全连接层交叉堆叠而成的前馈神经网络.
卷积神经网络有三个结构上的特性: 局部连接、权重共享和汇聚.
这些特性使得卷积神经网络具有一定程度上的平移、缩放和旋转不变性，更适用于图像相关的任务.
和前馈神经网络相比，卷积神经网络的参数更少.

3.2.1 卷积层

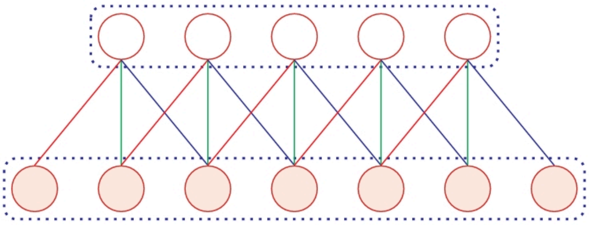
相对于全连接层来说，卷积层有两个重要性质:

- **① 局部连接:**
  卷积层的每个神经元都只与前一层中某个局部窗口内的神经元相连，有助于学习图像的局部特征.
- **② 权重共享:**
  作为参数的卷积核 $W^{(l)}$ 对于卷积层的所有神经元都是相同的.
  单个卷积核都可以视为在对输入的单个特征做提取.

以一维情况为例:



(a) 全连接层



(b) 卷积层

图 5.5　全连接层和卷积层对比

现考虑一般情况，假设卷积层的结构如下:

- 输入特征映射组: $\mathcal{X} \in \mathbb{R}^{M\times N\times D}$
  其中第 $1 \leq d \leq D$ 个切片矩阵 $X^d \in \mathbb{R}^{M\times N}$ 为一个输入特征映射.
  (图像相关的任务中，彩色图像通常有 RGB 三通道的特征映射，即 $D=3$)
- 输出特征映射组: $\mathcal{Y} \in \mathbb{R}^{M'\times N'\times P}$
  其中第 $1 \leq p \leq P$ 个切片矩阵 $Y^p \in \mathbb{R}^{M'\times N'}$ 为一个输出特征映射.
- 卷积核: $\mathcal{W} \in \mathbb{R}^{U\times V\times P\times D}$
  其中切片矩阵 $W^{p,d} \in \mathbb{R}^{U\times V}$ 为二维卷积核 (其中 $1 \leq p \leq P$, $1 \leq d \leq D$)
- 偏置: $b \in \mathbb{R}^P$
  (因此总参数量是 $U\times V\times P\times D + P$)

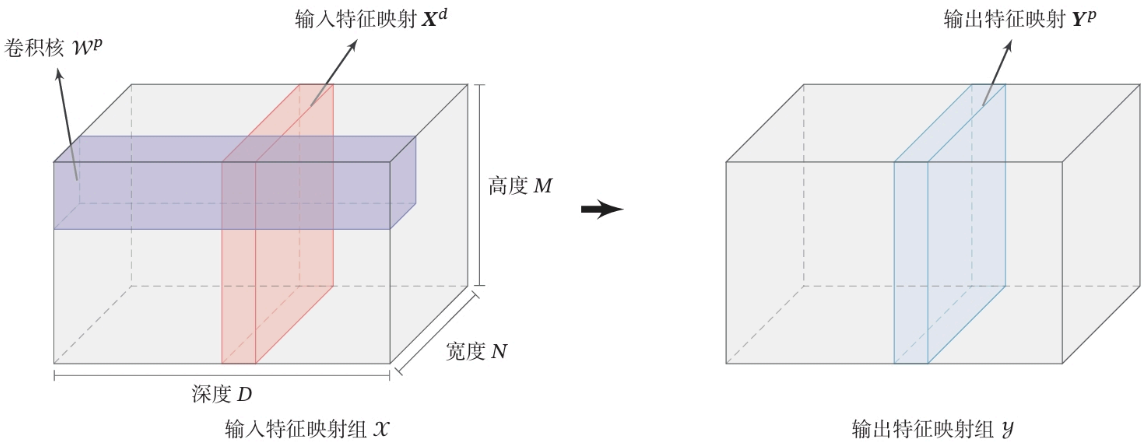


图 5.6　卷积层的三维结构表示

为计算输出特征映射 Y^p，用卷积核 $W^{p,1},\ldots,W^{p,D}$ 分别对输入特征映射 $X^1,\ldots,X^D$ 进行卷积，
然后将卷积结果相加，并加上一个标量偏置 b^p 得到卷积层的净输入 Z^p，
再经过非线性激活函数后得到输出特征映射 Y^p.

$$
\begin{aligned}
Z^p &= W^p \star \mathcal{X} \\
&= \sum_{d=1}^{D} W^{p,d} \star X^d + b^p \cdot \mathrm{ones}(M', N') \\
\hline
Y^p &= f(Z^p)
\end{aligned}
$$

其中 $W^p = [W^{p,1}, \dots, W^{p,D}] \in \mathbb{R}^{U \times V \times D}$ 为三维卷积核，$f(\cdot)$ 为激活函数.
重复上述过程 P 次，就得到 P 个输出特征映射 $Y^1, \dots, Y^P$.

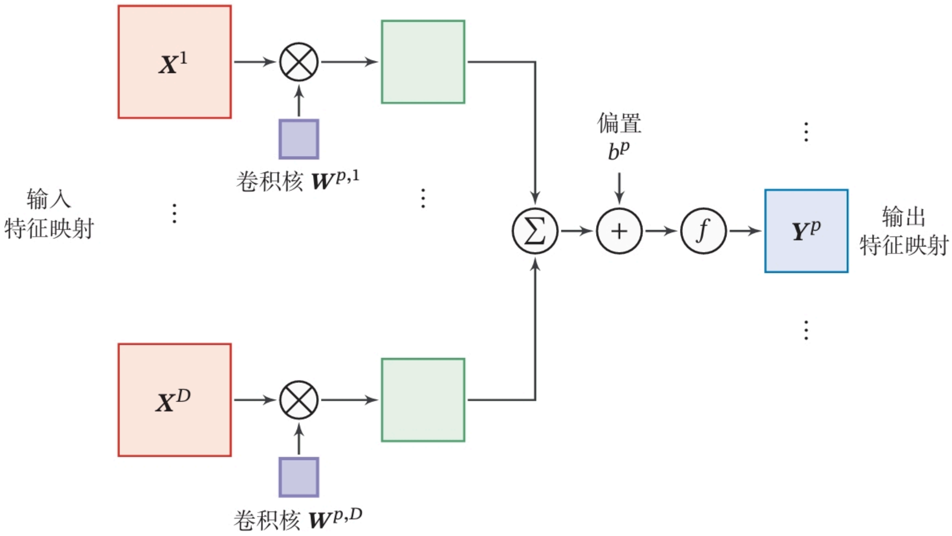


图 5.7　卷积层中从输入特征映射组 $\boldsymbol{X}$ 到输出特征映射 $\boldsymbol{Y}^p$ 的计算示例

3.2.2 汇聚层

汇聚层 (pooling layer) 又称**池化层**，用于特征筛选，减少参数数量.
卷积层虽然可以显著减少网络中连接的数量，但特征映射组中的神经元个数并没有显著减少.
如果后面直接接上全连接层，则全连接层的输入维数依然很高，很容易出现过拟合.
为了解决这个问题，可以在卷积层之后加上一个汇聚层，从而降低特征维数.

假设汇聚层的结构如下:

- 输入特征映射组: $\mathcal{X} \in \mathbb{R}^{M \times N \times D}$ (一般是卷积层的输出特征映射组)
  其中第 $1 \le d \le D$ 个切片矩阵 $X^d \in \mathbb{R}^{M \times N}$ 为一个输入特征映射.
- 输出特征映射组: $\mathcal{Y} \in \mathbb{R}^{M' \times N' \times D}$ (汇聚层不改变特征维数)
  其中第 $1 \le d \le D$ 个切片矩阵 $Y^d \in \mathbb{R}^{M' \times N'}$ 为一个输出特征映射.

对于第 d 个输入特征映射 $X^d \in \mathbb{R}^{M \times N}$，
我们将其划分为 $M' \times N'$ 个区域 $X^d_{m,n}$ (其中 $1 \le m \le M'$，$1 \le n \le N'$)
这些区域可以重叠，也可以不重叠.
汇聚 (pooling) 就是对每个区域进行**下采样** (down sampling) 得到一个值，作为整个区域的概括.
常用的汇聚函数有两种:

- ① 最大汇聚 (max pooling):

$$y^d_{m,n} := \text{maximum value of region } X^d_{m,n}$$

- ② 平均汇聚 (mean pooling):

$$y^d_{m,n} := \text{average value of region } X^d_{m,n}$$

典型的汇聚层是将每个特征映射划分为 2×2 大小的不重叠区域，然后使用最大汇聚的方式进行下采样.
目前主流的卷积网络中，汇聚层仅包含下采样操作，不包含激活函数.

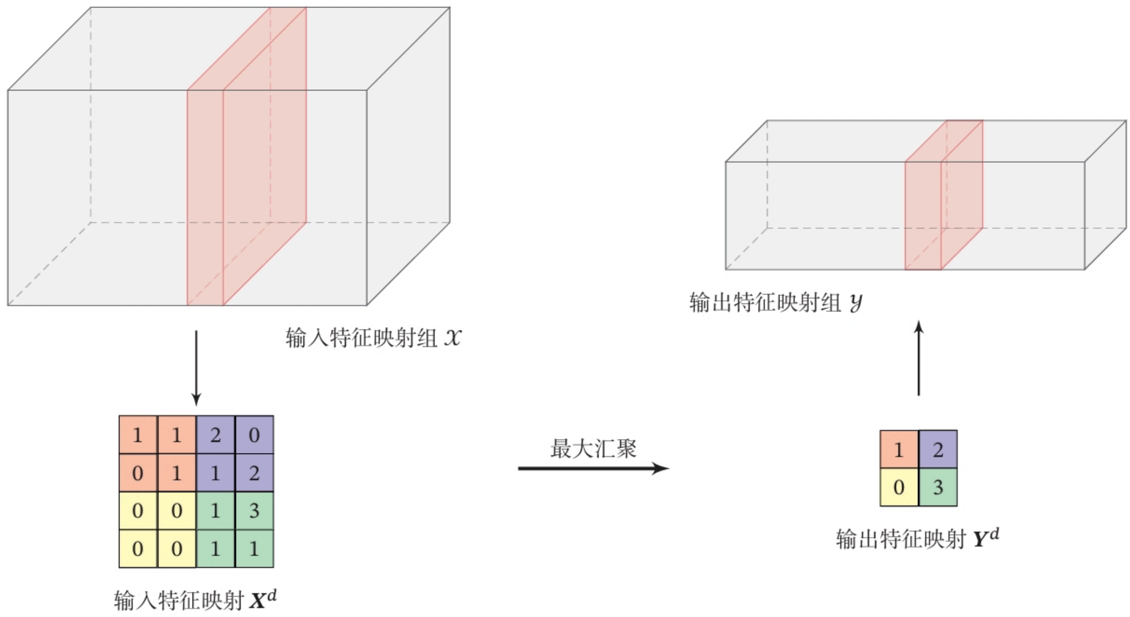


图 5.8 汇聚层中最大汇聚过程示例

汇聚层不但可以有效地减少神经元的数量，
还可以使得网络的神经元拥有更大的感受野，从而对于局部形态的改变具有更强的稳健性.

3.2.3 整体结构

一个典型的卷积网络是由卷积层、汇聚层、全连接层交叉堆叠而成.

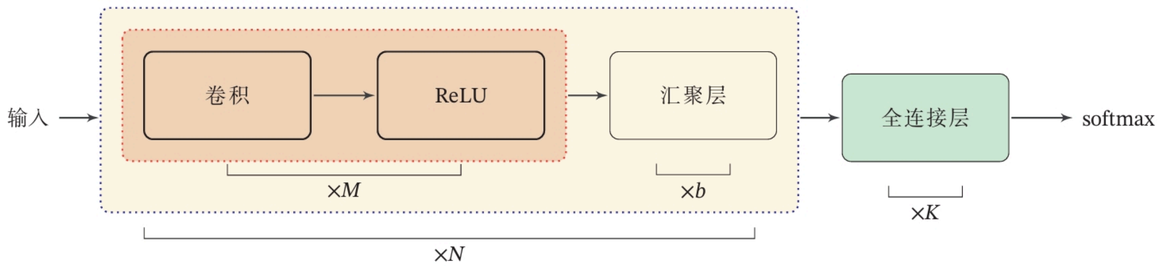


图 5.9 常用的卷积网络整体结构

卷积块为连续 M (通常取 $2\sim5$) 个卷积层和 b (通常取 0 或 1) 个汇聚层.
卷积网络中可以堆叠 N ($1\sim100$ 或更大) 个连续的卷积块，然后接上 K (通常为 $0\sim2$) 个全连接层.
目前卷积网络的整体结构趋向于使用更小的卷积核 (例如 1×1 和 3×3) 以及更深的结构 (例如 50 以上的层数).
此外，由于卷积的操作性越来越灵活 (例如不同的步长)，汇聚层的作用也变得越来越小，
因此在目前主流的卷积网络中，汇聚层的比例正在逐渐降低，趋向于全卷积网络.

3.2.4 反向传播

汇聚层:
设第 $l+1$ 层为汇聚层，来自第 l 层的输入特征映射组为 $\mathcal{X}^{(l)}\in\mathbb{R}^{M\times N\times D}$，
通过汇聚得到第 $l+1$ 层的特征映射净输出 $\mathcal{Z}^{(l+1)}=\mathrm{down}(\mathcal{X}^{(l)})\in\mathbb{R}^{M'\times N'\times P}$:
因为汇聚层是下采样操作 $\mathrm{down}(\cdot)$，所以反向传播时需要对偏导进行对应的上采样操作 $\mathrm{up}(\cdot)$:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial Z^{(l,p)}} &= \frac{\partial X^{(l,p)}}{\partial Z^{(l,p)}} \cdot \frac{\partial Z^{(l+1,p)}}{\partial X^{(l,p)}} \cdot \frac{\partial \mathcal{L}}{\partial Z^{(l+1,p)}} \quad (\text{note that } X^{(l,p)} = f_l(Z^{(l,p)})) \\
&= f_l'(Z^{(l,p)}) \odot \text{up}\left(\frac{\partial \mathcal{L}}{\partial Z^{(l+1,p)}}\right) \quad (1 \le p \le P)
\end{aligned}
$$

其中 $\text{up}(\cdot)$ 操作将 $\frac{\partial \mathcal{L}}{\partial Z^{(l,p)}}$ 上采样为与 $Z^{(l,p)}$ 同形的矩阵，而 $f_l(\cdot)$ 是第 l 层 (假设是卷积层) 的激活函数.

卷积层:
设第 l 层为卷积层，来自第 $l-1$ 层的输入特征映射组为 $\mathcal{X}^{(l-1)} \in \mathbb{R}^{M\times N\times D}$，
通过卷积计算得到第 l 层的特征映射净输出 $\mathcal{Z}^{(l)} \in \mathbb{R}^{M'\times N'\times P}$:

$$
\begin{array}{c}
\mathcal{Z}^{(l)} := [Z^{(l,1)}, \dots, Z^{(l,P)}] \in \mathbb{R}^{M'\times N'\times P} \\
\hline
\begin{aligned}
Z^{(l,p)} &= W^{(l,p)} \star \mathcal{X}^{(l-1)} + b^{(l,p)} \cdot \text{ones}(M', N') \\
&= \sum_{d=1}^{D} W^{(l,p,d)} \star X^{(l-1,p)} + b^{(l,p)} \cdot \text{ones}(M', N') \quad (1 \le p \le P)
\end{aligned} \\
\hline
\mathcal{X}^{(l)} = f_l(\mathcal{Z}^{(l)}) \in \mathbb{R}^{M'\times N'\times P}
\end{array}
$$

其中 $W^{(l,p,d)}$ 为四维张量 $\mathcal{W}^{(l)} \in \mathbb{R}^{U\times V\times P\times D}$ 的第 (p,d) 个卷积核，
而 $b^{(l,p)}$ 为偏置向量 $b^{(l)} \in \mathbb{R}^P$ 的第 p 个分量，$f_l(\cdot)$ 是第 l 层的激活函数.

> **写在前面:**
> 按最优化的记号，关于 z 的梯度应该用 ∇_z 符号表示 (标量对列向量求梯度得到的是同形的列向量)，
> 但为了方便我们使用 $\frac{\partial}{\partial z}$ 来代表梯度，尽管 $\frac{\partial}{\partial z}$ 符号在数学上代表导数 (标量对列向量求导数得到的是转置后同形的行向量)
> 机器学习的记号真是太混乱啦! 而且与数学的记号不兼容.
> 无论如何，现在我们认为 $\frac{\partial}{\partial z}$ 代表梯度.
> (邱锡鹏老师书上的记号也是这样默认的，不过着实让我困惑了一阵)

- 考虑损失函数 $\mathcal{L}$ 关于第 l 层卷积核 $W^{(l,p,d)}$ 的偏导: **(存疑)**

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial W^{(l,p,d)}} &= \frac{\partial Z^{(l,p)}}{\partial W^{(l,p,d)}} \cdot \frac{\partial \mathcal{L}}{\partial Z^{(l,p)}} \\
&= \frac{\partial (W^{(l,p,d)} \star X^{(l-1,p)})}{\partial W^{(l,p,d)}} \cdot \frac{\partial \mathcal{L}}{\partial Z^{(l,p)}} \\
&= \left(\frac{\partial W^{(l,p,d)}}{\partial W^{(l,p,d)}} * X^{(l-1,p)}\right) \cdot \frac{\partial \mathcal{L}}{\partial Z^{(l,p)}} \quad (1 \le d \le D,\ 1 \le p \le P) \\
&= \frac{\partial \mathcal{L}}{\partial Z^{(l,p)}} * X^{(l-1,p)}
\end{aligned}
$$

  其中 $*$ 是互相关运算.

- 考虑损失函数 $\mathcal{L}$ 关于第 l 层偏置向量 $b^{(l)}$ 的偏导: **(存疑)**

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b^{(l,p)}} &= \text{sum}\left(\frac{\partial Z^{(l,p)}}{\partial b^{(l,p)}} \cdot \frac{\partial \mathcal{L}}{\partial Z^{(l,p)}}\right) \\
&= \text{sum}\left(\text{ones}(M', N') \cdot \frac{\partial \mathcal{L}}{\partial Z^{(l,p)}}\right) \quad (1 \le p \le P) \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N}\left[\frac{\partial \mathcal{L}}{\partial Z^{(l,p)}}\right]_{i,j}
\end{aligned}
$$

- 考虑到上一层 (第 $l-1$ 层) 的反向传播: **(存疑)**

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial Z^{(l-1,d)}} &= \frac{\partial X^{(l-1,d)}}{\partial Z^{(l-1,d)}} \cdot \frac{\partial \mathcal{Z}^{(l)}}{\partial X^{(l-1,d)}} \cdot \frac{\partial \mathcal{L}}{\partial \mathcal{Z}^{(l)}} \\
&= f'_{l-1}(Z^{(l-1,d)}) \odot \sum_{p=1}^{P} \left(W^{(l,p,d)} * \frac{\partial \mathcal{L}}{\partial Z^{(l,p)}} \right)
\end{aligned}
$$

其中我们假设第 $l-1$ 层也是卷积层，激活函数为 $f_{l-1}(\cdot)$，而 $*$ 是自相关操作.

The End