介绍

本章主要介绍 2D 卷积的前向传播的卷积计算逻辑，以及反向传播时，计算对权重、偏置、以及对输入的梯度

基本概念

1. 卷积的数学定义
   从数学上讲，连续卷积定义为:
   $$(f * g)(t) = \int_{-\infty}^{+\infty} f(\tau) g(t-\tau) d\tau$$
   对于数字信号和图像处理，我们处理的是离散数据，因此使用离散卷积，其定义为：
   $$(f * g)[n] = \sum_{m=-\infty}^{\infty} f[m] g[n-m]$$
   这个公式可以分解为以下几个步骤来理解：
   i. 翻转：将函数 g 沿着 y 轴翻转，得到 $g[-m]$
   ii. 平移：将翻转后的函数 g 平移 n 个单位，得到 $g[n-m]$
   iii. 相乘：将函数 $f[m]$ 和 平移后的 $g[n-m]$ 对应相乘
   iv. 求和：将所有乘积结果求和，得到在位置 n 处的卷积值
   在 CNN 中，卷积核（或滤波器）扮演函数 g的角色，而输入图像的一部分则扮演函数 f的角色。
2. CNN 中的卷积与互相关
   虽然从数学上讲 CNN 中的操作被称为“卷积”，但它实际上执行的是**互相关** (cross-correlation) 运算。互相关与卷积非常相似，唯一的区别在于它不涉及翻转操作。
   互相关的离散形式定义为：
   $$(f * g)[n] = \sum_{m=-\infty}^{\infty} f[m] g[m+n]$$
   在 CNN 中，由于卷积核的参数是在训练过程中学习所得，卷积核翻转不翻转并不会影响模型的学习能力。翻转操作只是一个数学上的约定，在实践中，为了简化计算，大家通常使用互相关来代替卷积。因此，当在深度学习框架中看到“卷积”，它通常指的是互相关运算。
3. 在 CNN 中的卷积计算公式
   对于输入特征图 X 和卷积核 W，输出特征图 Y 的计算公式：
   $$Y[i,j] = (X * W)[i,j] = \sum_{m=1}^{K_h} \sum_{n=1}^{K_w} X[i \cdot S_h + m, j \cdot S_w + n] \cdot W[m,n] + b$$
   其中：

- (i,j)：输出特征图的坐标
- S_h, S_w：卷积的高度和宽度步长
- K_h, K_w: 卷积核的高度和宽度
- b：偏置项

简单的说，就是将卷积核放在输入特征图的卷积窗口上，将卷积核中的每个数值与输入特征图中对应位置的数值相乘，然后将所有乘积结果相加，得到输出特征图上的某个位置的输出值。
这个乘加过程如同将输入特征图卷积窗口中的数值展平成向量，与展平的卷积核向量内积，得到一个标量值，作为输出特征图的某个元素。最后还要加上一个偏置项，作为输出特征图的最终值。

CNN 中的几个概念

1. 输入
   CNN 卷积层的输入一般是图像的像素数值，未必是原始数值，比如[0~255]，图像数值首先被归一化到 [0,1] 区间，然后还可能需要用指定的均值和方差正则化，然后才能作为卷积层的输入。
   通常，图像数据打包成一个 4D 张量，形状为 $[N, C_{in}, H_{in}, W_{in}]$, 还有其他的形状：$[N, H_{in}, W_{in}, C_{in}]$, 代表不同的数据数据的存储格式，本文中，我们以形状 $[N, C_{in}, H_{in}, W_{in}]$ 为例，这是 pytorch 中常见的存储格式。
   其中:
   N: 批次，尤其在训练时，一次输入多个批次
   C_{in}: 输入特征图数量，也就是图像的通道数，比如灰度图像只有一个通道，而彩色图像有三个通道。
   H_{in}: 图像的高度
   W_{in}: 图像的宽度
2. 卷积核
   卷积核一般是 4D 张量，形状为 $[C_{out}, C_{in}, H_k, W_k]$. 其中：
   C_{out}: 卷积输出通道数
   C_{in}: 卷积核输入通道数
   H_k：卷积核高度
   W_k: 卷积核宽度
   我们可以这么看：$[C_{in}, H_k, W_k]$ 表示一个卷积核，它实际上是一个 3D 张量。那么我们在每个卷积层需要定义有多少个卷积核，C_{out} 就是卷积核的数量，它也是卷积运算之后，输出的特征图的数量。
   其中，C_{in} 需要等同于输入中的通道数 C_{in}，否则无法进行运算。
   另外，卷积核的宽高通常为奇数。
3. 填充
   填充（Padding）是对输入图像特征图的一种操作，具体操作过程就是在图像的上下左右填充一些零，其作用有：
   - 保持输出特征图的尺寸，有时候为了使得输出特征图保持某个尺寸，需要对原始图像进行填充
   - 保留边缘信息，减少边界信息的丢失

   在填充时，我们通常在左右和上下分别保持相同的填充尺寸，也可以上下左右使用同样的尺寸填充。原始图像经过填充之后，输出特征图的尺寸会随之变化。
4. 卷积步长
   卷积步长 (Stride)，表示，每次卷积运算，卷积核在输入特征图上移动多少个像素。它有以下几个作用：
   - 控制输出尺寸，不同的步长会生成不同尺寸的输出特征图
   - 降维和下采样，比如步长越大，输出特征图尺寸越小
   - 减少计算量，步长大，参与计算的像素就少

- 增加感受野，更大的步长意味着输出的每个像素对应输入的更大区域

特别说明

为了方便阐述下面的前向传播和反向传播的原理，我们会忽略批次维度(N)，因为批次维度对计算过程没有影响，多个批次就是多次重复下面的计算结果，并将最终的结果堆叠成一个更高维的张量。实际计算的时候，就是在最外层套一层循环。

但凡需要前向和后向传播的，都有批次维度，卷积核和偏置不包含批次维度。

参数说明

我们规定下文中提到的各种参数如下：

- 输入特征图 (X): 形状为 $[C_{in}, H_{in}, W_{in}]$
- 卷积核 (W): 形状为 $[C_{out}, C_{in}, H_k, W_k]$
- 偏置 (b): 形状为 $[C_{out}]$
- 填充 (Padding): P_h 表示上下填充的像素，P_w 表示左右填充的像素
- 步长 (Stride): S_h，表示上下移动的步长，P_w 表示左右移动的步长
- 输出特征图 (Y): 形状为 $[C_{out}, H_{out}, W_{out}]$

其中，输出特征图的尺寸由如下公式计算所得：

$$H_{out} = floor\left(\frac{H_{in} + 2 \cdot P_h - H_k}{S_h}\right) + 1$$
$$W_{out} = floor\left(\frac{W_{in} + 2 \cdot P_w - W_k}{S_w}\right) + 1$$

floor: 向下取整函数

这里需要注意，输入特征图 (X) 中的 C_{in} 表示有几个通道，比如灰度图像为 1，RGB 为 3，则卷积核 (W) 需要有相同的输入通道 C_{in}。

输出特征图 (Y) 中的 C_{out} 与卷积核 (W) 中的 C_{out} 须一致。

前向传播

我们用实际的形状来说明计算过程：

假设:

- 输入特征图 (X)，形状为 $[3, 3, 3]$
- 卷积核 (W), 形状为 $[2, 3, 3, 3]$
- 偏置 (b), 形状为 $[2]$，这是一维向量

- 填充 (Padding), $P_h = 1$ 表示上下填充, $P_w = 1$ 表示左右填充
- 步长 (Stride), $S_h = 2$ 表示上下移动的步长，$S_w = 2$ 表示左右移动的步长
- 输出 (Y), 形状为可由上述公式计算所得 $[2, 2, 2]$

输入特征图 (X):

$$\begin{bmatrix} x^1_{(1,1)} & x^1_{(1,2)} & x^1_{(1,3)} \\ x^1_{(2,1)} & x^1_{(2,2)} & x^1_{(2,3)} \\ x^1_{(3,1)} & x^1_{(3,2)} & x^1_{(3,3)} \end{bmatrix} \begin{bmatrix} x^2_{(1,1)} & x^2_{(1,2)} & x^2_{(1,3)} \\ x^2_{(2,1)} & x^2_{(2,2)} & x^2_{(2,3)} \\ x^2_{(3,1)} & x^2_{(3,2)} & x^2_{(3,3)} \end{bmatrix} \begin{bmatrix} x^3_{(1,1)} & x^3_{(1,2)} & x^3_{(1,3)} \\ x^3_{(2,1)} & x^3_{(2,2)} & x^3_{(2,3)} \\ x^3_{(3,1)} & x^3_{(3,2)} & x^3_{(3,3)} \end{bmatrix}$$

第一个卷积核 (W^1):

$$\begin{bmatrix} w^{(1,1)}_{(1,1)} & w^{(1,1)}_{(1,2)} & w^{(1,1)}_{(1,3)} \\ w^{(1,1)}_{(2,1)} & w^{(1,1)}_{(2,2)} & w^{(1,1)}_{(2,3)} \\ w^{(1,1)}_{(3,1)} & w^{(1,1)}_{(3,2)} & w^{(1,1)}_{(3,3)} \end{bmatrix} \begin{bmatrix} w^{(1,2)}_{(1,1)} & w^{(1,2)}_{(1,2)} & w^{(1,2)}_{(1,3)} \\ w^{(1,2)}_{(2,1)} & w^{(1,2)}_{(2,2)} & w^{(1,2)}_{(2,3)} \\ w^{(1,2)}_{(3,1)} & w^{(1,2)}_{(3,2)} & w^{(1,2)}_{(3,3)} \end{bmatrix} \begin{bmatrix} w^{(1,3)}_{(1,1)} & w^{(1,3)}_{(1,2)} & w^{(1,3)}_{(1,3)} \\ w^{(1,3)}_{(2,1)} & w^{(1,3)}_{(2,2)} & w^{(1,3)}_{(2,3)} \\ w^{(1,3)}_{(3,1)} & w^{(1,3)}_{(3,2)} & w^{(1,3)}_{(3,3)} \end{bmatrix}$$

第二个卷积核 (W^2):

$$\begin{bmatrix} w^{(2,1)}_{(1,1)} & w^{(2,1)}_{(1,2)} & w^{(2,1)}_{(1,3)} \\ w^{(2,1)}_{(2,1)} & w^{(2,1)}_{(2,2)} & w^{(2,1)}_{(2,3)} \\ w^{(2,1)}_{(3,1)} & w^{(2,1)}_{(3,2)} & w^{(2,1)}_{(3,3)} \end{bmatrix} \begin{bmatrix} w^{(2,2)}_{(1,1)} & w^{(2,2)}_{(1,2)} & w^{(2,2)}_{(1,3)} \\ w^{(2,2)}_{(2,1)} & w^{(2,2)}_{(2,2)} & w^{(2,2)}_{(2,3)} \\ w^{(2,2)}_{(3,1)} & w^{(2,2)}_{(3,2)} & w^{(2,2)}_{(3,3)} \end{bmatrix} \begin{bmatrix} w^{(2,3)}_{(1,1)} & w^{(2,3)}_{(1,2)} & w^{(2,3)}_{(1,3)} \\ w^{(2,3)}_{(2,1)} & w^{(2,3)}_{(2,2)} & w^{(2,3)}_{(2,3)} \\ w^{(2,3)}_{(3,1)} & w^{(2,3)}_{(3,2)} & w^{(2,3)}_{(3,3)} \end{bmatrix}$$

输出特征图 (Y):

$$\begin{bmatrix} y^{(1)}_{(1,1)} & y^{(1)}_{(1,2)} \\ y^{(1)}_{(2,1)} & y^{(1)}_{(2,2)} \end{bmatrix} \begin{bmatrix} y^{(2)}_{(1,1)} & y^{(2)}_{(1,2)} \\ y^{(2)}_{(2,1)} & y^{(2)}_{(2,2)} \end{bmatrix}$$

偏置 (b):

$$[b_1, b_2]$$

计算过程

前面说过，每次卷积过程，相当于展开卷积核与输入特征图卷积窗口中逐元素相乘，然后再累加。所以，这个过程如何向量内积，由此，我们想到可以将计算过程变成矩阵乘法，从而可以利用现有的线性代数的函数库，加速计算过程。

具体而言，就是模拟整个卷积过程，将输入特征图 (X) 重塑为一个矩阵，再将卷积核 (W) 也重塑为一个矩阵，将二者相乘，最后再按照输出特征图的形状对结果进行重塑，即可完成卷积计算过程。

我们分为如下几步：

1. 对输入特征图填充
   $P_h = 1, P_w = 1$，填充之后如下所示：

$$\begin{bmatrix}0&0&0&0&0\\0&x^1_{(1,1)}&x^1_{(1,2)}&x^1_{(1,3)}&0\\0&x^1_{(2,1)}&x^1_{(2,2)}&x^1_{(2,3)}&0\\0&x^1_{(3,1)}&x^1_{(3,2)}&x^1_{(3,3)}&0\\0&0&0&0&0\end{bmatrix}\begin{bmatrix}0&0&0&0&0\\0&x^2_{(1,1)}&x^2_{(1,2)}&x^2_{(1,3)}&0\\0&x^2_{(2,1)}&x^2_{(2,2)}&x^2_{(2,3)}&0\\0&x^2_{(3,1)}&x^2_{(3,2)}&x^2_{(3,3)}&0\\0&0&0&0&0\end{bmatrix}\begin{bmatrix}0&0&0&0&0\\0&x^3_{(1,1)}&x^3_{(1,2)}&x^3_{(1,3)}&0\\0&x^3_{(2,1)}&x^3_{(2,2)}&x^3_{(2,3)}&0\\0&x^3_{(3,1)}&x^3_{(3,2)}&x^3_{(3,3)}&0\\0&0&0&0&0\end{bmatrix}$$

2. 重塑输入特征图
   这个算法称为 **im2col**，在各个深度学习框架中比较常见，就是模拟卷积过程，将输入特征图重塑为二维矩阵: $\hat{X}$ 其中矩阵 $\hat{X}$ 的行列计算方式如下：

$$H_{\hat{x}} = H_{out} \cdot W_{out} = 2 \cdot 2 = 4$$
$$W_{\hat{x}} = C_{in} \cdot H_k \cdot W_k = 3 \cdot 3 \cdot 3 = 27$$

   结合 $S_h = 2$ 和 $S_w = 2$，其重排结果如下：
   鉴于这个矩阵行宽过大，我们将其以转置的形式输出：

$$\begin{bmatrix}
0 & 0 & 0 & x^1_{(2,2)}\\
0 & 0 & x^1_{(2,1)} & x^1_{(2,3)}\\
0 & 0 & x^1_{(2,2)} & 0\\
0 & x^1_{(1,2)} & 0 & x^1_{(3,2)}\\
x^1_{(1,1)} & x^1_{(1,3)} & x^1_{(3,1)} & x^1_{(3,3)}\\
x^1_{(1,2)} & 0 & x^1_{(3,2)} & 0\\
0 & x^1_{(2,2)} & 0 & 0\\
x^1_{(2,1)} & x^1_{(2,3)} & 0 & 0\\
x^1_{(2,2)} & 0 & 0 & 0\\
0 & 0 & 0 & x^2_{(2,2)}\\
0 & 0 & x^2_{(2,1)} & x^2_{(2,3)}\\
0 & 0 & x^2_{(2,2)} & 0\\
0 & x^2_{(1,2)} & 0 & x^2_{(3,2)}\\
x^2_{(1,1)} & x^2_{(1,3)} & x^2_{(3,1)} & x^2_{(3,3)}\\
x^2_{(1,2)} & 0 & x^2_{(3,2)} & 0\\
0 & x^2_{(2,2)} & 0 & 0\\
x^2_{(2,1)} & x^2_{(2,3)} & 0 & 0\\
x^2_{(2,2)} & 0 & 0 & 0\\
0 & 0 & 0 & x^3_{(2,2)}\\
0 & 0 & x^3_{(2,1)} & x^3_{(2,3)}\\
0 & 0 & x^3_{(2,2)} & 0\\
0 & x^3_{(1,2)} & 0 & x^3_{(3,2)}\\
x^3_{(1,1)} & x^3_{(1,3)} & x^3_{(3,1)} & x^3_{(3,3)}\\
x^3_{(1,2)} & 0 & x^3_{(3,2)} & 0\\
0 & x^3_{(2,2)} & 0 & 0\\
x^3_{(2,1)} & x^3_{(2,3)} & 0 & 0\\
x^3_{(2,2)} & 0 & 0 & 0
\end{bmatrix}$$

3. 重塑 (W) 为 $\hat{W}$:

$\hat{W}$ 的形状为:

$$
\begin{aligned}
H_{\hat{w}} &= C_{out} = 4 \\
W_{\hat{w}} &= C_{in} \cdot H_k \cdot W_k = 3 \cdot 3 \cdot 3 = 27
\end{aligned}
$$

本质上，就是将每个卷积核的所有元素依次展平成一个行向量，我们这里有两个卷积核，所以有 2 行，每个卷积核有 3 个通道，所以每行有 27 个元素。

鉴于这个矩阵行宽过大，我们将其以转置的形式输出：

$$
\begin{bmatrix}
w_{(1,1)}^{(1,1)} & w_{(1,1)}^{(2,1)} \\
w_{(1,2)}^{(1,1)} & w_{(1,2)}^{(2,1)} \\
w_{(1,3)}^{(1,1)} & w_{(1,3)}^{(2,1)} \\
w_{(2,1)}^{(1,1)} & w_{(2,1)}^{(2,1)} \\
w_{(2,2)}^{(1,1)} & w_{(2,2)}^{(2,1)} \\
w_{(2,3)}^{(1,1)} & w_{(2,3)}^{(2,1)} \\
w_{(3,1)}^{(1,1)} & w_{(3,1)}^{(2,1)} \\
w_{(3,2)}^{(1,1)} & w_{(3,2)}^{(2,1)} \\
w_{(3,3)}^{(1,1)} & w_{(3,3)}^{(2,1)} \\
w_{(1,1)}^{(1,2)} & w_{(1,1)}^{(2,2)} \\
w_{(1,2)}^{(1,2)} & w_{(1,2)}^{(2,2)} \\
w_{(1,3)}^{(1,2)} & w_{(1,3)}^{(2,2)} \\
w_{(2,1)}^{(1,2)} & w_{(2,1)}^{(2,2)} \\
w_{(2,2)}^{(1,2)} & w_{(2,2)}^{(2,2)} \\
w_{(2,3)}^{(1,2)} & w_{(2,3)}^{(2,2)} \\
w_{(3,1)}^{(1,2)} & w_{(3,1)}^{(2,2)} \\
w_{(3,2)}^{(1,2)} & w_{(3,2)}^{(2,2)} \\
w_{(3,3)}^{(1,2)} & w_{(3,3)}^{(2,2)} \\
w_{(1,1)}^{(1,3)} & w_{(1,1)}^{(2,3)} \\
w_{(1,2)}^{(1,3)} & w_{(1,2)}^{(2,3)} \\
w_{(1,3)}^{(1,3)} & w_{(1,3)}^{(2,3)} \\
w_{(2,1)}^{(1,3)} & w_{(2,1)}^{(2,3)} \\
w_{(2,2)}^{(1,3)} & w_{(2,2)}^{(2,3)} \\
w_{(2,3)}^{(1,3)} & w_{(2,3)}^{(2,3)} \\
w_{(3,1)}^{(1,3)} & w_{(3,1)}^{(2,3)} \\
w_{(3,3)}^{(1,3)} & w_{(3,3)}^{(2,3)} \\
w_{(3,3)}^{(1,3)} & w_{(3,3)}^{(2,3)}
\end{bmatrix}
$$

4. 矩阵乘法：
   公式为：

$$\hat{Y} = \hat{W}\hat{X}^T$$

我们考察输出形状：

- $\hat{W}$ 的形状为 $[2, 27]$
- $\hat{X}$ 的形状为 $[4, 27]$，转置之后 $[27, 4]$
- $\hat{Y}$ 的形状为 $[2, 4]$

5. 加上偏置项

$$\begin{bmatrix} y_{(1,1)}^{(1)} + b_1 & y_{(1,2)}^{(1)} + b_1 & y_{(2,1)}^{(1)} + b_1 & y_{(2,2)}^{(1)} + b_1 \\ y_{(1,1)}^{(2)} + b_2 & y_{(1,2)}^{(2)} + b_2 & y_{(2,1)}^{(2)} + b_2 & y_{(2,2)}^{(2)} + b_2 \end{bmatrix}$$

6. 重塑 $\hat{Y}$ 为 $Y : [2, 2, 2]$, 这与我们开始计算的输出形状一致。

可以手工计算下矩阵乘法的输出结果，应该与按照公式计算方法结果一致。说明这种转化为矩阵乘法的方法是正确的，它只是一种数学技巧而已。

反向传播

反向传播，我们需要计算损失函数对卷积核权重、偏置、以及传播给前一层的梯度。另外，在训练场景下，我们需要缓存前向传播中 **im2col** 的结果，因为，在下面的计算中要用到。

反向传播中，从上游传回来的梯度张量的形状一般是 $[N, C_{out}, H_{out}, W_{out}]$, 这个形状与我们前向传播的输出给下一层的张量形状一致。具体到本例而言，在不考虑批次的情况下，上游梯度 $\frac{\partial L}{\partial Y}$ 的形状为 $[2, 2, 2]$, 与前向传播的最终输出形状一致。

$$\begin{bmatrix} \frac{\partial L}{\partial y_{(1,1)}^{(1)}} & \frac{\partial L}{\partial y_{(1,2)}^{(1)}} \\ \frac{\partial L}{\partial y_{(2,1)}^{(1)}} & \frac{\partial L}{\partial y_{(2,2)}^{(1)}} \end{bmatrix} \begin{bmatrix} \frac{\partial L}{\partial y_{(1,1)}^{(2)}} & \frac{\partial L}{\partial y_{(1,2)}^{(2)}} \\ \frac{\partial L}{\partial y_{(2,1)}^{(2)}} & \frac{\partial L}{\partial y_{(2,2)}^{(2)}} \end{bmatrix}$$

计算 $\frac{\partial L}{\partial b}$

在前向传播中，我们看到偏置是如何应用到输出特征图的，对于某个输出输出通道的特征图，其中每个输出元素都加上了该通道对应的偏置。

在前向传播中，我们可以看到， 输出特征图的数量与偏置项的数量是一致的。

我们计算 $\frac{\partial L}{\partial b_k}$：

$$\frac{\partial L}{\partial b_k} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\partial L}{\partial y_{(i,j)}^{(k)}} \frac{\partial y_{(i,j)}^{(k)}}{\partial b_k}$$

其中：

$$y_{(i,j)}^{(k)} = \hat{y}_{(i,j)}^{(k)} + b_k$$

- $\hat{y}^{(k)}_{(i,j)}$ 是第 k 个输出特征图索引为 (i,j) 元素，是卷积所得
- b_k 是第 k 个输出通道的偏置元素

因此：

$$\frac{\partial y^{(k)}_{(i,j)}}{\partial b_k}=1$$

从而:

$$\frac{\partial L}{\partial b_k}=\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\partial L}{\partial y^{(k)}_{(i,j)}}\cdot 1$$

简单的说，$\frac{\partial L}{\partial b_k}$ 就是对第 k 个通道的输出特征图的梯度逐个元素求和。

计算 $\frac{\partial L}{\partial W}$

就是计算损失函数对卷积核参数的梯度。

在前向传播中，我们通过 **im2col** 重排输入 (X) 和 卷积核 (W)，然后通过矩阵乘法得到输出特征图。其公式为：

$$\hat{Y}=\hat{W}\hat{X}^T$$

根据矩阵乘法的微分法则，我们可以得到：

$$\frac{\partial L}{\partial \hat{W}}=\frac{\partial L}{\partial \hat{Y}}\hat{X}$$

在前向传播时， $\hat{X}$ 的形状为 $[4,27]$, $\hat{W}$ 的形状为 $[2,27]$，$\hat{Y}$ 的形状为 $[2,4]$。

为了利用上述的微分法则，我们需要将 $\frac{\partial L}{\partial Y}$ 重新塑形成 $\hat{Y}$ 的形状，也就是 $[2,4]$。

那么 $\frac{\partial L}{\partial \hat{W}}$ 的形状为 $[2,27]$, 然后重新塑形为 $[2,3,3,3]$, 就得到了 $\frac{\partial L}{\partial W}$

计算 $\frac{\partial L}{\partial X}$

通常使用“转置卷积”的方法来计算 $\frac{\partial L}{\partial X}$.

具体来说，一般有两种常用的方法来计算 $\frac{\partial L}{\partial X}$：

通过前向传播时的矩阵方法计算

前向传播时的矩阵计算公式：

$$\hat{Y}=\hat{W}\hat{X}^T$$

根据矩阵微分法则，我们得到：

$$\frac{\partial L}{\partial \hat{X}} = \left(\frac{\partial L}{\partial \hat{Y}}\right)^T \hat{W}$$

1. 重排 $\frac{\partial L}{\partial Y}$, 使之与 $\frac{\partial L}{\partial \hat{Y}}$ 形状一致。
   这里 $\frac{\partial L}{\partial Y}$ 的形状为 $[2,2,2]$, $\frac{\partial L}{\partial \hat{Y}}$ 的形状为 $[2,4]$。就是就是将梯度特征图从 $[2,2]$ 展开为一个 $[1,4]$ 的行向量。
   展开之后的形式如下:

$$\begin{bmatrix} \frac{\partial L}{\partial y^{(1)}_{(1,1)}} & \frac{\partial L}{\partial y^{(1)}_{(1,2)}} & \frac{\partial L}{\partial y^{(1)}_{(2,1)}} & \frac{\partial L}{\partial y^{(1)}_{(2,2)}} \\ \frac{\partial L}{\partial y^{(2)}_{(1,1)}} & \frac{\partial L}{\partial y^{(2)}_{(1,2)}} & \frac{\partial L}{\partial y^{(2)}_{(2,1)}} & \frac{\partial L}{\partial y^{(2)}_{(2,2)}} \end{bmatrix}$$

2. 进行矩阵乘法
   我们分析输出形状:
   - 经过第一步的重塑，$\frac{\partial L}{\partial \hat{Y}}$ 的形状为 $[2,4]$，转置之后的形状为 $[4,2]$
   - 根据前面的推导，$\hat{W}$ 的形状为 $[2,27]$
   - $\frac{\partial L}{\partial \hat{X}}$ 的形状为 $[4,27]$
3. 将 $\frac{\partial L}{\partial \hat{X}}$ 还原到 $\frac{\partial L}{\partial X}$
   到了这里，虽然 $\frac{\partial L}{\partial \hat{X}}$ 包含了 $\frac{\partial L}{\partial X}$ 的全部信息，但是形状与输入 X 不同，我们需要重塑回其原始的形状 $[3,3,3]$。这个过程称之为 **col2im**，因为 $\hat{X}$ 是输入特征图 X 经过 **im2col** 重塑所得，那么 **col2im** 就是逆运算过程。原始输入特征图中的元素 $x^{(k)}_{(i,j)}$ 可能在 $\hat{X}$ 中出现多次，那么还原的时候，需要将多次的梯度进行累加。
   最后，考虑到我们在前向传播时的填充 $P_h = 1, P_w = 1$，所以，在执行 **col2im** 时需要将填充的部分去掉，最终得到 $\frac{\partial L}{\partial X}$ 的形状为 $[3,3,3]$

通过反卷积的方法计算

为了显示这种计算方法的由来，我们从一个小型的特征图来推导其计算方法。

设:

- 输入特征图的形状为 $[1,3,3]$

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix}$$

- 卷积核形状为 $[1,1,3,3]$

$$\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix}$$

- $P_h = 1, P_w = 1, S_h = 2, S_w = 2$

前向传播

1. 填充

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & x_{11} & x_{12} & x_{13} & 0 \\ 0 & x_{21} & x_{22} & x_{23} & 0 \\ 0 & x_{31} & x_{32} & x_{33} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

2. 输出 Y

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

3. 手工计算 Y：

$$\begin{aligned}
y_{11} &= w_{11} \cdot 0 + w_{12} \cdot 0 + w_{13} \cdot 0 + w_{21} \cdot 0 + w_{22} \cdot x_{11} + w_{23} \cdot x_{12} + w_{31} \cdot 0 + w_{32} \cdot x_{21} + w_{33} \cdot x_{22} \\
y_{12} &= w_{11} \cdot 0 + w_{12} \cdot 0 + w_{13} \cdot 0 + w_{21} \cdot x_{12} + w_{22} \cdot x_{13} + w_{23} \cdot 0 + w_{31} \cdot x_{22} + w_{32} \cdot x_{23} + w_{33} \cdot 0 \\
y_{21} &= w_{11} \cdot 0 + w_{12} \cdot x_{21} + w_{13} \cdot x_{22} + w_{21} \cdot 0 + w_{22} \cdot x_{31} + w_{23} \cdot x_{32} + w_{31} \cdot 0 + w_{32} \cdot 0 + w_{33} \cdot 0 \\
y_{22} &= w_{11} \cdot x_{22} + w_{12} \cdot x_{23} + w_{13} \cdot 0 + w_{21} \cdot x_{32} + w_{22} \cdot x_{33} + w_{23} \cdot 0 + w_{31} \cdot 0 + w_{32} \cdot 0 + w_{33} \cdot 0
\end{aligned}$$

整理之后得到：

$$\begin{aligned}
y_{11} &= w_{22} \cdot x_{11} + w_{23} \cdot x_{12} + w_{32} \cdot x_{21} + w_{33} \cdot x_{22} \\
y_{12} &= w_{21} \cdot x_{12} + w_{22} \cdot x_{13} + w_{31} \cdot x_{22} + w_{32} \cdot x_{23} \\
y_{21} &= w_{12} \cdot x_{21} + w_{13} \cdot x_{22} + w_{22} \cdot x_{31} + w_{23} \cdot x_{32} \\
y_{22} &= w_{11} \cdot x_{22} + w_{12} \cdot x_{23} + w_{21} \cdot x_{32} + w_{22} \cdot x_{33}
\end{aligned}$$

反向传播

设 $\frac{\partial L}{\partial Y}$ 如下所示：

$$\begin{bmatrix} \frac{\partial L}{\partial y_{11}} & \frac{\partial L}{\partial y_{12}} \\ \frac{\partial L}{\partial y_{21}} & \frac{\partial L}{\partial y_{22}} \end{bmatrix}$$

我们计算 $\frac{\partial L}{\partial X}$：

$$\frac{\partial L}{\partial x_{11}} = \frac{\partial L}{\partial y_{11}} \frac{\partial y_{11}}{\partial x_{11}} = \frac{\partial L}{\partial y_{11}} w_{22}$$

$$\frac{\partial L}{\partial x_{12}} = \frac{\partial L}{\partial y_{11}} \frac{\partial y_{11}}{\partial x_{12}} + \frac{\partial L}{\partial y_{12}} \frac{\partial y_{12}}{\partial x_{12}} = \frac{\partial L}{\partial y_{11}} w_{23} + \frac{\partial L}{\partial y_{12}} w_{21}$$

$$\frac{\partial L}{\partial x_{13}} = \frac{\partial L}{\partial y_{12}} \frac{\partial y_{12}}{\partial x_{13}} = \frac{\partial L}{\partial y_{12}} w_{22}$$

$$\frac{\partial L}{\partial x_{21}} = \frac{\partial L}{\partial y_{11}} \frac{\partial y_{11}}{\partial x_{13}} + \frac{\partial L}{\partial y_{21}} \frac{\partial y_{21}}{\partial x_{21}} = \frac{\partial L}{\partial y_{11}} w_{32} + \frac{\partial L}{\partial y_{21}} w_{12}$$

$$\frac{\partial L}{\partial x_{22}} = \frac{\partial L}{\partial y_{11}} \frac{\partial y_{11}}{\partial x_{22}} + \frac{\partial L}{\partial y_{12}} \frac{\partial y_{12}}{\partial x_{22}} + \frac{\partial L}{\partial y_{21}} \frac{\partial y_{21}}{\partial x_{22}} + \frac{\partial L}{\partial y_{22}} \frac{\partial y_{22}}{\partial x_{22}} = \frac{\partial L}{\partial y_{11}} w_{33} + \frac{\partial L}{\partial y_{12}} w_{31} + \frac{\partial L}{\partial y_{21}} w_{13} + \frac{\partial L}{\partial y_{22}} w_{11}$$

$$\frac{\partial L}{\partial x_{23}} = \frac{\partial L}{\partial y_{12}} \frac{\partial y_{12}}{\partial x_{23}} + \frac{\partial L}{\partial y_{22}} \frac{\partial y_{22}}{\partial x_{23}} = \frac{\partial L}{\partial y_{12}} w_{32} + \frac{\partial L}{\partial y_{22}} w_{12}$$

$$\frac{\partial L}{\partial x_{31}} = \frac{\partial L}{\partial y_{21}} \frac{\partial y_{21}}{\partial x_{31}} = \frac{\partial L}{\partial y_{21}} w_{22}$$

$$\frac{\partial L}{\partial x_{32}} = \frac{\partial L}{\partial y_{21}} \frac{\partial y_{21}}{\partial x_{32}} + \frac{\partial L}{\partial y_{22}} \frac{\partial y_{22}}{\partial x_{32}} = \frac{\partial L}{\partial y_{21}} w_{23} + \frac{\partial L}{\partial y_{22}} w_{21}$$

$$\frac{\partial L}{\partial x_{33}} = \frac{\partial L}{\partial y_{22}} \frac{\partial y_{22}}{\partial x_{33}} = \frac{\partial L}{\partial y_{22}} w_{22}$$

初看起来，好像不大容易找到其中的模式，但是，按照如下的调整，我们可以通过卷积的方式得到 $\frac{\partial L}{\partial X}$, 其结果等同于上述的手工计算所得。

1. 将 $\frac{\partial L}{\partial Y}$ 中间增加一个 0 填充的行和列，然后再用 $P_h = 1, P_w = 1$ 的 0 填充四周，我们得到：

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{\partial L}{\partial y_{11}} & 0 & \frac{\partial L}{\partial y_{12}} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{\partial L}{\partial y_{21}} & 0 & \frac{\partial L}{\partial y_{22}} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

2. 再将卷积核翻转 180°，得到：

$$\begin{bmatrix} w_{33} & w_{32} & w_{31} \\ w_{23} & w_{22} & w_{21} \\ w_{13} & w_{12} & w_{11} \end{bmatrix}$$

3. 以第一步所得 $\hat{Y}$ 为输入特征图，第二步所得的 $\hat{W}$ 为卷积核，$S_h = 1, S_w = 1$ 进行卷积运算，最终得到形如 $[3, 3]$ 的 $\frac{\partial L}{\partial X}$，而且其结果与上述手工计算的一致。

因此，我们可以根据上述的简单推导，得到通过反卷积运算 $\frac{\partial L}{\partial X}$ 的梯度。

具体的步骤：

1. 上述在 $\frac{\partial L}{\partial Y}$ 中间插入零的过程称之为**上采样**, 但是上采样和四周零填充不是必须的，需要经过计算所得。
   计算四周填充的公式：

$$E_h = H_k - 1 - P_h$$
$$E_w = W_k - 1 - P_w$$

   - H_k, W_k 是卷积核的尺寸
   - P_h, P_w 是前向传播时填充尺寸

   计算上采样之后的尺寸：

$$U_h = H_{out} + (H_{out} - 1)(S_h - 1)$$
$$U_w = W_{out} + (W_{out} - 1)(S_w - 1)$$

   - H_{out}, W_{out} 是 $\frac{\partial L}{\partial Y}$ 的特征图尺寸
   - S_h, S_w 是前向传播时的卷积步长

   从公式可以看出，当 $S_h = S_w = 1$ 时不需要上采样，当 $E_h = E_w = 0$ 时不需要零填充。
   另外，我们可以将填充和上采样一次完成，最终经过填充和上采样之后的尺寸：

$$F_h = U_h + 2 * E_h$$
$$F_w = U_w + 2 * E_w$$

2. 卷积核翻转 180°
3. 将第一步所得作为输入，用翻转后的卷积核，以步长 $S_h = S_w = 1$ 进行卷积运算。

这里同样要对输入用 **im2col** 展开，然后与卷积核进行矩阵乘法，这里的卷积不需要偏置项。

最终得到的输出应当与前一种方法计算结果一致。