第二章　卷积神经网络

2.1　卷积神经网络初见

2.1.1　动机

卷积神经网络主要面向图像数据处理，针对图像数据的特征表示学习，有哪些数据特征上的先验呢？或者说如何设计一种神经网络模型来更有效地处理图像数据呢？

图像特征描述子的应该具有的关键性质如下：

- 平移不变性。图像空间中感兴趣目标对象发生平移后的不同图片所对应的特征表示空间中也应该呈现出相应的平移规律，或者说一个有效的特征描述子会正确捕获到图像中的感兴趣目标。

- 局部性。图像数据呈现出典型的冗余结构，词袋模型表明了可以用一定数量的视觉字 (visual words) 来表示一大类图像。好的特征描述子应该能够捕获这种冗余的图像局部模式。

- 整体性。图像的高级语义通常需要整合局部的信息进行推理而知，比如，根据多个盲人摸象的不同结果，可以推测出“大象”的身份。

图像特征描述子与视觉模式的关系。　早期特征工程侧重于提取一个更加低层的特征，比如，梯度、边、角特征，并不能称为具有一定语义含义的典型视觉模式，由此，视觉语义内容理解的重任落在了模型设计阶段，从而限制了机器学习中学习能力的发挥。基于深度神经网络模型的表示学习则为特征描述子的自动学习打开了天窗。因此，理解神经网络到底学习到了什么样的特征表示是一个值得思考的问题。或者说如何基于图像数据的上

述特点来合理的设计适应图像数据特征表示学习的神经网络成为关键。能否让神经网络来自动学习出视觉模式呢？

多层感知机的局限性。 思考一下，图像数据处理中的上述先验特性可以通过最具普遍意义的多层感知机(全连接神经网络)来实现吗？

- 全连接的特征变换方式使用了全局的输入信息。
- 主体对象发生位置变化后的两张图像中主体对象位置处的特征表示不一样。

2.1.2 卷积操作的定义

卷积的标准定义。 在数学中，两个(多元)函数（比如 $f, g : \mathbb{R}^b \to \mathbb{R}$ ）之间的“卷积”被定义为

$$(f * g)(\mathbf{x}) = \int f(\mathbf{z}) g(\mathbf{x} - \mathbf{z}) d\mathbf{z}$$

从等式右边的积分表达式中可以看出：$\mathbf{x}$ 可以看做是一个常量，两个函数都可以看成是关于 $\mathbf{z}$ 的函数，$f(\mathbf{z})$ 通常称为卷积核，作为一个权重向量施加到 $g(\mathbf{x} - \mathbf{z})$ 的值域上，但是该值域可以看做是对函数 $g(\mathbf{z})$ 先以原点为中心进行翻转 (旋转 180 度) 再平移 $\mathbf{x}$ 个单位。

卷积运算的定义使得两个函数 f 和 g 具有可交换性，即，

$$(f * g)(\mathbf{x}) = (g * f)(\mathbf{x})$$

卷积运算事实上可以看做是理解傅里叶变换所设定的子操作概念，使得两个函数的傅里叶变换的乘积等于它们卷积后的傅里叶变换。

抛开卷积的标准定义，仅关注其运算形式，可以理解为“用卷积核来加权另一组值”，由此就对应到了如下的“互相关运算”。

深度神经网络中的卷积层—— 互相关运算。 互相关运算的定义为：

$$(f * g)(\mathbf{x}) = \int f(\mathbf{z}) g(\mathbf{x} + \mathbf{z}) d\mathbf{z}$$

可以看出，相比于卷积的标准操作，缺少了函数的翻转。

对于图像而言，通常设定为一个邻域来计算，即

$$[\mathbf{H}]_{i,j} = C_{i,j} + \sum_{a=-\Delta}^{\Delta} \sum_{b=-\Delta}^{\Delta} [\mathbf{V}]_{a,b} [\mathbf{X}]_{i+a,j+b}$$

这就是深度深度网络模型中的卷积操作，其中 $C_{i,j}$ 为偏置量，全连接网络中也存在偏置量。

通道的概念。 在讲卷积的标准定义时，使用的函数形式为多元函数。一元函数可以理解为二维平面上一维数轴上的函数曲线，二元函数可以理解为三维空间中二维平面上堆放了一个的曲面，三元函数可以理解为一个 RGB 彩色图像。但是，对于彩色图像而言，相同空间位置出的 RGB 值称为一个像素的三个值，针对于图像上的卷积运算，更加看重于空间位置对应的特征表示，因此，不能简单地看成是三元函数，而是应该把第三个维度看做是同一空间位置出的不同表现形式，称为通道。因此卷积运算应该是由空间位置指定的多通道上的联合运算，如下图 2.1 所示。

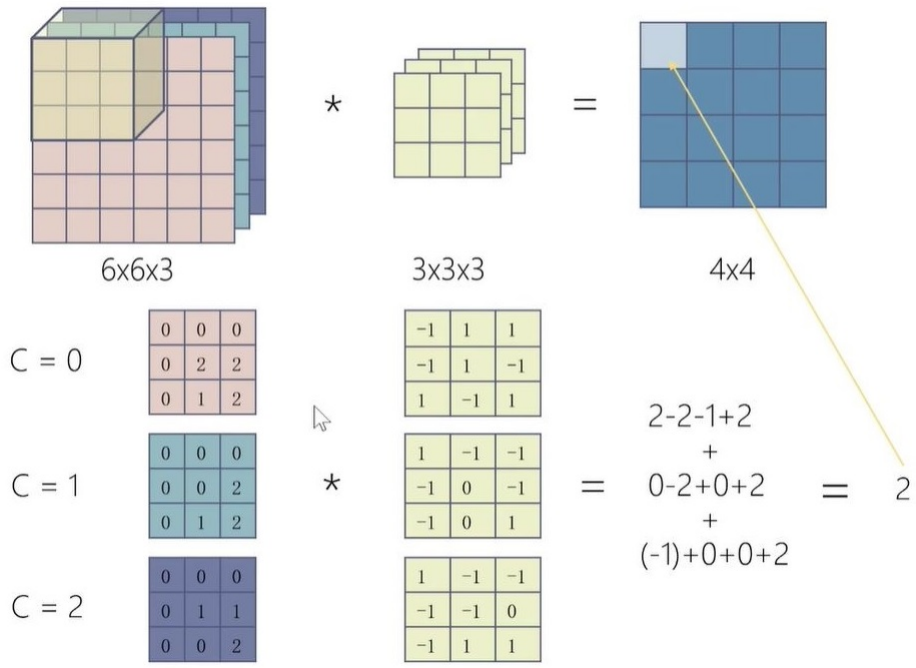

图 2.1: 卷积操作示意图。

卷积操作的可变设置。 比如，边缘填充、步幅。

如何理解 1*1 的卷积核？

2.1.3 汇聚操作 (pooling)

在基于深度神经网络模型的图像特征抽取过程中，希望信息越来越聚集，同时，空间分辨率也越来越小，当然，后续网络层中的神经元所对应的感受野也就越来越大，这样也就越来越接近于终极任务的输出，比如，判定图片中是否含有一只猫。

如何做到空间分辨率越来越低呢？仅依赖卷积操作能实现吗？比如，增大步长。显然是可行的。但是考虑到被大步长跳过的信息也就会忽略。所以，可以融合小步长下的邻域值来提升信息的聚合度，同时可以提升特征表示抽取上的平移不变形 (本来聚焦于局部操作的卷积已经具有平移不变性，这里更加侧重于卷积核大小空间上的平移不变性，在后续大感受野上更加显著。)

常见融合方案：平均汇聚和最大值汇聚。

卷积操作与全连接操作的关系对比。

- 权重稀疏化
- 权重共享：同一通道上的各个神经元的权重相同
- ...

2.1.4 代表性模型—— LeNet

1989 年，由 AT&T 贝尔实验室的研究员 Yann LeCun 提出，并以其名字命名。LeNet 是最早发布的卷积神经网络之一，主要用来解决手写体数字字符的识别问题。其结构如下：

- 2 个卷积层：每个卷积块中的基本单元是一个卷积层、一个 sigmoid 激活函数和平均汇聚层。(现代使用更加有效的 ReLU 和最大汇聚层)。
- 3 个全连接层。使用反向传播算法进行梯度下降优化。

其详细的参数结构图如 2.3 所示：

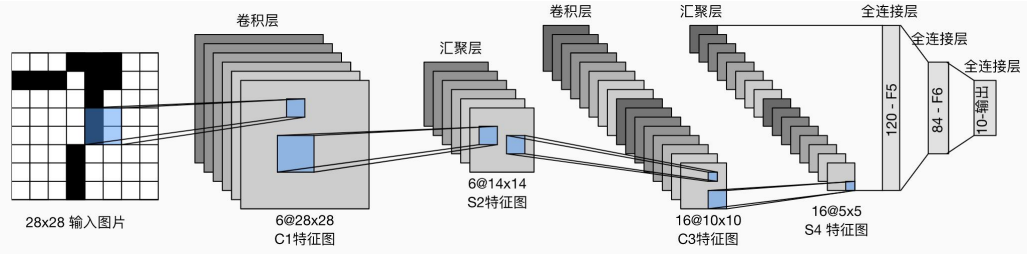

图 2.2: LeNet 的结构示意图。

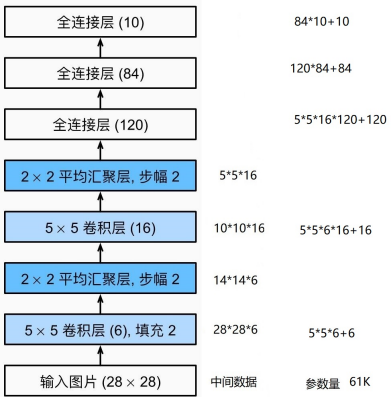

图 2.3: LeNet 的参数结构图。

卷积网络中的常用计算规则

- 卷积/汇聚核后的空间尺寸的计算公式为：$\frac{w-k/2+p}{s} \times \frac{h-k/2+p}{s}$，其中，k为核尺寸，p为padding尺寸，s为步长
- 参数量的计算为：c*k*k*c'+c，其中c为当前卷积层通道数，c'为上一层卷积层通道数，k是核尺寸
- FLOPS 计算主要聚焦于前向过程中的矩阵乘法与加法运算，主要包括卷积层与全连接层：对于卷积层而言，由于共享参数，所以为：参数数量*当前层的空间尺寸；对于全连接层而言，等同于参数数量。

作业：实验代码测试

可参照 d2i.ai ，在 AI stuido 上实现 paddlepaddle 版本，并测试 LeNet

在 Fashion-MNIST 数据集上的性能表现。

2.2 现代卷积神经网络

2.2.1 AlextNet —— 飞跃

类似于 LeNet 这样的浅层小规模神经网络在较小的数据集上取得了与 SVM 相媲美的结果，但是在大型数据集上仍然无法取得优势。原因在于传统模型通常使用了人工精心设计的特征描述子，如，LBP、SIFT、SURF、HOG，但是神经网络模型则处理的是原始的图像像素信息。

因此，势必要赋予神经网络模型更多的使命，比如，进行特征表示学习。这样一来，就需要增加模型容量，进而对算力提出更高的要求，对据量也应该具有一定的规模。

21世纪，已经具备了 ImageNet 这样大规模的图像数据集(2009CVPR)，采用众包方式进行了图片类别标注。另外，算力方面，Nvida 公司把原本用来图形加速的 GPU 做成了面向通用并行计算的 GPGPU (general-purpose GPUs)，并有了配套的编程运算平台库 CUDA。使得卷积神经网络中大量存在的卷积运算和矩阵乘法运算能够借用通用 GPU 快速运算。

2012 年，AlexNet 在 ImageNet 图像分类竞赛上拔得头筹，所取得的图片分类精度远超第二名，首次表明了特征表示学习可以超越人工特征描述子。AlexNet 和 LeNet 的架构非常相似，其结构如下：

- 8 个卷积层：使用 ReLU 激活函数和最大汇聚层。ReLU 具有计算简单，梯度稳定的特点。
- 采用 dropout 控制全连接层的模型复杂度。
- 训练时增加了大量的图像增强数据，如翻转、裁切和变色。增强了模型的健壮性，有效减少过拟合。

从历史发展的角度来总结一下 AlexNet (摘抄自d2l.ai)：

- AlexNet的架构与LeNet相似，但使用了更多的卷积层和更多的参数来拟合大规模的ImageNet数据集。
- 今天，AlexNet已经被更有效的架构所超越，但它是从浅层网络到深层网络的关键一步。