



南京航空航天大學

NANJING UNIVERSITY OF
AERONAUTICS AND ASTRONAUTICS

卷积神经网络 入门基础

李明磊

南京航空航天大学 电子信息工程学院

E-mail: minglei_li@nuaa.edu.cn

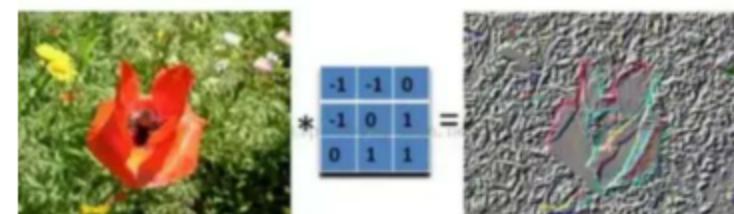
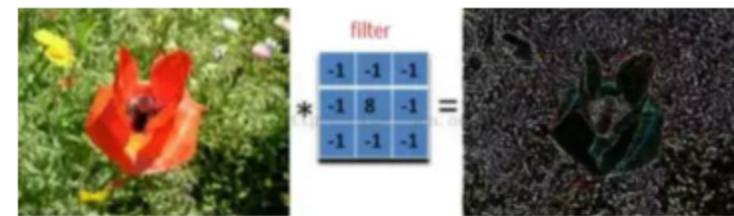
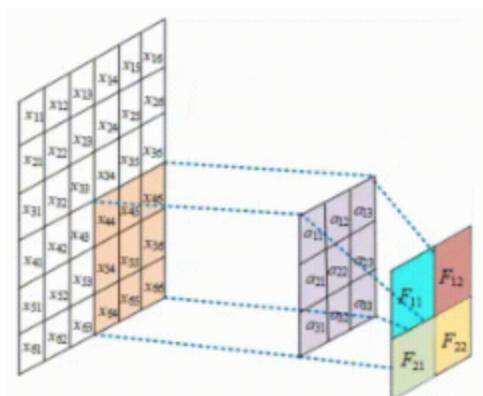
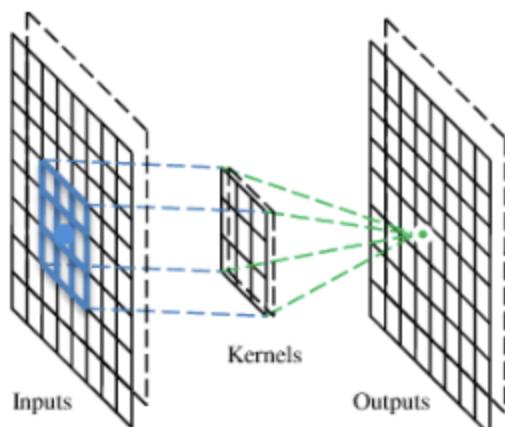
卷积神经网络

卷积神经网络（Convolutional Neural Networks,CNN）是一种深度学习模型或类似于人工神经网络的多层感知器，常用来分析视觉图像。计算机科学家Yann LeCun是在1998年发布的LeNet，第一次喊出了卷积网络(Convolutions network)。



Yann LeCun(1960年-)

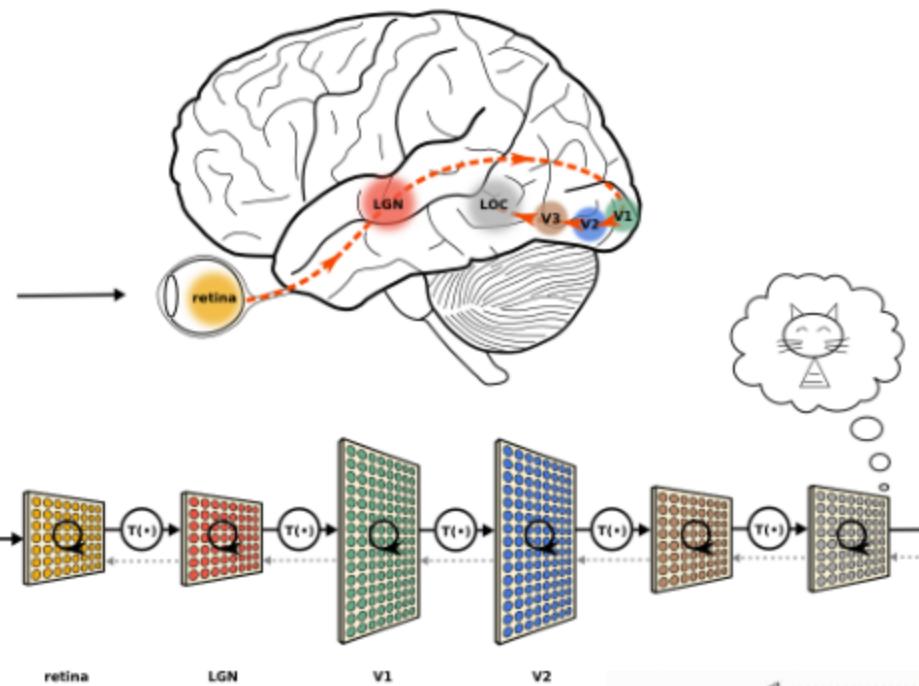
[卷积神经网络可视化_哔哩哔哩_bilibili](#)



卷积动画

卷积神经网络

A CNN is a neural network with some convolutional layers (and some other layers). A convolutional layer has a number of filters that does convolutional operation.



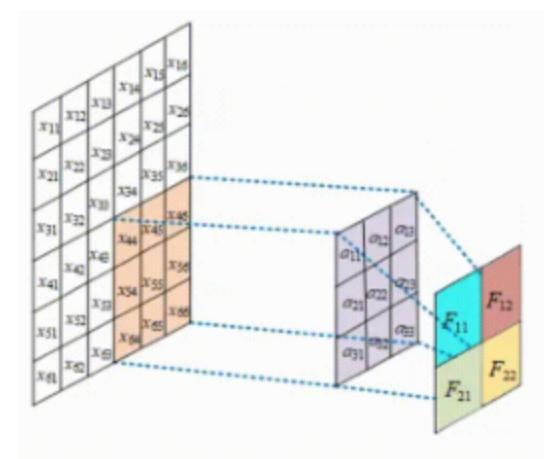
$$\begin{array}{|c|c|} \hline O_{11} & O_{12} \\ \hline O_{21} & O_{22} \\ \hline \end{array} = \text{Convolution} \left(\begin{array}{|c|c|c|} \hline X_{11} & X_{12} & X_{13} \\ \hline X_{21} & X_{22} & X_{23} \\ \hline X_{31} & X_{32} & X_{33} \\ \hline \end{array}, \begin{array}{|c|c|} \hline F_{11} & F_{12} \\ \hline F_{21} & F_{22} \\ \hline \end{array} \right)$$

$$O_{11} = F_{11}X_{11} + F_{12}X_{12} + F_{21}X_{21} + F_{22}X_{22}$$

$$O_{12} = F_{11}X_{12} + F_{12}X_{13} + F_{21}X_{22} + F_{22}X_{23}$$

$$O_{21} = F_{11}X_{21} + F_{12}X_{22} + F_{21}X_{31} + F_{22}X_{32}$$

$$O_{22} = F_{11}X_{22} + F_{12}X_{23} + F_{21}X_{32} + F_{22}X_{33}$$



卷积神经网络和深度学习的历史

卷积神经网络在深度学习的历史中发挥了重要作用。它们是将研究**大脑获得的深刻理解**成功应用于机器学习应用的**关键例子**，也是**第一个表现良好的深度模型之一**。**是第一个解决重要商业应用的神经网络**，并且仍然是当今深度学习应用的前沿。

在20世纪90年代，A T & T 的神经网络研究小组开发了一个用于读取支票的卷积神经网络，到90年代末，NEC部署的这个系统用于读取美国所有支票的10%。后来，微软部署了若干个基于卷积神经网络的**OCR和手写识别系统(MNIST)**。

卷积神经网络也被用来赢得许多比赛。当前对深度学习的商业热潮始于2012年。当时**Alex Krizhevsky**使用新型卷积神经网络(**AlexNet**)赢得了当年的**ImageNet大赛**第一名，TOP-5分类错误率比第二名小约**10%**，引起轰动。

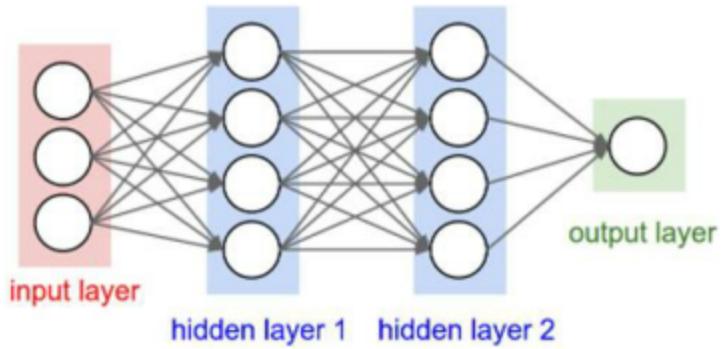
深度学习以及卷积神经网络的适用需要大量的有效训练数据，过去的互联网时代为深度学习提供了大量的**训练数据**，同时随着几十年来**硬件技术**的发展，为利用和计算大量数据提供了条件。所以，近年来，每一次**模型算法**的更新，都取得了良好的效果，为深度学习这把火炬增添了燃料。

卷积神经网络

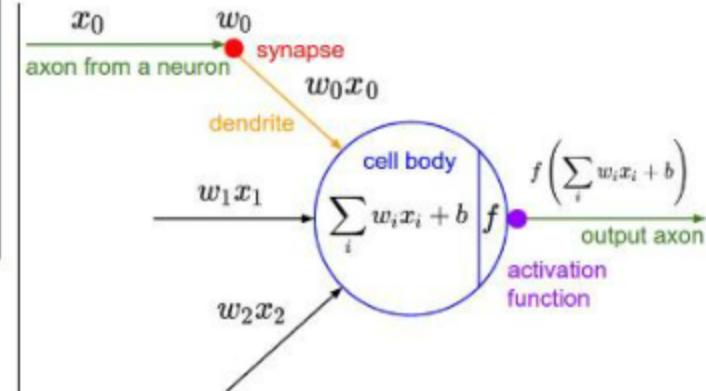
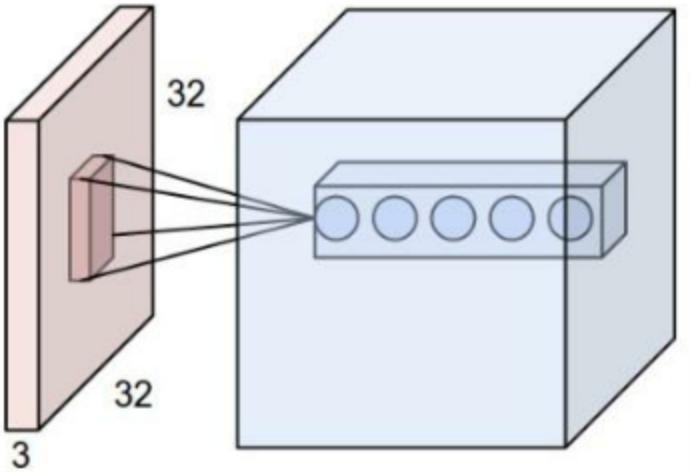
卷积的矩阵转换

感受野和卷积核是卷积运算的一种特殊设定和直观表示，卷积核和感受野之间的卷积运算使用向量矩阵的形式实现，提高了计算效率。

传统神经网络



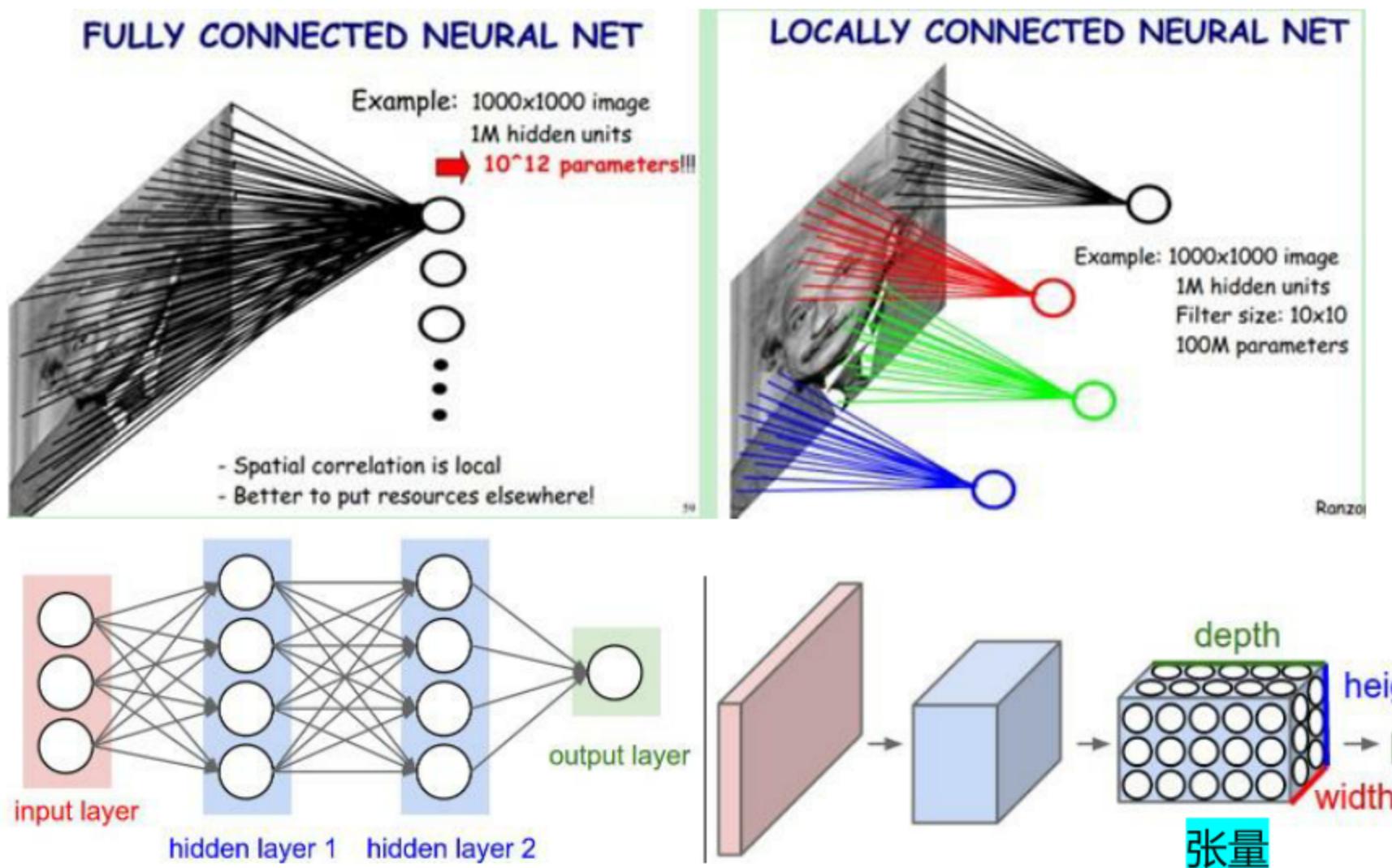
卷积神经网络



卷积神经网络

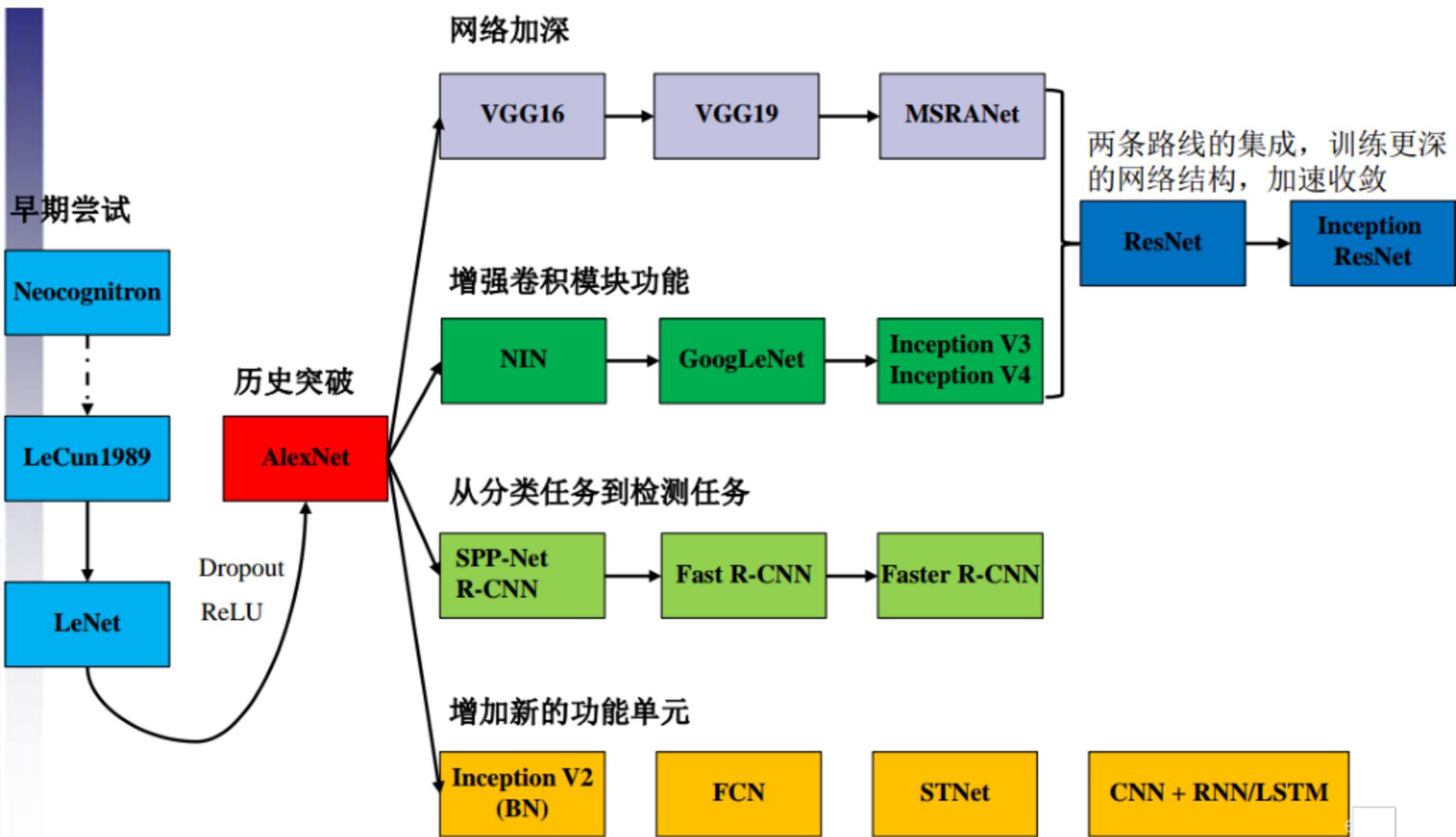
本质：用卷积的方式，提取特征。

卷积神经网络的计算效率提升，参数量： 10^{12} -> 10^6



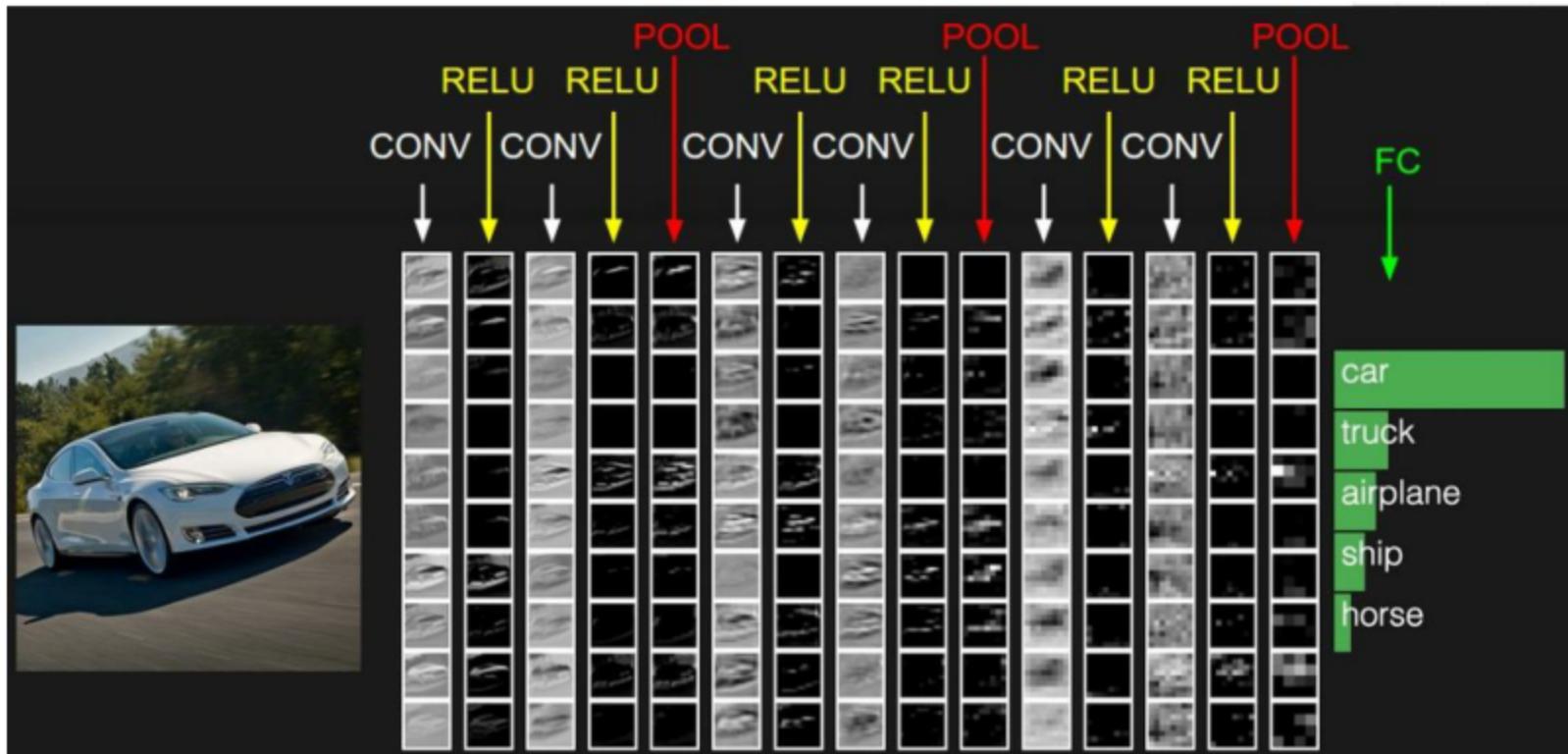
图像的每个像素值都是对客观世界的观测，可以认为是一种离散在像素位置的数据特征

卷积神经网络



卷积神经网络

A simple CNN structure



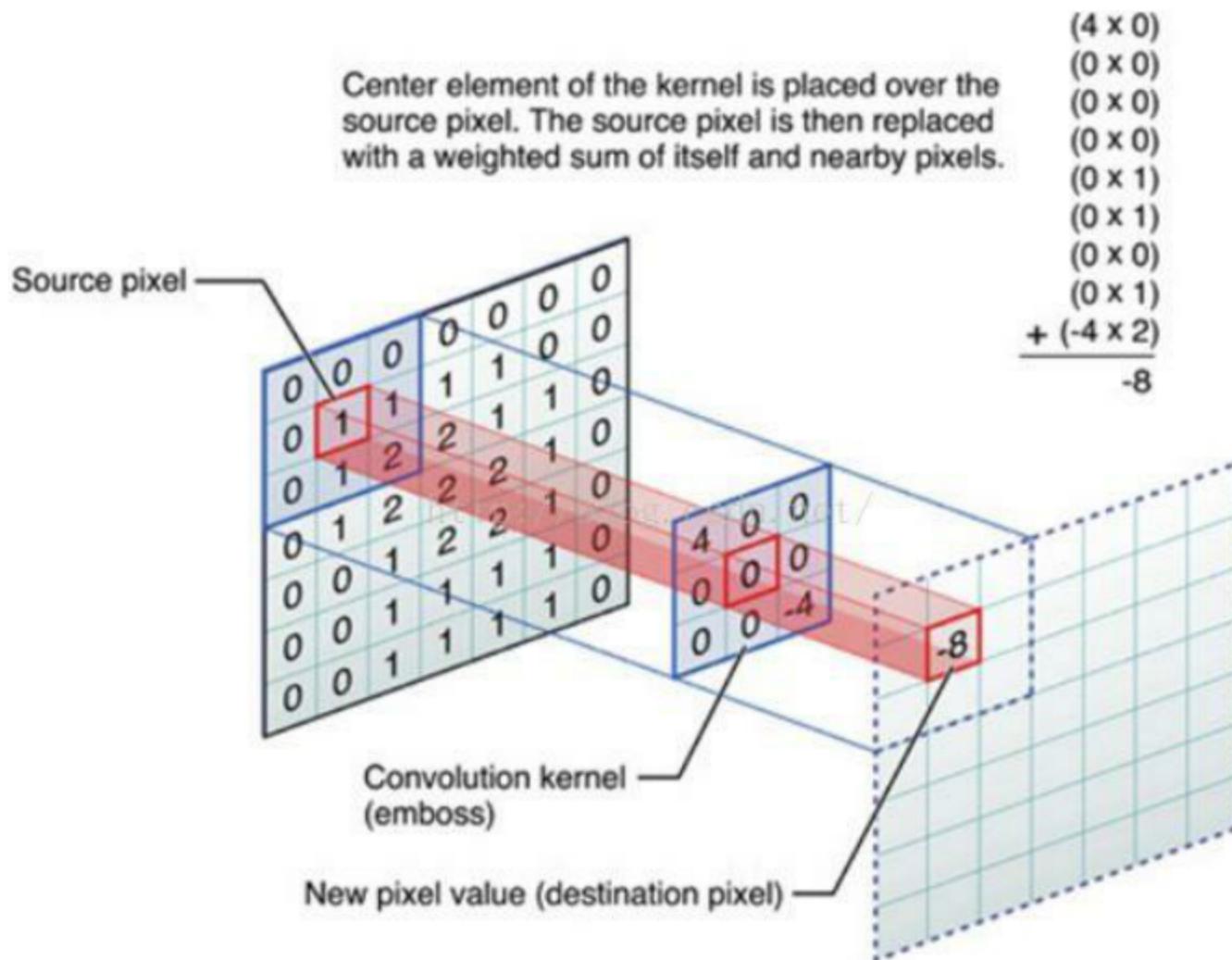
CONV: Convolutional kernel layer
RELU: Activation function
POOL: Dimension reduction layer
FC: Fully connection layer

- 数据输入层 / Input layer
- 卷积计算层 / CONV layer
- ReLU激励层 / ReLU layer
- 池化层 / Pooling layer
- 全连接层 / FC layer

学会数卷积层数

卷积运算层/ CONV layer

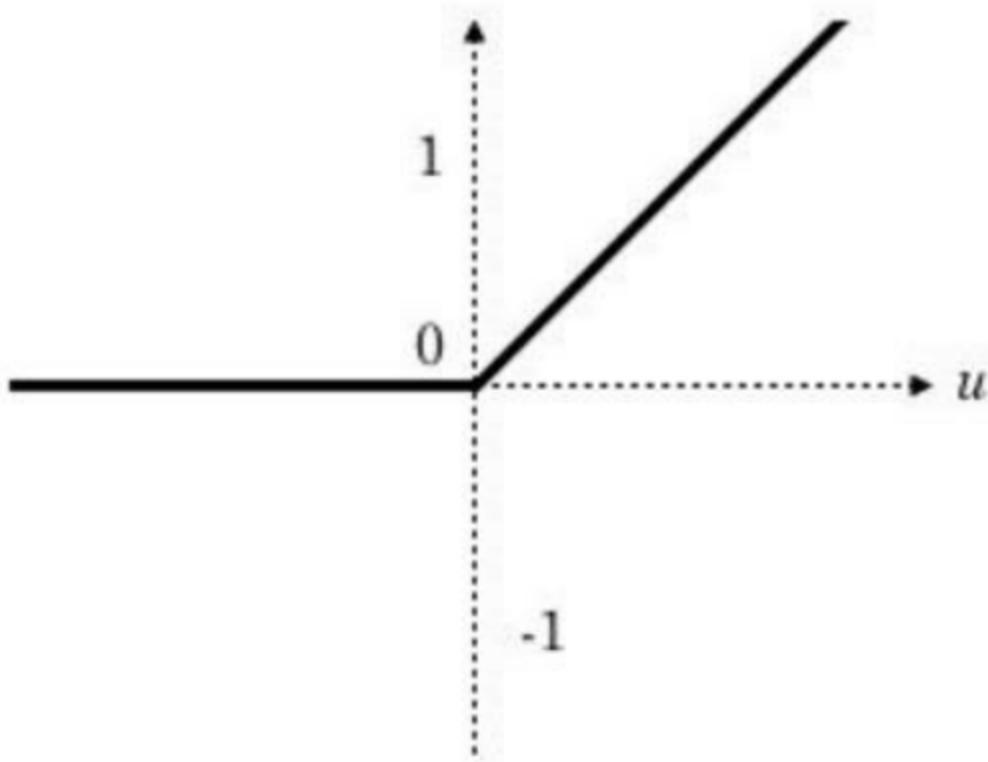
卷积核 Convolutional kernel



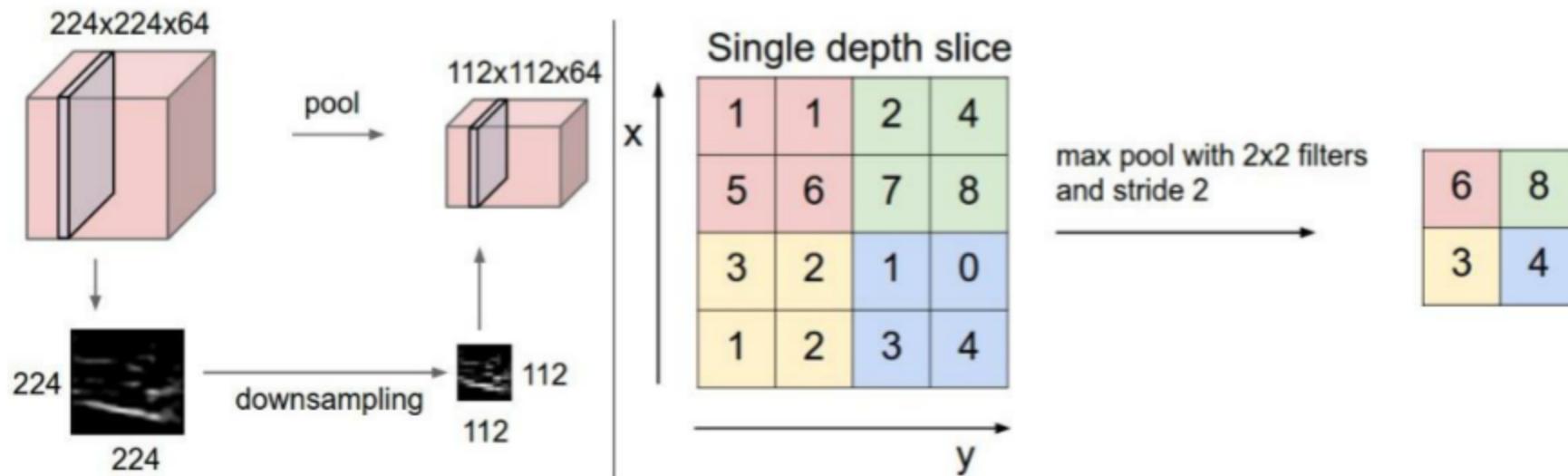
ReLU激励层 / ReLU layer

ReLU (Rectified linear unit)

$$f(x) = \max(0, x)$$



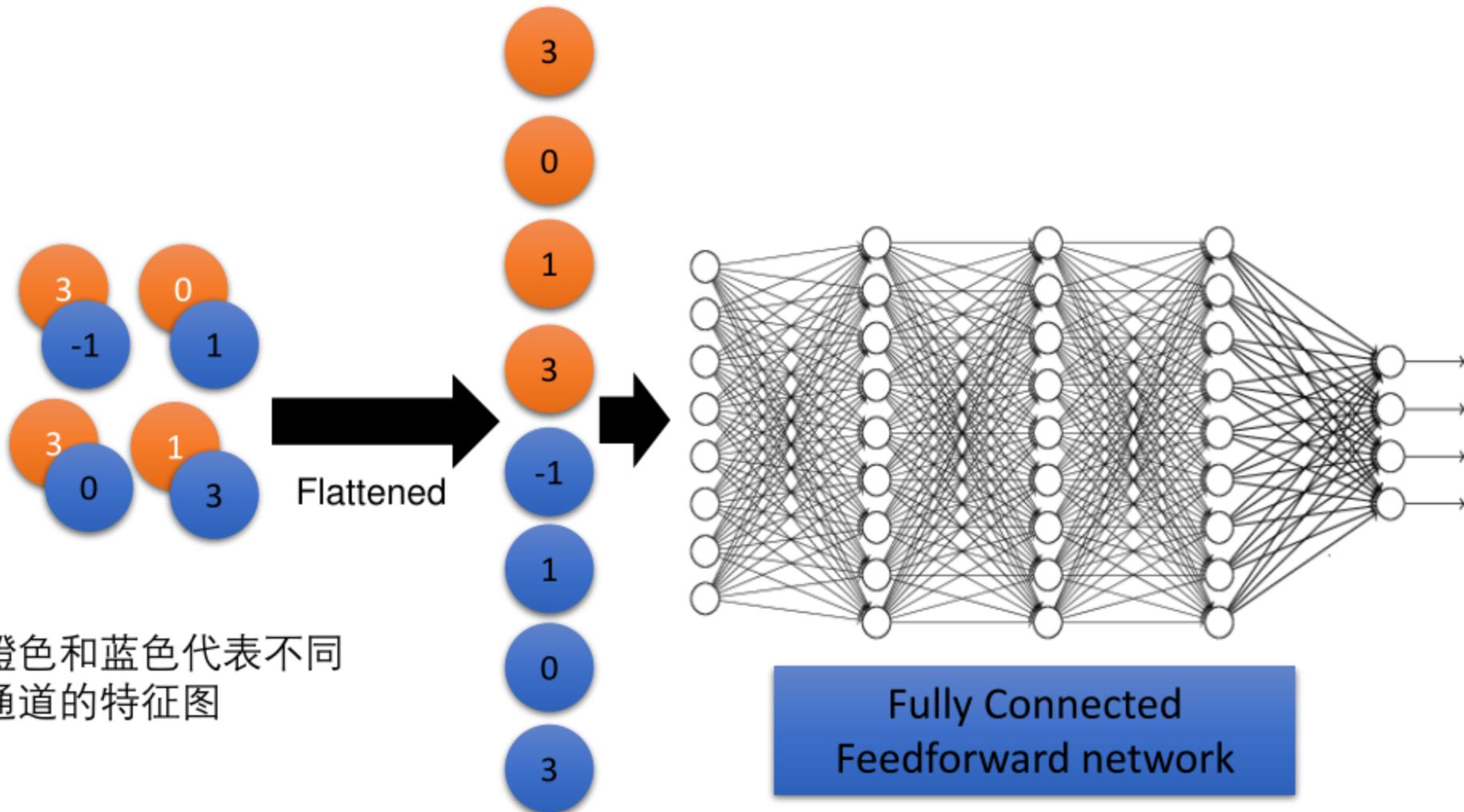
Pooling layer



Pooling layer downsamples the volume spatially, independently in each depth slice of the input volume. **Left:** In this example, the input volume of size $[224 \times 224 \times 64]$ is pooled with filter size 2, stride 2 into output volume of size $[112 \times 112 \times 64]$. Notice that the volume depth is preserved. **Right:** The most common downsampling operation is max, giving rise to **max pooling**, here shown with a stride of 2. That is, each max is taken over 4 numbers (little 2×2 square).

Flattening层

Flatten层用来将输入“压平”，**张量扁平化**，即把多维的输入一维化，常用在从卷积层到**全连接层**的过渡。



Components: loss layer

loss functions used for any (non-CNN) classifier/regressor → type of loss function depends on the kind of task that is learned

- **Softmax loss or multinomial logistic loss:** prediction of a class out of N mutually exclusive classes (e.g., multi-class classifiers)
- **Sigmoid cross-entropy loss:** prediction of N independent probability values in $[0, 1]$
- **Euclidean loss:** for regression, i.e. continuous, real-valued labels

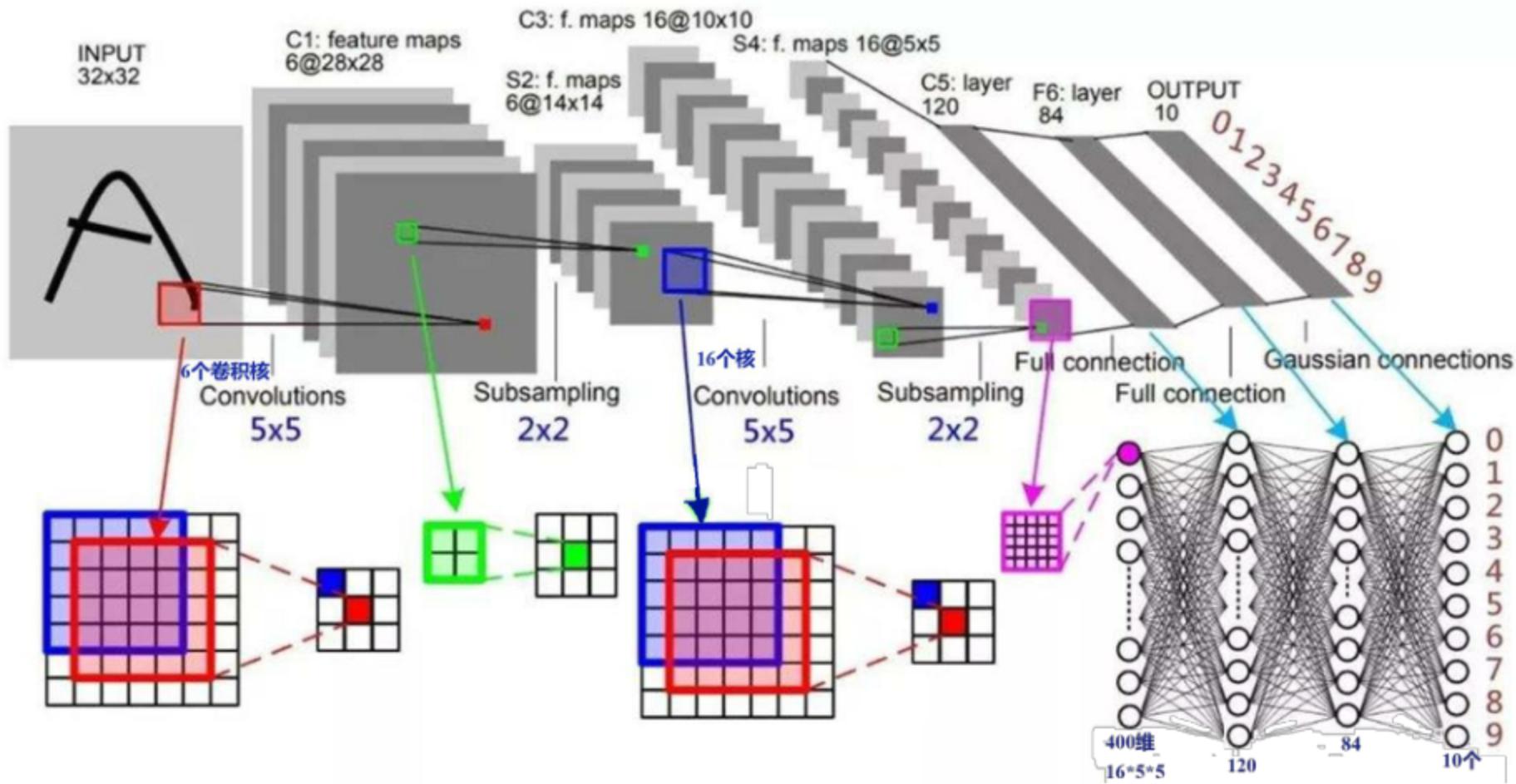
Example: Semantic pixel-wise classification

- Compute (multinomial logistic) loss for each pixel of the input
- Sum all pixel-wise multinomial logistic losses to one image-wide multinomial logistic loss which evaluates the "cost" associated with the entire image.

经典卷积神经网络示例

Architecture of LeNet-5

包括卷积层、池化层、全连接层、损失函数等内容



卷积神经网络架构与常规人工神经网络架构非常相似，特别是在网络的最后几层，即全连接。此外，还注意到卷积神经网络能够接受多个特征图作为输入，而不是向量。

经典卷积神经网络示例

Architecture of LeNet-5

LeNet-5的基本结构包括7层网络结构（不含输入层），其中包括2个卷积层、2个降采样层（池化层）、2个全连接层和1个输出层。

1、输入层（Input layer）

输入层接收大小为 32×32 的手写数字图像，其中包括灰度值（0-255）。在实际应用中，我们通常会对输入图像进行预处理，例如对像素值进行归一化，以加快训练速度和提高模型的准确性。

2、卷积层C1（Convolutional layer C1）

卷积层C1包括6个卷积核，每个卷积核的大小为 5×5 ，步长为1，填充为0。因此，每个卷积核会产生一个大小为 28×28 的特征图（输出通道数为6）。

3、采样层S2（Subsampling layer S2）

采样层S2采用最大池化（max-pooling）操作，每个窗口的大小为 2×2 ，步长为2。因此，每个池化操作会从4个相邻的特征图中选择最大值，产生一个大小为 14×14 的特征图（输出通道数为6）。这样可以减少特征图的大小，提高计算效率，并且对于轻微的位置变化可以保持一定的不变性。

4、卷积层C3（Convolutional layer C3）

卷积层C3包括16个卷积核，每个卷积核的大小为 5×5 ，步长为1，填充为0。因此，每个卷积核会产生一个大小为 10×10 的特征图（输出通道数为16）。

5、采样层S4（Subsampling layer S4）

采样层S4采用最大池化操作，每个窗口的大小为 2×2 ，步长为2。因此，每个池化操作会从4个相邻的特征图中选择最大值，产生一个大小为 5×5 的特征图（输出通道数为16）。

6、全连接层C5（Fully connected layer C5）

C5将每个大小为 5×5 的特征图拉成一个长度为400的向量，并通过一个带有120个神经元的全连接层进行连接。120是由LeNet-5的设计者根据实验得到的最佳值。

7、全连接层F6（Fully connected layer F6）

F6将120个神经元连接到84个神经元。

8、输出层（Output layer）

输出层由10个神经元组成，每个神经元对应0-9中的一个数字，并输出最终的分类结果。在训练过程中，使用交叉熵损失函数计算输出层的误差，并通过反向传播算法更新卷积核和全连接层的权重参数。

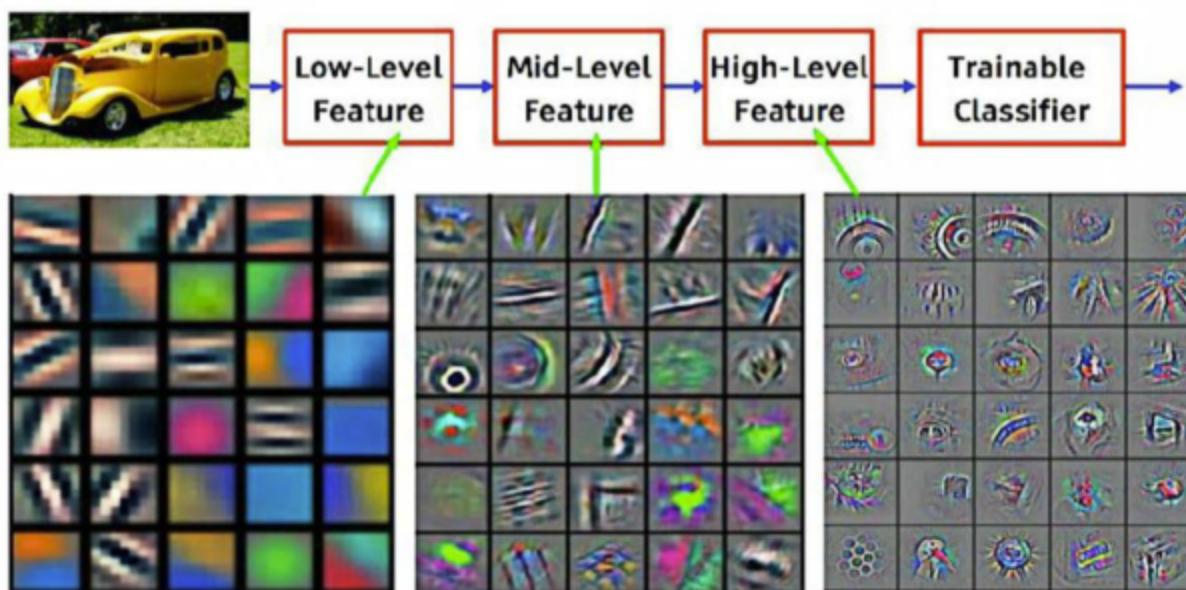
然而，在实际应用中，通常会对LeNet-5进行一些改进，例如增加网络深度、增加卷积核数量、添加正则化等方法，以进一步提高模型的准确性和泛化能力。

卷积网络 卷积网络结构的三个核心思想保证了数据在变换、缩放和扭曲情况下的一致性。基于此思想，提出了LeNet-5的网络结构，包括卷积层、池化层、全连接层、损失函数等内容。

卷积神经网络

Components: CONV example

[From recent Yann LeCun slides]



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Low-level filters (features): learn primitive patterns, usually blobs and gradients

High-level filters (features): capture object parts (e.g., wheels, faces) and entire objects

Components: learned filters!

All filters that convolve the image are learned discriminatively (via stochastic gradient descent and backpropagation) such that they adapt to the specific data and specific task!

Krizhevsky et al. 2012



Backpropagation:

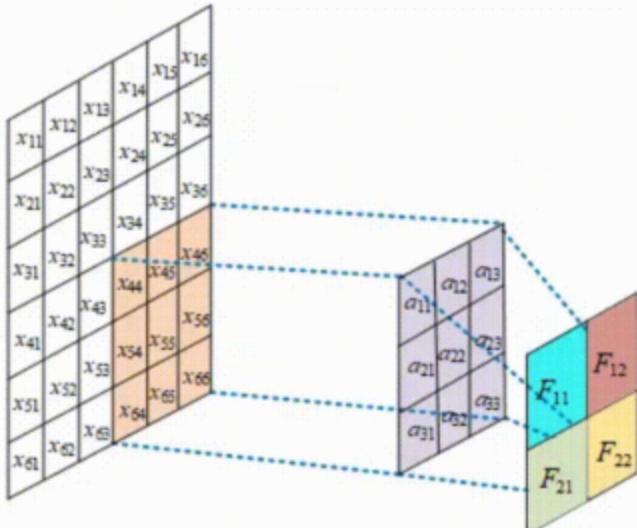
- For learning filters (given class labels) we use backpropagation.
- A backward pass for a convolution operation (for both the data and the weights) is also a convolution but with spatially-flipped filters.

This is the main power of deep learning!

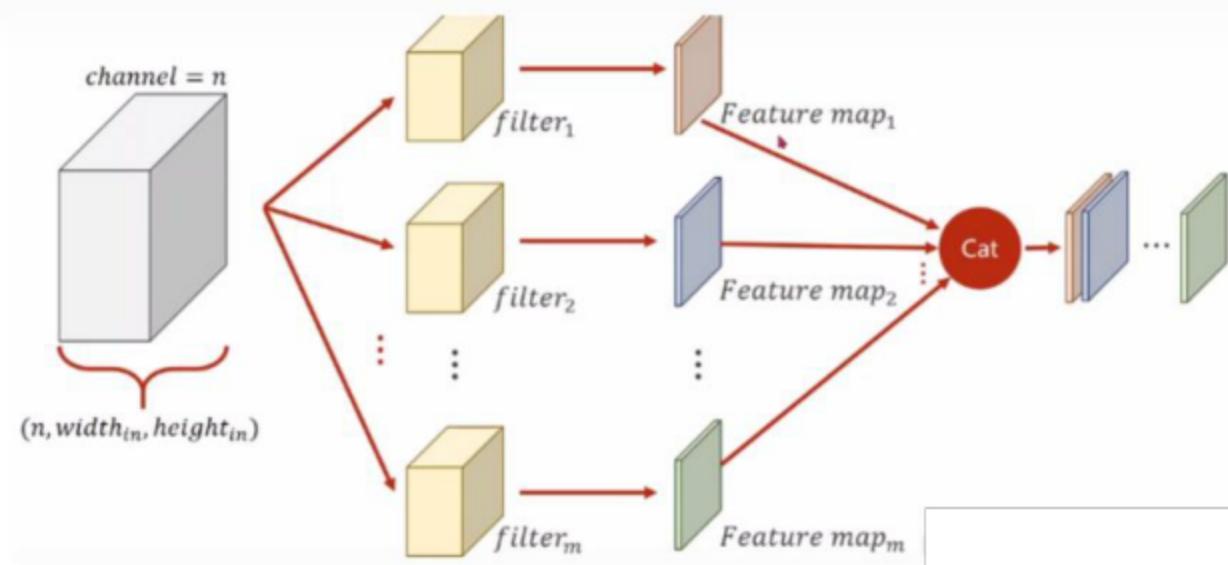
卷积神经网络

卷积神经网络之所以计算效率高，对特征提取的效果好
卷积神经网络具有3个特性：权值共享，多卷积核，池化。

1、权值共享



2、多通道多卷积核



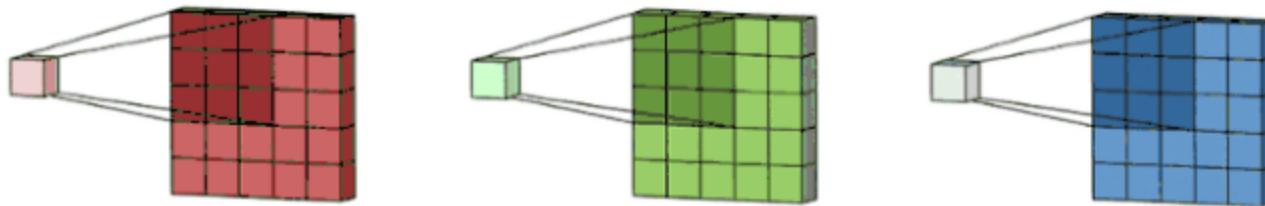
所谓的权值共享就是说，给一张输入图片，用一个filter去扫这张图，filter里面的数就叫权重^Q，这张图每个位置是被同样的filter扫的，所以权重是一样的，也就是共享。

多通道多卷积核：每一个卷积核的通道数量，要求与输入的通道数量一致。因为卷积运算要对每一个通道的像素值要进行处理，所以每一个卷积核的通道数量必须要与输入通道数量保持一致

卷积神经网络

2、 CNN多通道和多卷积核

动图



每个卷积核都应用到输入层的 3 个通道，执行 3 次卷积后得到了尺寸为 3×3 的 3 个通道。

动图

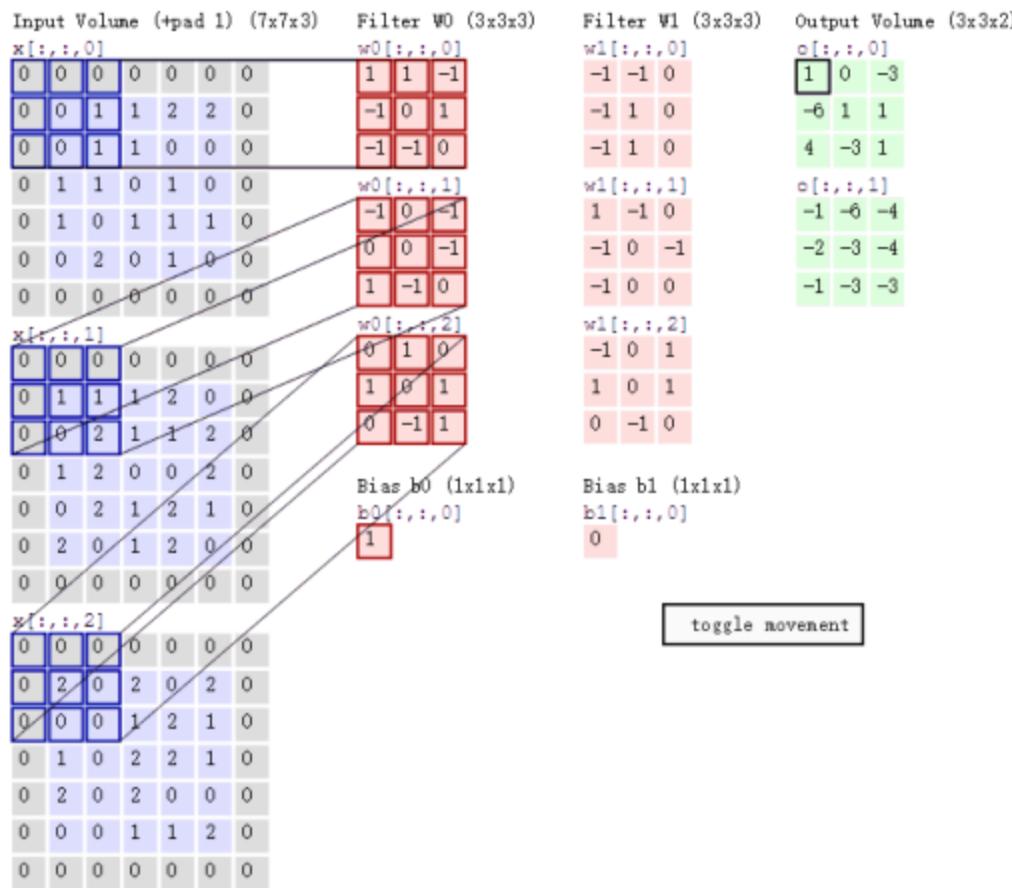


3 个通道都合并到一起（元素级别的加法）组成了一个大小为 $3 \times 3 \times 1$ 的单通道。这个通道是输入层 ($5 \times 5 \times 3$ 矩阵) 使用了过滤器 ($3 \times 3 \times 3$ 矩阵) 后得到的结果。

卷积神经网络

卷积神经网络之所以计算效率高，对特征提取的效果好
卷积神经网络具有3个特性：权值共享，多卷积核，池化。

2、 CNN多通道和多卷积核

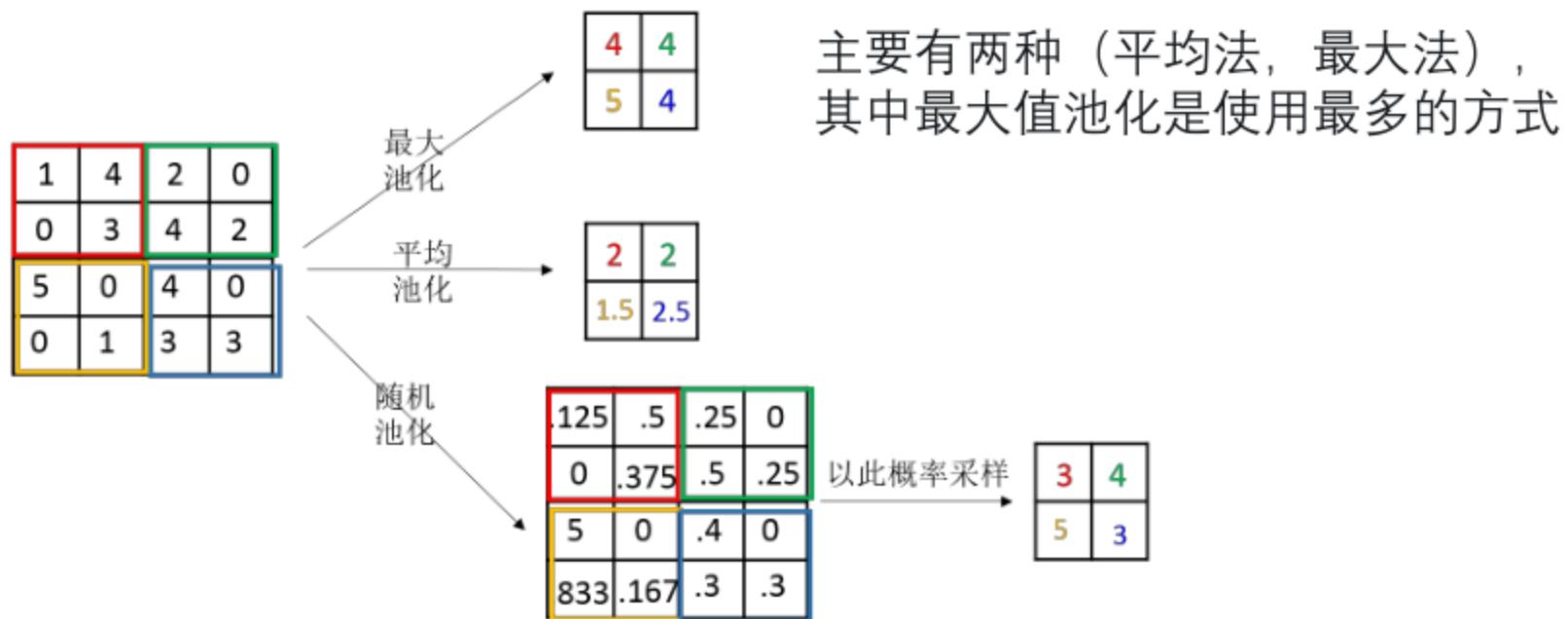


卷积神经网络

卷积神经网络之所以计算效率高，对特征提取的效果好
卷积神经网络具有3个特性：权值共享，多卷积核，池化。

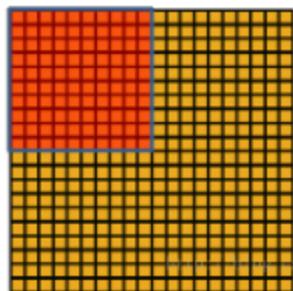
3、池化

- 图像深度学习中，图像尺寸过大，引入池化以减少参数矩阵的尺寸，从而减少最后全连层中的参数数量——根本目的为了防止过拟合
- 在图像识别领域，池化还能提供平移和旋转不变性。若对某个区域做了池化，即使图像平移/旋转几个像素，得到的输出值也基本一样，因为每次最大值运算得到的结果总是一样的。
- 常出现的场合：卷积层后一般会跟上一个池化层
- 实际上是一种降采样的方式

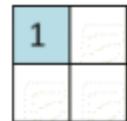


CNN 特性 – 池化

卷积神经网络池化有最大池化(max_pool)和平均池化(avg_pool)，顾名思义，最大池化取区域内最大值，平均池化取区域内平均值。其它池化包括L₂范数以及依靠据中心像素距离的加权平均池化。



Convolved
feature



Pooled
feature

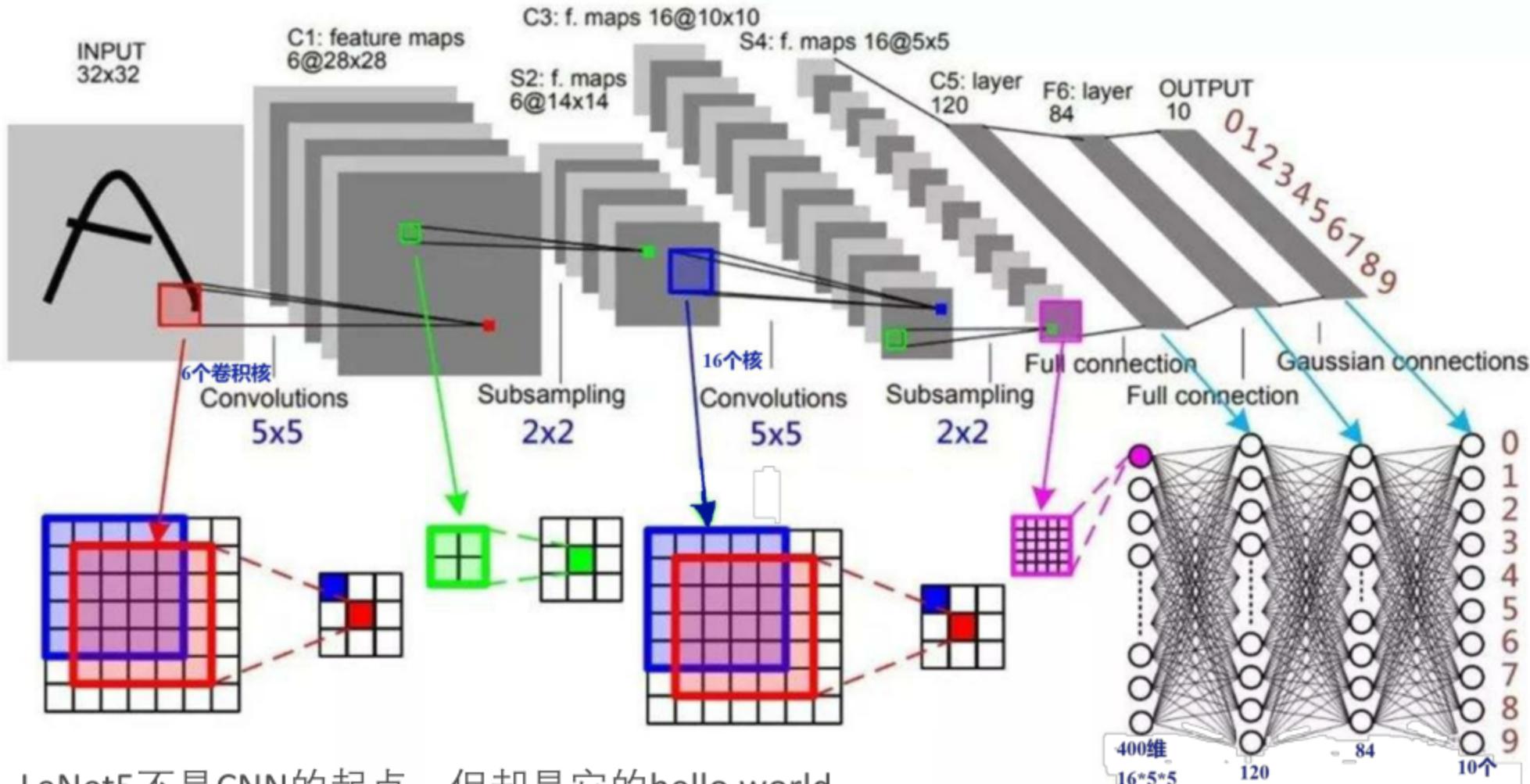
CNN池化过程

为什么要池化？

1. 减少参数的量，**提高计算效率**。
2. 最大池化能显著增强局部特征，平均池化可**减少噪声**。
(最大池化提取轮廓特征，平均池化可模糊图像)
3. **提高局部平移不变性**

经典的CNN模型

Architecture of LeNet (LeCun 1989年、1998年)



LeNet5不是CNN的起点，但却是它的hello world
让大家看到了卷积神经网络商用的前景。

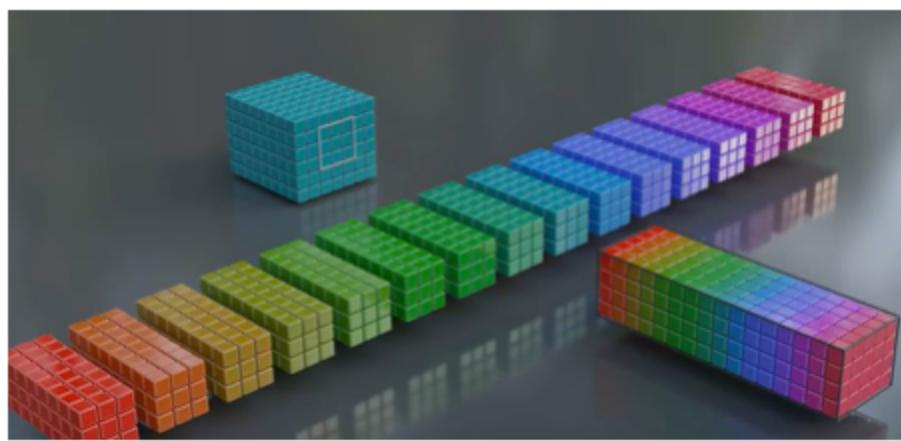
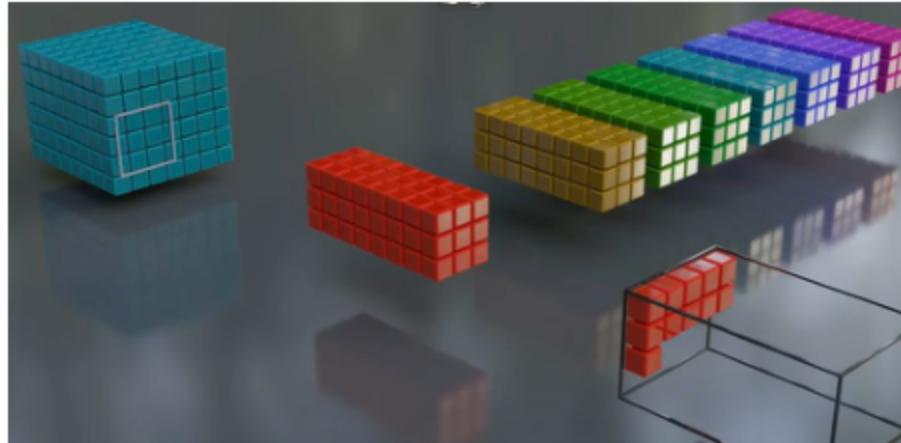
Quiz: 有几次卷积?

一共2层 (2个卷积层+2个全连接层+1个softmax层)

回头看看什么是卷积

所有的卷积神经网络动画都是错的！除了这个动画_哔哩哔哩_bilibili

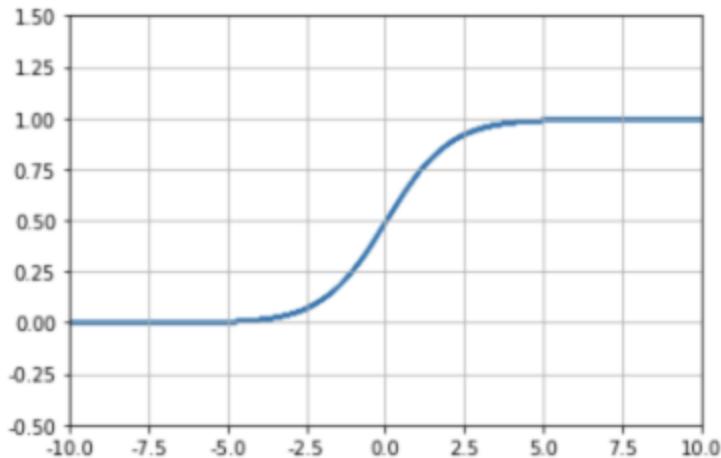
卷积神经网络中的滤波器要选择多少个呢？！3D动画_哔哩哔哩_bilibili



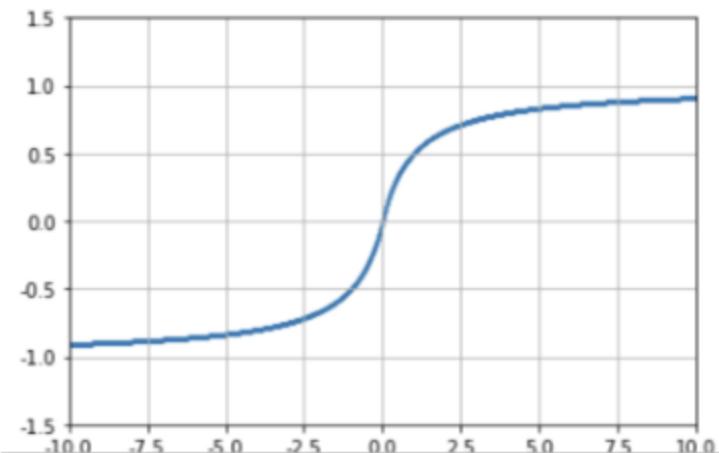
卷积神经网络中的反向传播动画演示_哔哩哔哩_bilibili

激活函数

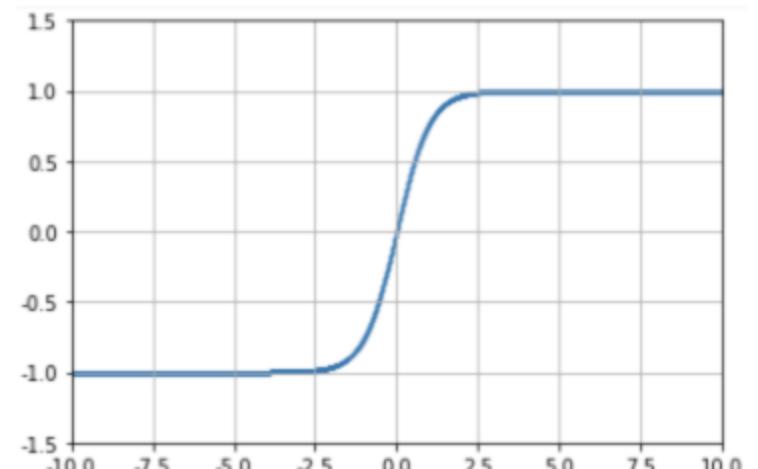
$$\text{Sigmoid}(x) = \frac{1}{1 + \exp(-x)}$$



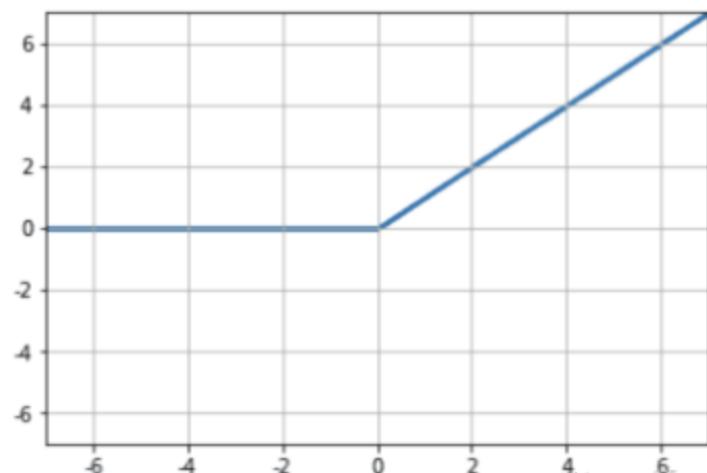
$$\text{SoftSign}(x) = \frac{x}{1 + |x|}$$



$$\text{Tanh}(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

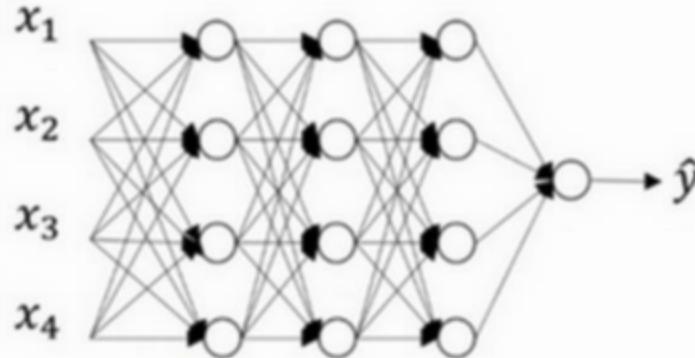


$$\text{ReLU}(x) = \max(0, x)$$



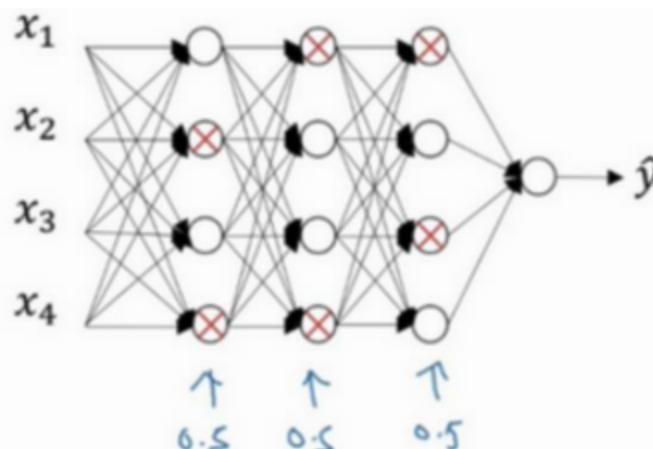
Dropout

Dropout属于一种在神经网络下的正则化，对于一个正常的神经网络如下图所示：



假如针对说这个神经网络存在过拟合，dropout通过在训练的过程中随机丢掉部分神经元来减小神经网络的规模从而防止过拟合。

随机丢掉部分神经元，这里设置一个概率P，它表示针对网络中每一层消除的神经网络节点的概率。如下图所示红色X的神经元表示已经丢弃的，然后神经网络也删除一部分神经元连接的线。



MNIST dataset

The MNIST database of handwritten digits, available from this page, has a training set of 60,000 examples, and a test set of 10,000 examples.

It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image.

CIFAR10 dataset

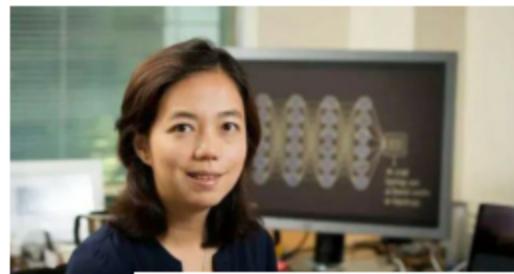
60000 32x32 color images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.

ImageNet: 一个大型的图像分类数据库

ImageNet数据集为深度学习的兴起做了相当大的贡献。该数据集是使用Amazon Mechanical Turk将分类任务外包给工人来构建的，这也使得这个天文级别的数据集成为可能。

ImageNet大型视觉识别挑战赛(ILSVRC, ImageNet Large Scale Visual Recognition Challenge)是以ImageNet数据库为对象的图像分类算法竞赛，同时它也推动了计算机视觉领域其他许多创新的发展。

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9



airplane



bird



cat



deer



dog



frog



horse



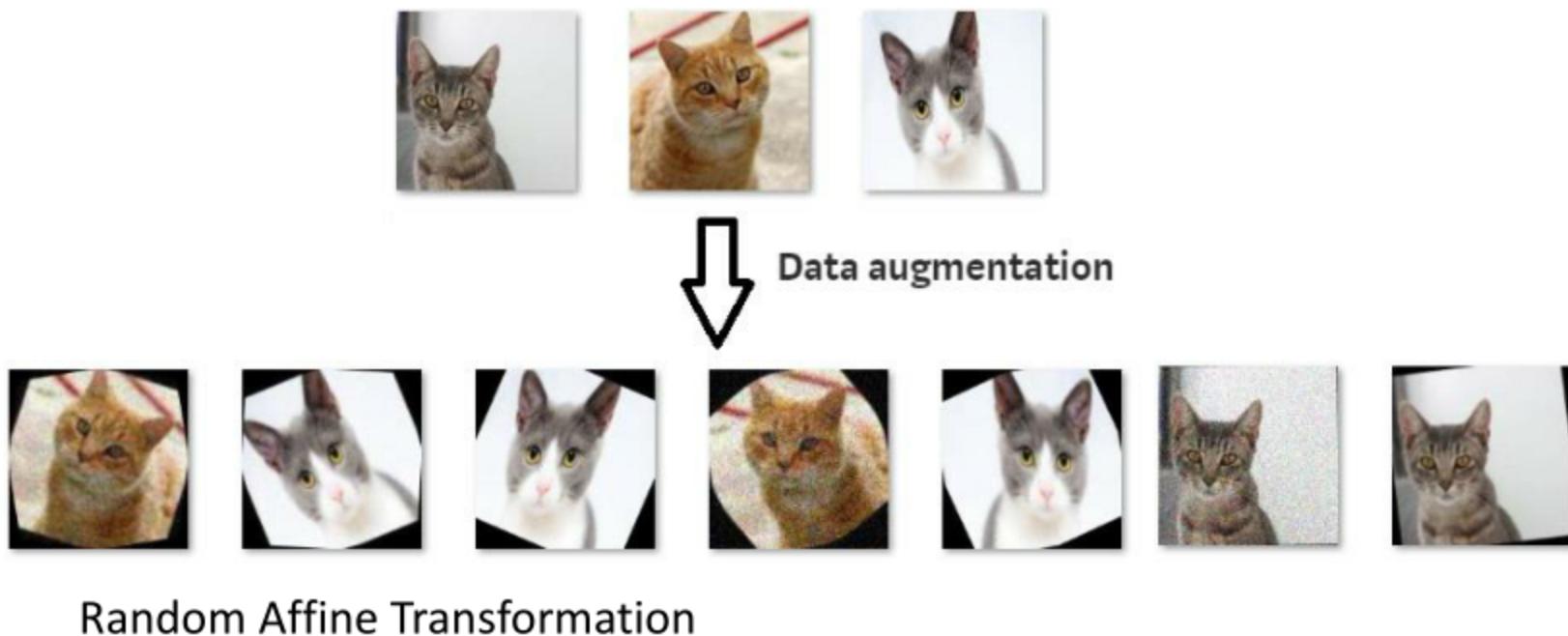
ship



truck

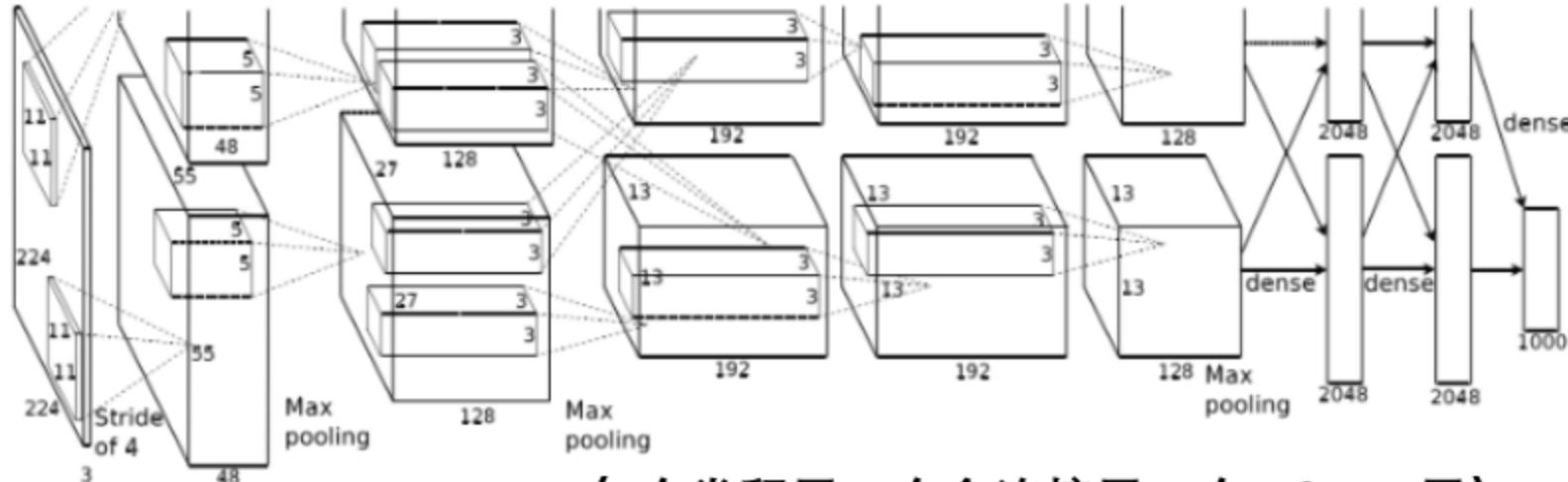


Data Augmentation



经典的CNN模型

Architecture of AlexNet (2012年)



(5个卷积层+3个全连接层+1个softmax层)

ImageNet LSVRC-2012的冠军,1000类,120万高清图像

Top5Error:26.2% → 15.3%.

结构:

由6000万个参数和650,000个神经元。由5个卷积层和其后的max-pooling层以及3个全连接层,1000-way的softmax层组成。开创性的使用“**dropout**”技术,避免了过拟合。

计算开销问题:

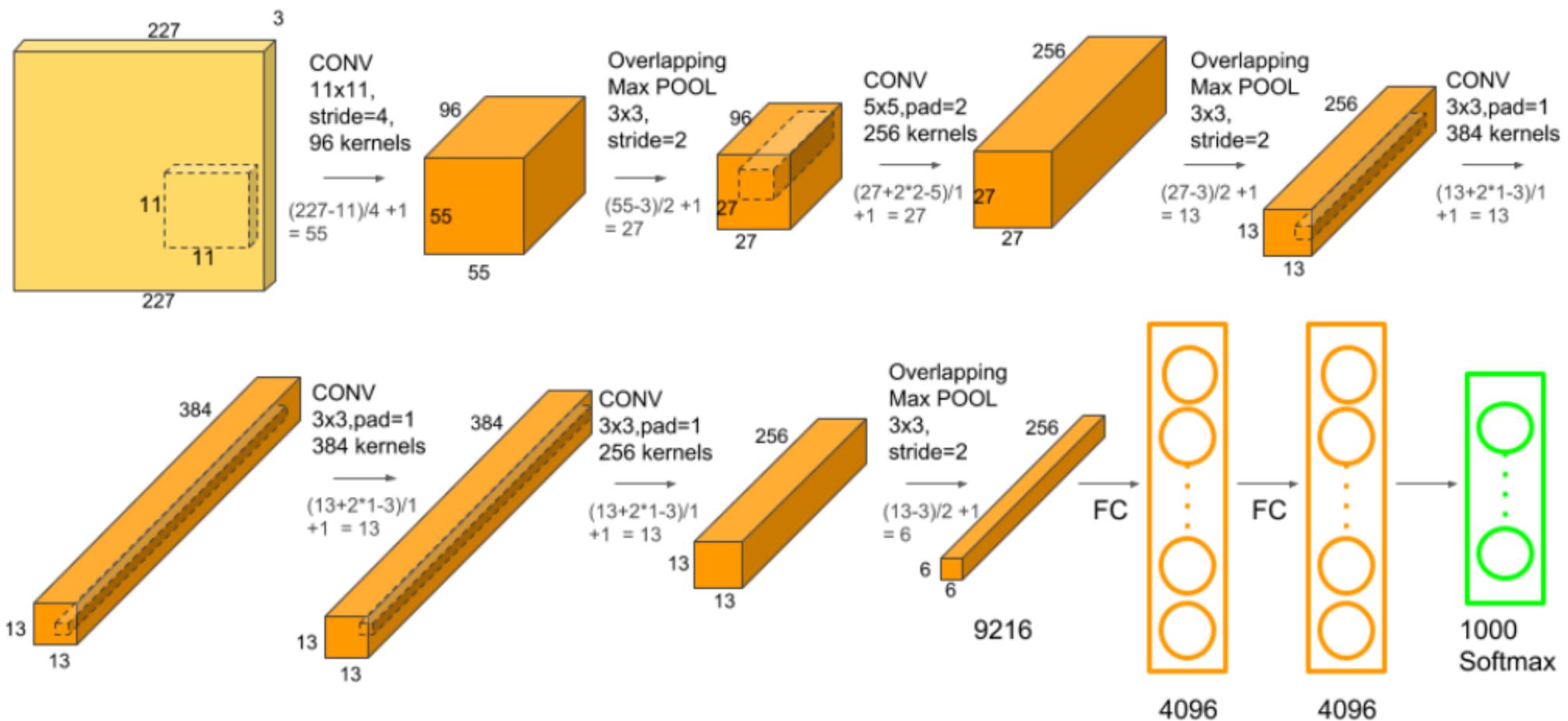
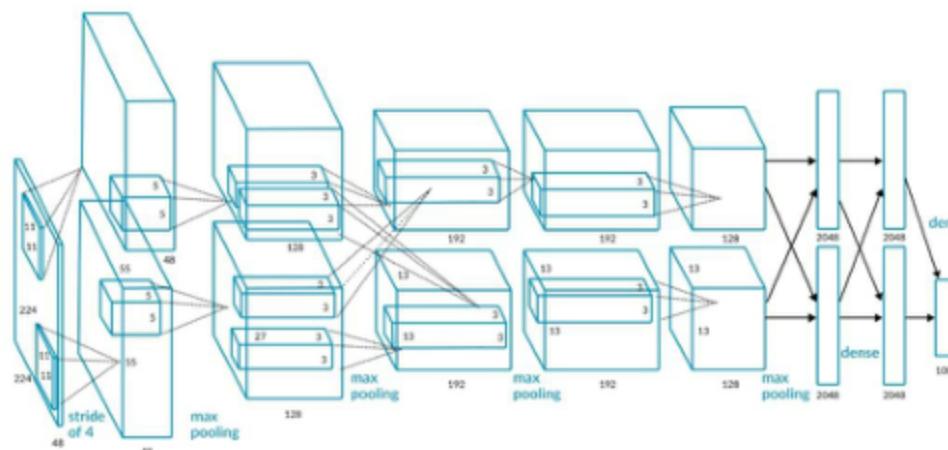
采取将网络分布在两个GPU上,在每个GPU中放置一半核(或神经元),额外的技巧: GPU间的通讯只在某些层进行。

相比于Lenet5,优势在于:

1. 网络加深
2. 同时解决过拟合
(dropout, data augmentation, LRN)
3. 利用多GPU加速计算

经典的CNN模型

Architecture of AlexNet (2012年)



经典的CNN模型

VGG (2014年)

《Very Deep Convolutional Networks for Large-Scale Image Recognition》

ILSVRC 2014比赛分类项目的第2名，
定位项目的第1名

VGGNet网络结构简洁，迁移到其他图片数据上的泛化性能非常好。VGGNet现在依然经常被用来提取图像特征，该网络训练后的模型参数在其官网上开源了，可以用来在图像分类任务上进行在训练，即：提供了非常好的初始化权重，使用较为广泛。

- VGG16包含了16个隐藏层（13个卷积层和3个全连接层），图中的D列
 - VGG19包含了19个隐藏层（16个卷积层和3个全连接层），图中的E列
- VGG网络的结构非常一致，从头到尾全部使用的是3x3的卷积和2x2的max-pooling

VGG不同的版本

VGG16包含了16个隐藏层（13个卷积层和3个全连接层） D列

VGG19包含了19个隐藏层（16个卷积层和3个全连接层） E列

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 LRN	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 conv1-256	conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 conv3-256
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

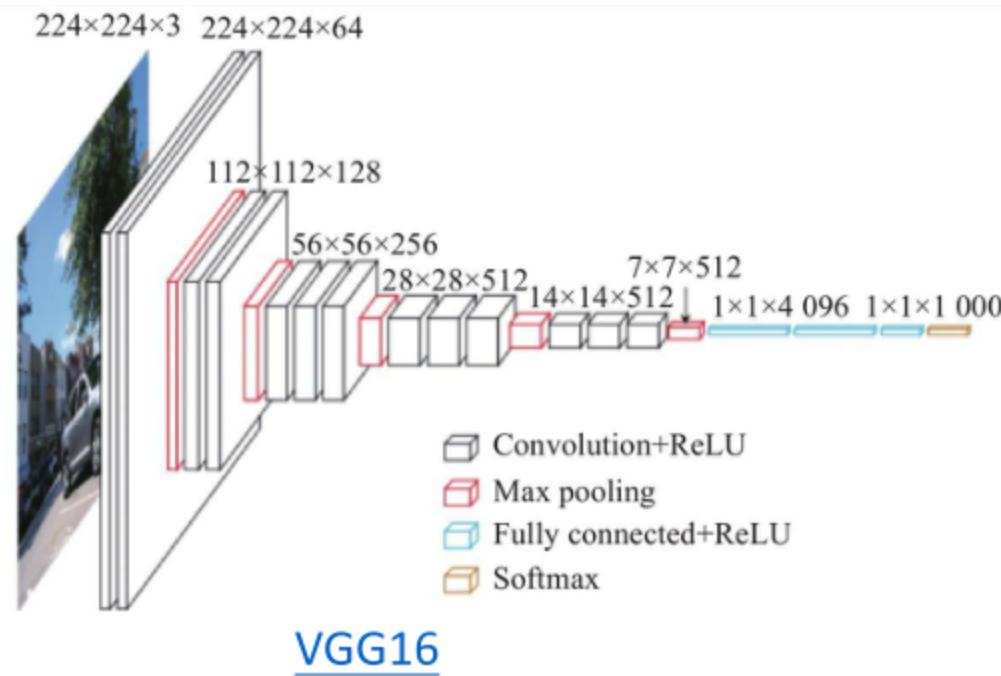
Table 2: Number of parameters (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

经典的CNN模型

VGG (2014年)

类型	Kernel 尺寸/步长(或注释)	输入尺寸	输出尺寸
第一段卷积层	3 x 3 / 1	224 x 224 x 3	224 x 224 x 64
	3 x 3 / 1	224 x 224 x 64	224 x 224 x 64
Maxpool	2 x 2 / 2	224 x 224 x 64	112 x 112 x 64
第二段卷积层	3 x 3 / 1	112 x 112 x 64	112 x 112 x 128
	3 x 3 / 1	112 x 112 x 128	112 x 112 x 128
Maxpool	2 x 2 / 2	112 x 112 x 128	56 x 56 x 128
第三段卷积层	3 x 3 / 1	56 x 56 x 128	56 x 56 x 256
	3 x 3 / 1	56 x 56 x 256	56 x 56 x 256
	3 x 3 / 1	56 x 56 x 256	56 x 56 x 256
Maxpool	2 x 2 / 2	56 x 56 x 256	28 x 28 x 256
第四段卷积层	3 x 3 / 1	28 x 28 x 256	28 x 28 x 512
	3 x 3 / 1	28 x 28 x 512	28 x 28 x 512
	3 x 3 / 1	28 x 28 x 512	28 x 28 x 512
Maxpool	2 x 2 / 2	28 x 28 x 512	14 x 14 x 512
第五段卷积层	3 x 3 / 1	14 x 14 x 512	14 x 14 x 512
	3 x 3 / 1	14 x 14 x 512	14 x 14 x 512
	3 x 3 / 1	14 x 14 x 512	14 x 14 x 512
Maxpool	2 x 2 / 2	14 x 14 x 512	7 x 7 x 512
Fc1	ReLU	25088	4096
Fc2		4096	4096
Fc3		4096	1000



VGG16

VGG优缺点
VGG优点
VGGNet的结构非常简洁，整个网络都使用了同样大小的卷积核尺寸（3x3）和最大池化尺寸（2x2）。几个小滤波器（3x3）卷积层的组合比一个大滤波器（5x5或7x7）卷积层好：验证了通过不断加深网络结构可以提升性能。VGG缺点
VGG耗费更多计算资源，并且使用了更多的参数（这里不是3x3卷积的锅），导致更多的内存占用（140M）。其中绝大多数的参数都是来自于第一个全连接层。VGG可是有3个全连接层啊！简单来说，在VGG中，使用了3个3x3卷积核来代替7x7卷积核，使用了2个3x3卷积核来代替5*5卷积核，这样做的主要目的是在保证具有相同感知野的条件下，提升了网络的深度，在一定程度上提升了神经网络的效果。为什么使用2个3x3卷积核可以来代替5*5卷积核
5x5卷积看做一个小的全连接网络在5x5区域滑动，我们可以先用一个3x3的卷积滤波器卷积，然后再用一个全连接层连接这个3x3卷积输出，这个全连接层我们也可以看做一个3x3卷积层。这样我们就可以用两个3x3卷积级联（叠加）起来代替一个5x5卷积。

经典的CNN模型

GoogLeNet (2014年)

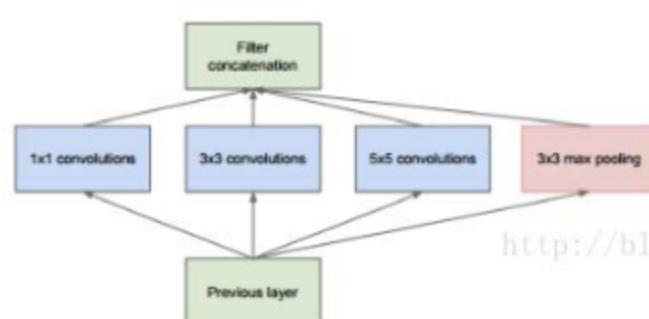
ImageNet**2014**年分类冠军，也被称为**Inception V1**。
Inception V1有22层深，参数量为5M

文章提出获得高质量模型最保险的做法就是**增加模型的深度（层数）或者是其宽度（层核或者神经元数）**，但是这里一般设计思路的情况下会出现缺陷

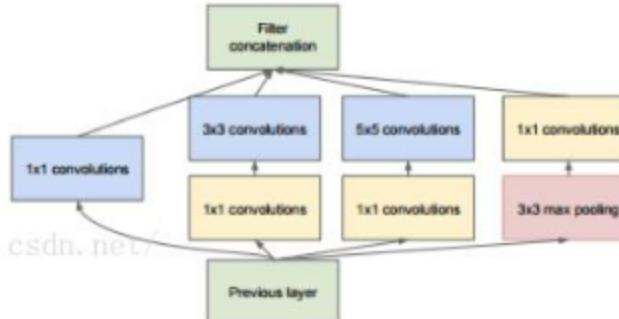
1. 若训练数据集有限，参数太多，容易过拟合；
2. 网络越大计算复杂度越大，难以应用；
3. 网络越深，梯度越往后穿越容易消失，难以优化模型。

googlenet的主要思想就是围绕这两个思路去做的：

1. 深度，**层数更深**，文章采用了22层，googlenet巧妙的在不同深度处增加了两个loss来避免上述提到的梯度消失问题，。
2. 宽度，**增加了多种核** 1×1 , 3×3 , 5×5 ，在 3×3 前， 5×5 前，max pooling后分别加上了 1×1 的卷积核起到了降低feature map厚度的作用。以下是googlenet用的inception可以称之为 **inception v1**，如下图所示：



(a) Inception module, naïve version

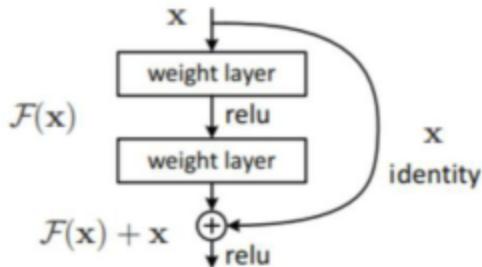
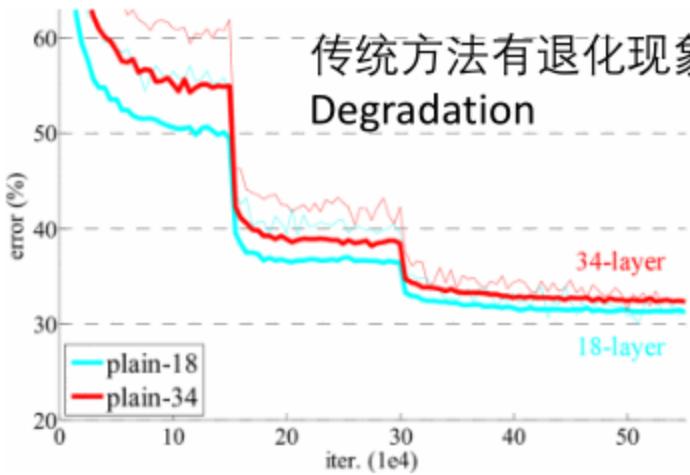


(b) Inception module with dimension reductions

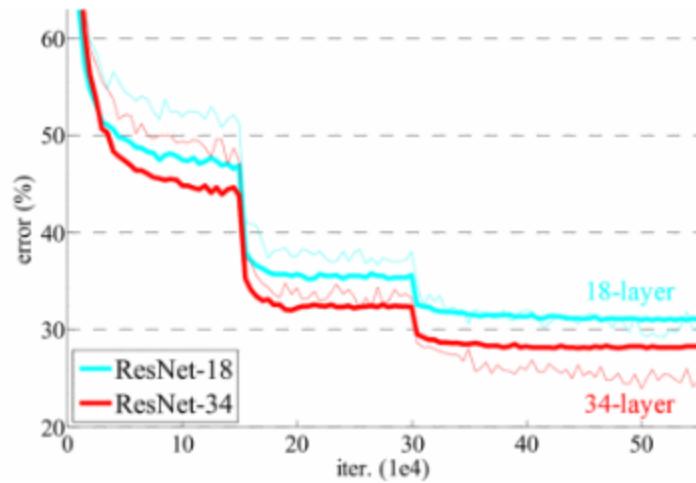
经典的CNN模型

ResNet (2015)

传统方法有退化现象
Degradation

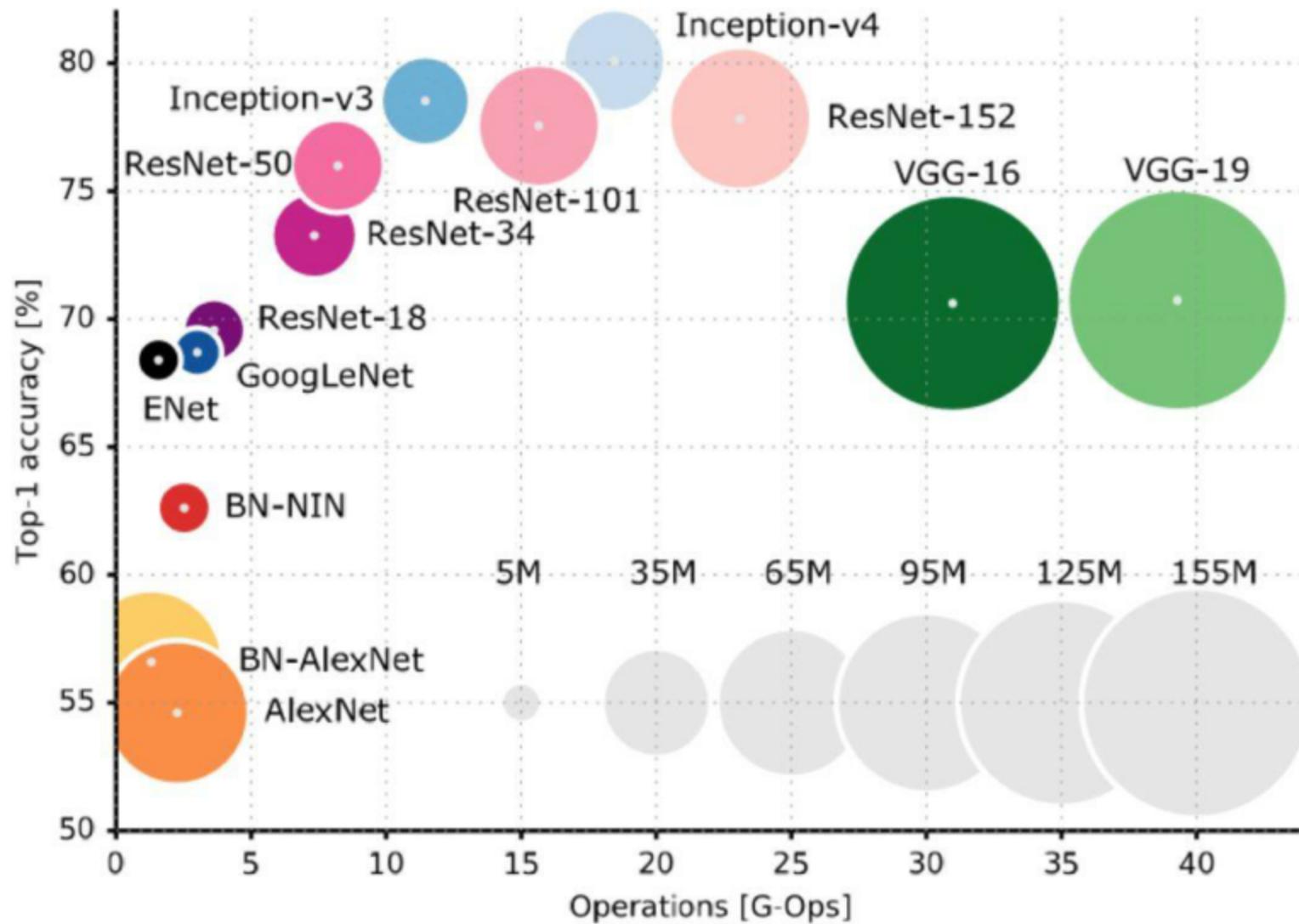


何恺明 (Kaiming He)



ResNet 在2015 年的ILSVRC中取得了冠军。

残差神经网络主要贡献是克服了深层网络的退化现象，针对退化现象发明了“快捷连接 Shortcut connection”，极大的消除了深度过大的神经网络训练困难问题。神经网络的“深度”首次突破了100层、最大的神经网络甚至超过了1000层。

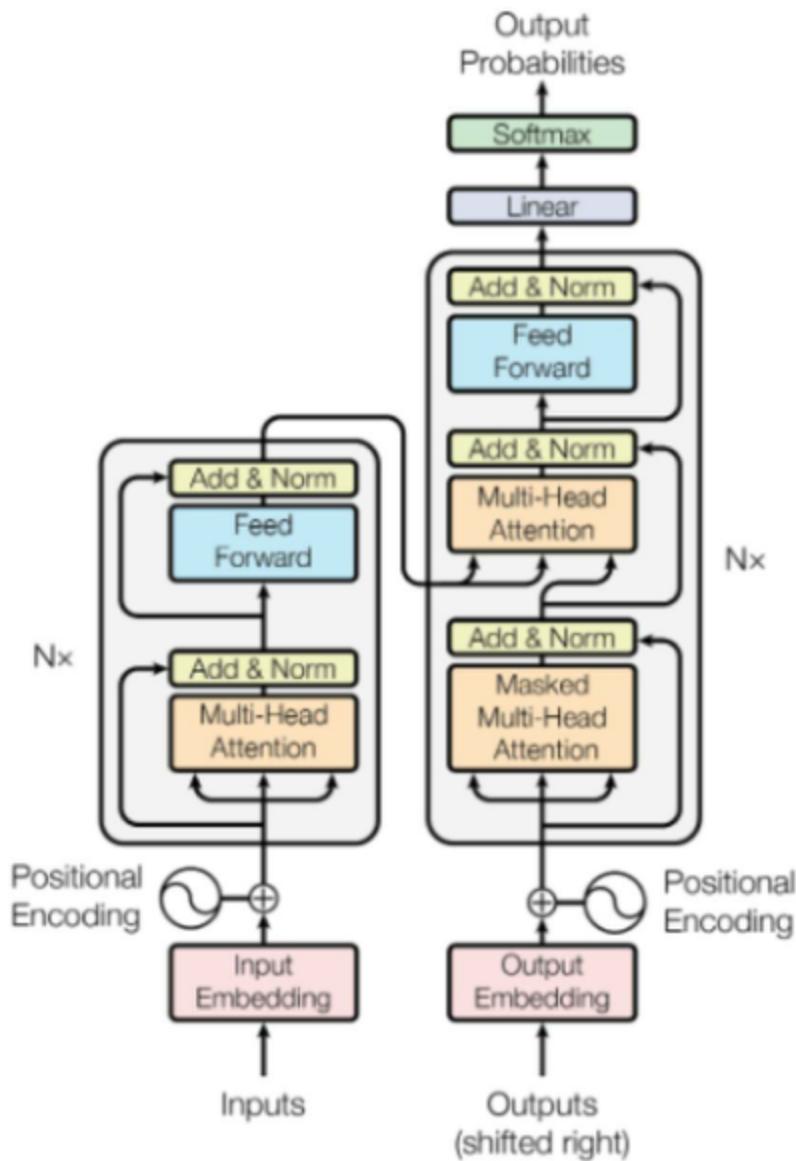


TOPS (Tera Operations Per Second) : 处理器每秒钟可进行的操作次数
1 TOPS = 1024 GOPS = 1024×1024 MOPS

卷积神经网络

如何用深度学习提取图片特征？这样做有什么好处？哔哩哔哩_bilibili

超强动画，一步一步深入浅出解释
Transformer原理！哔哩哔哩_bilibili



Attention is all you need (2017)