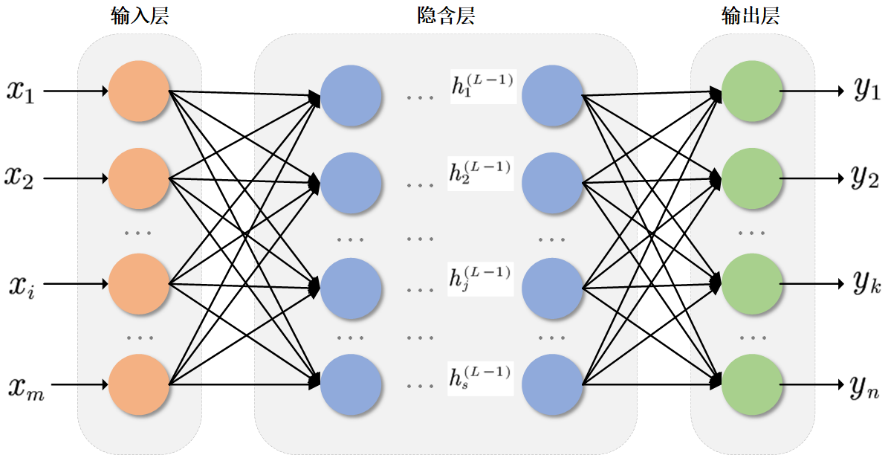
深度学习如何为半导体缺陷把脉

想象一下，在一块指甲大小的半导体芯片上，密布着数十亿个微型元件，它们是现代科技的心脏。这些芯片的完美无瑕直接关系到智能手机、电脑乃至火箭的正常运行。然而，即便是最先进的制造工艺，也难以避免微小缺陷的产生。此时我们需要**快速甄别不同的缺陷类型**，以优化我们的制造工艺。这时，深度学习这位“超级侦探”便大显身手了。

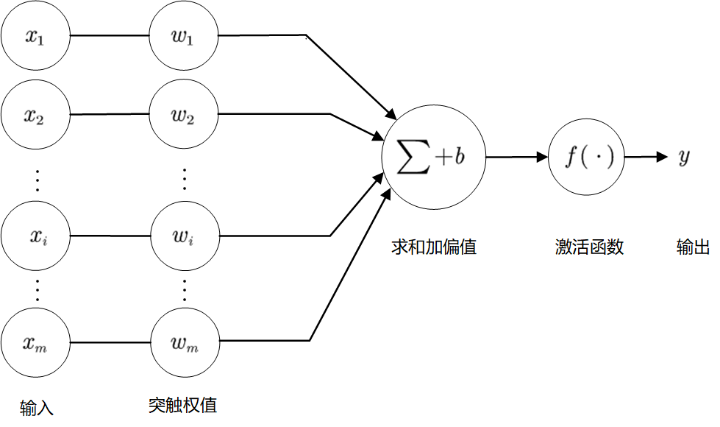
### 如何简单的理解深度学习？

想象你正在教一个孩子识别苹果，开始时，你可能需要告诉他：“看，这是红色的，圆的，上面有叶子。”但如果是深度学习，它会通过无数张苹果图片自我学习，逐渐理解苹果的特征，甚至能区分不同品种和成熟度。深度学习是一种**AI**人工智能技术，它通过多层的**神经网络**模拟人脑的工作方式，自动从大量数据中学习到模型参数，并通过模型做出决策或预测。

1. 神经网络的基础构造：神经元模型



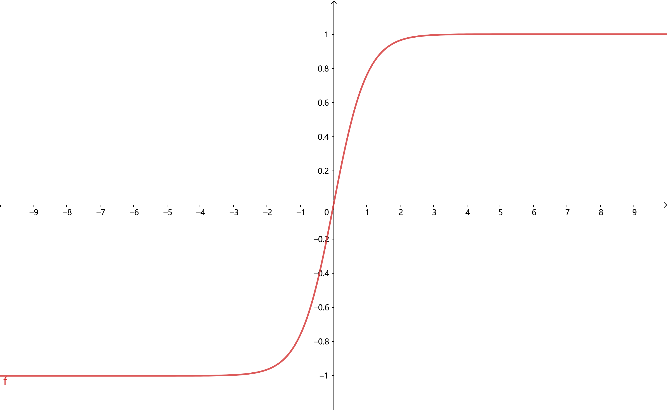
通常一个多层神经网络（如上图所示），由L层**神经元**（或称为节点）组成。每个神经元接收一组输入信号，通过加权求和后，使用一个激活函数（如Sigmoid、Tanh函数等）转换这些信号，决定是否应该“激发”并传递给下一个神经元或作为最终输出。这个过程模拟了生物神经元的兴奋或抑制行为。单个**神经元**的模型如下：



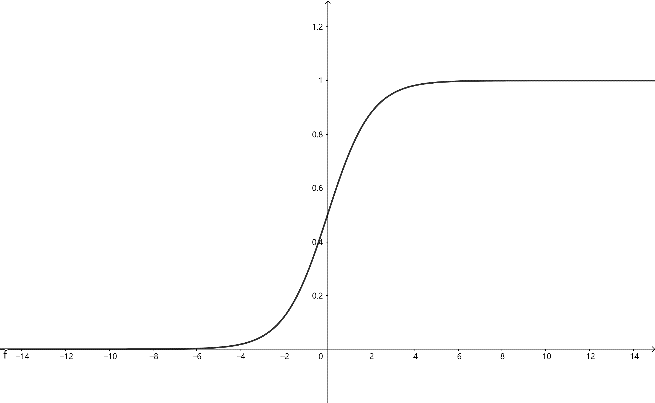
其中：X为输入的样本向量值，w为连接权重，b为偏置，f(·)为激活函数

1. 激活函数的作用

激活函数就像是神经元的一个**开关或者增压器**。在神经网络中，信息以数字信号的形式在各个神经元间传递。每个神经元会接收到一些输入值，然后对这些输入值做一番计算。**如果没有激活函数，这个计算过程就仅仅是线性的——就像数学里的加减乘除，结果很容易预测。**但现实世界的问题往往充满非线性，比如苹果和橙子的区别不仅仅在于尺寸加颜色那么简单。下面两个是常见的激活函数图像：



Tanh函数



Sigmod函数

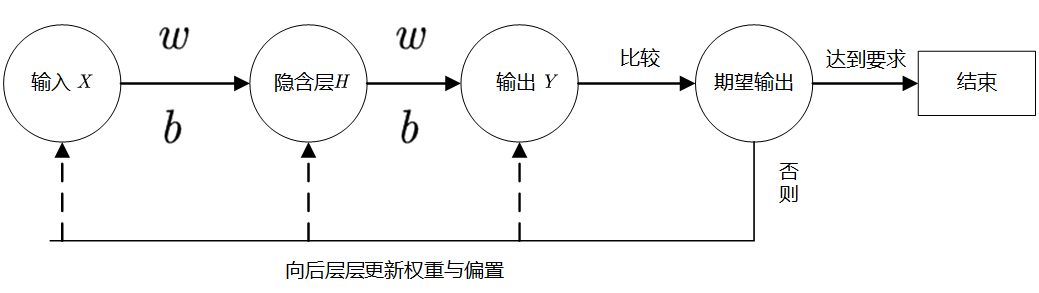
1. 神经网络的架构

神经网络主要分为三种基本类型：

1. **前馈神经网络**（Feedforward Neural Networks）：这是最简单的神经网络形式，信息流动方向单一，从输入层经过一个或多个隐藏层，最后到达输出层。每个层的神经元只与下一层的神经元相连，没有反馈回路。
2. **反馈神经网络**（Feedback Neural Networks）：与前馈网络不同，这类网络允许信息反馈，即神经元的输出可以影响其自身或同一层、前一层的神经元。这种循环连接使得网络能够进行动态处理和学习序列信息，如RNN（循环神经网络）。
3. **图神经网络**（Graph Neural Networks）：特别适用于处理非欧几里得数据，如社交网络、分子结构等。它在图结构上进行操作，每个节点都有自己的特征向量，并且可以通过边与相邻节点的信息交流。
4. 模型的学习与训练

神经网络的学习过程主要依赖于**反向传播算法**（Backpropagation）。在训练阶段，网络会接收一批已知正确答案的输入数据（称为训练样本集），通过不断调整神经元之间的连接权重，最小化预测输出与真实结果之间的差异。这个调整过程基于梯度下降法，即沿着误差减少最快的方向逐步调整权重。

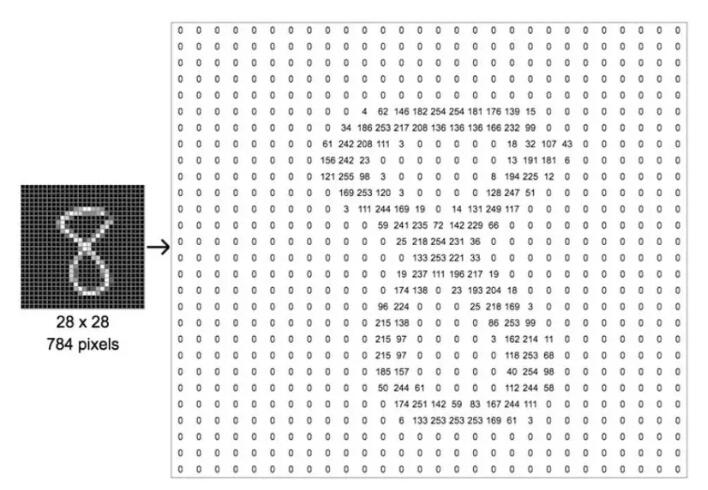
训练过程如下：



### 深度学习是如何应用到defect图像分类的？

1. 图像转化成数值计算

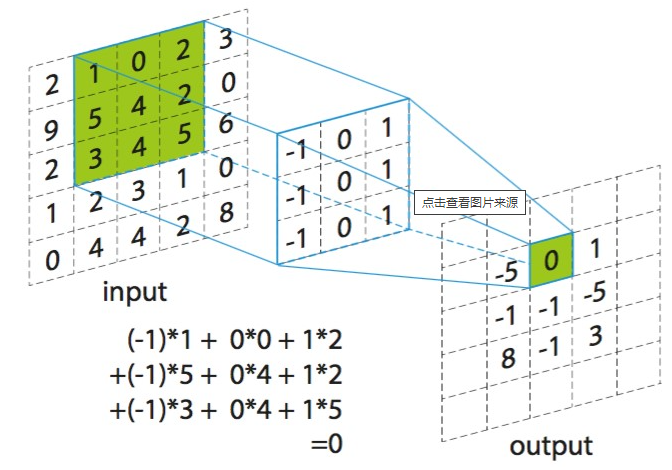
图像到像素矩阵：首先，想象你有一张彩色或灰度的图片，这张图片在计算机看来其实就是一个由像素点组成的巨大网格。每个像素点都有自己的颜色值，**如果是灰度图像，这个值代表从黑到白（0 ~ 255）的灰度程度**；如果是彩色图像，则会有红、绿、蓝三个颜色通道的值。神经网络处理图片前的第一步，就是将这张图片转换成一个数字矩阵（或多个矩阵，对于彩色图像来说），每个数字对应着一个像素的颜色强度。如下图的手写数字“8”，白色字迹部分对应到矩阵中的数值就明显偏大，黑色背景的值则为0：



1. 图像的卷积神经网络运算

接下来，就要请出我们的主角——卷积神经网络（CNN）了。CNN的工作方式就像是使用一系列的“滤镜”（称为卷积核或滤波器）在图像上滑动，这些滤镜旨在捕捉特定的特征，比如边缘、纹理或形状。

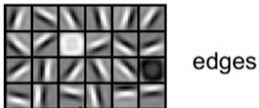
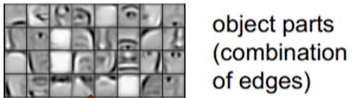
想象你有一个小的方块模板（卷积核），它在图像上滑动，每次停顿时就与下方的像素矩阵做点乘运算，然后把这些结果相加，得到一个新的数字。这个过程重复进行，直到覆盖整个图像，这样你就得到了一个新的、但尺寸较小的特征图。这个过程就是卷积运算，它能提取图像的局部特征。



每完成一次卷积运算后，激活函数就会介入，它给这些新的数字赋予生命力。就像之前提到的，激活函数确保只有那些重要的信息被加强并传递下去，不重要的信息则被削弱或过滤掉。ReLU（Rectified Linear Unit）是最常用的激活函数之一，它简单地将所有负数变为0，保留正数原样不变。

池化层：为了减少计算量并保持重要特征，网络可能会包含池化层。这一步就像是对特征图进行压缩，只保留关键部分，比如取每个小区域的最大值或平均值，这样图像变得更小，但关键特征得以保留。

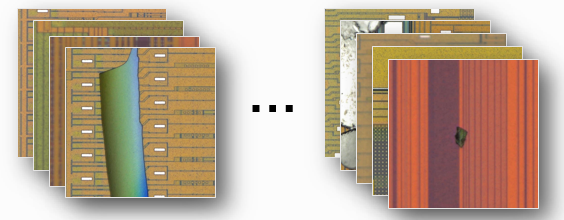
多层叠加：上面的过程会在多个卷积层中重复，每个层级的滤镜会越来越能捕捉复杂的特征，从最初的边缘到后来的轮廓、形状，乃至更高层次的概念（如下图人脸特征的提取）。

全连接层：最终，经过几轮特征提取后，图像的高级抽象信息会被送到神经网络的最后几层，通常是全连接层。这里，信息被整合并用于做出最终决策，比如判断这幅图像是人还是其他动物。

1. 大量数据准备

我们的资深工程师，可以快速的分辨不同的defect类型，这是由于长时间的经验积累，见多识广，熟能生巧，对于深度学习系统也是这样，需要“见多识广”，因此我们积累了大量的学习样本，我们的资深工程师会对收集到的defect数据进行标注归类，其中既包括正常的图片，也包括有缺陷的图片。这样就可以喂给我们的模型进行特征参数的学习。但我们必须保证样本是被正确标注的，否则模型将学到错误信息，导致模型判断能力不足。



**大量的学习样本**

1. 模型训练

接着，构建深度神经网络，就像搭建层层叠叠的过滤器。每一层负责学习defect的不同特征，从简单的边缘、纹理到复杂的图案结构。

训练完成后，模型导入生产系统，机台在拍摄图像后，模型可以快速预测缺陷类型，大大节省工程师的人力，并不知疲倦的一直工作。