基于深度学习的图像内容检索架构及其在K12教育资源中的应用

0 概述

以图搜图，又称基于内容的图像检索或反向图片搜索。它是互联网领域兴起的一种新型的图像搜索方式，可以更加直观的表达用户的搜索意图，并且能客观的帮助用户实现搜索。“ 以图搜图 ” 已经成为图像搜索引擎的发展方向。从鼻祖 Tineye 开始，google、百度、搜狗、360 等越来越多的国内外搜索引擎公司甚至其他互联网公司均加入到这一领域。

2.1 发展现状 据统计，目前国内外拥有以图搜图技术并已经成功推广 应用的搜索引擎公司及相关信息如表 1 所示。 以图搜图技术帮助互联网用户通过某张图片的一部分或 者全部，来搜索这张图片的相关信息。就目前而言，它主要 应用在以下两个方面： （1）用户通过上传图片可以进行全网同图片或相似图片 的搜索，可以搜索到高分辨率图片、无水印图片、图片背景、 发布时间来源以及使用过这张图的所有网站等信息。 （2）用户可以通过上传图片进行全网同款或者相似商品 的搜索和比价。

2.2 发展前景 以图搜图技术除了用于全网图片搜索、相似商品的比价 外，还在以下方面发挥着重要作用。 （1）便利群众生活的工具：通过随拍，随时了解想要的 信息。例如通过拍卧室设计照片，能够自动识别照片中各个 家具的价格等信息，并显示出来。 （2）行业的技术辅助手段：通过以图搜图的图像特征提 取识别技术，能够帮助各行业各系统精确识别个体目标，提 高工作效率。例如，在交通系统，通过以图搜图技术，能够 快速识别交通肇事车辆；在公安系统，通过图像特征识别在 2017.12 53 人脸数据库中快速识别符合罪犯特征的个体，能够有效缩小 排查范围，节约各方资源。 （3）扩展实现基于图片的视频搜索：视频搜索的应用场 景覆盖我们生活和工作中的方方面面。简单来说，大家目前 在各大互联网视频网站、电影视频网站搜索自己感兴趣视频 的时候，只能通过文字关键字来搜索。而如果集成了以图搜 视频的技术，就可以通过某个电影场景、片段、海报等图片 内容来直接搜索。

以图搜图传统方法：

图像搜索的算法一般分为三个步骤：

（1）将目标图片进行特征提取。描述图像的算法很多，用的比较多的有 SIFT 描述子，指纹算法函数，bundling features 算法，hash function（散列函数）等。也可以根据不同的图像，设计不同的算法，比如图像局部 N 阶矩的方法提

取图像特征。

（2）将图像特征信息进行编码，并将海量图像编码做成查找表。对于目标图像，可以对分辨率较大的图像进行降采样，减少运算量后再进行图像特征提取和编码处理。

（3）相似度匹配运算。利用目标图像的编码值，在图像搜索引擎中的图像数据库进行全局或是局部的相似度计算，再根据所需要的鲁棒性，设定阈值，然后将相似度高的图片预保留下来，最后筛选最佳匹配图片。

传统方法的缺陷：

特征提取的不充分或无用特征的提取，这将直接导致根据提取得到的特征匹配的到的图片不是我们实际所需要的图片，在教育资源的搜索上直接反映为颜色、画风等的相同但是搜索到的资源的内容相关性却不高，具体表现为文字等内容的无关性

1.1深度学习简介：

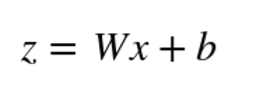
深度学习的基本思想是通过有监督或无监督的方式学习层次化的特征表达，对目标进行从底层到高层的描述，各层的特征不是利用人工工程来设计的，而是使用一种通用的学习过程从数据中学到的，含有隐层的多层感知器就是一种深度学习结构。深度学习模型的训练过程是在大量样本数据的基础上进行的，采用反向传播技术进行权值的迭代优化。在视频图像检测与识别领域，卷积神经网络（CNN）是主流的实现手段，通过设计出多层的人工神经网络进行目标的特征提取和识别检测，目前，基于深度学习技术在行人结构化、车辆结构化、人脸识别等方面取得了突破性的进步。

深度学习架构可以由若干类型的层组成。例如，AlxNETs 包含卷积、全连接层、归一化层和最大池化层。现在我们将依次查看每一层的内容。

**全连接层**

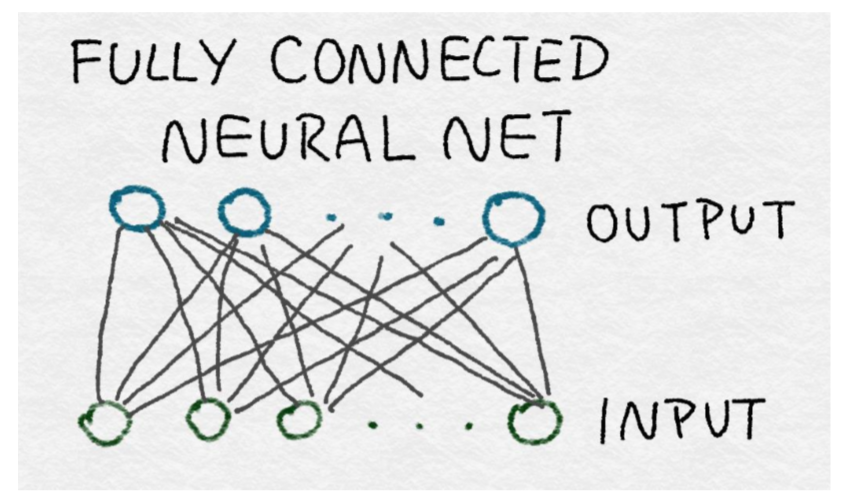
所有神经网络的核心是输入的线性函数。我们在第4章中遇到的逻辑回归是神经网络的一个示例。全连接的神经网络只是所有输入特征的一组线性函数。回想一个线性函数可以被写为输入特征向量与权重向量之间的内积，加上一个可能的常数项。线性函数的集合可以表示为矩阵向量乘积，其中权重向量成为权重矩阵。

**全连接层的数学定义**



W的每一行是将整个输入向量X映射成Z中的单个输出的权重向量。b是表示每个神经元恒定偏移（或偏置）的标量。

全连接层之所以如此命名，是因为在每一个输入都要在每个输出中使用。在数学上，这意味着对矩阵W中的值没有限制。（如我们将很快看到的，卷积层仅利用每个输出的一小部分输入。）在图中，一个完全连接的神经网络可以由一个完整的二部图表示，其中前一层的每个结点输出都连接到下一层的每个输入。

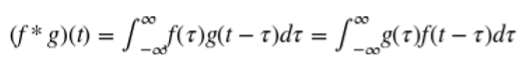


全连接层包含尽可能多的参数。因此，它们是昂贵的。这种密集连接允许网络检测可能涉及所有输入的全局模式。由于这个原因，AlexNet 的最后两层完全连接。在输入为条件下输出仍然是相互独立的。

**卷积层**

与全连接层相反，卷积层仅使用每个输出的输入子集。通过在输入上移动窗，每次使用几个特征产生输出。为了简单起见，可以对输入的不同集合使用相同的权重，而不是重新学习新权重。数学上，卷积算子以两个函数作为输入，并产生一个函数作为输出。它翻转一个输入函数，将其移动到另一个函数上，并在每个点上在乘法曲线下输出总面积。计算曲线下总面积的方法是取其积分。操作符在输入中是对称的，这意味着不管我们翻转第一个输入还是第二个输入，输出都是一样的。

卷积定义为：



我们已经看到了一个简单的卷积的示例，当我们看着图像梯度（“图像梯度”）。但是卷积的数学定义似乎仍有点复杂。用信号处理的示例来解释卷积后的结果是最容易的。

想象一下，我们有一个小黑匣子。为了看到黑匣子的作用，我们通过一个单一的刺激单位。无论输出看起来如何，我们记录在一张纸上。我们等到对最初的刺激没有反应为止。随时间变化的函数称为响应函数；我们称之为响应函数g(t)。

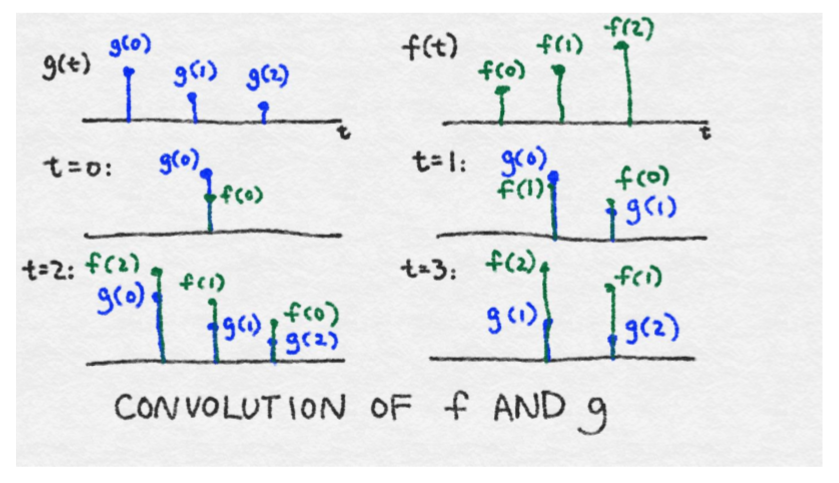
想想一下现在我们有一个疯狂的函数f(t)，随后将它输入黑盒中。在时间t=0时，f(0)与黑盒进行通讯，随后用f(0)乘以g(0)。在时间t=1，f(1)进入黑盒，随后与g(0)相乘。在相同的时间，黑盒持续回复信号f(0)，它现在是月g(1)相乘了。所以在t=1时的输出为(f(0)\*g(1))+(f(1)\*g(0))。在t=2时，输出会变得更复杂，当f(2)进入图片后，这时f(0)与f(1)持续产生回复。所以在t=2时的总输出为(f(0)\*g(2))+(f(1)\*g(1))+(f(2)\*g(0))。通过这种方式，响应函数在时间上有效地被翻转，其中τ=0总是与当前进入黑匣子的信息进行交互，并且响应函数的尾部与先前发生的函数进行交互。图 8—8 展示了这一过程。到目前为止，为了便于描述，我们已经把时间离散了。在现实中，时间是连续的，所以总和是一个积分。

这个黑箱被称为线性系统，因为它不比标量乘法和求和更疯狂。卷积算子清楚地捕捉线性系统的影响。

**卷积的思想**

卷积算子捕获线性系统的效果，该线性系统将输入信号与其响应函数相乘，求出所有过去输入响应的和。

在上面的示例中，g(t)用来表示响应函数，f(t)表示输入。但是由于卷积是对称的，响应和输出实际上并不重要。输出只是两者的结合。g(t)也称为滤波器。



图像是二维信号，所以我们需要一个二维滤波器。二维卷积滤波器通过取两个变量的积分来推广一维情形。由于数字图像具有离散像素，卷积积分变成离散和。此外，由于像素的数量是有限的，滤波函数只需要有限数量的元素。在图像处理中，二维卷积滤波器也被称为核或掩模。

2 维卷积的离散定义

当将卷积滤波器应用于图像时，我们不需要定义一个覆盖整个图像的巨型滤波器。相反，只覆盖几个像素的滤波器就够了，并且在图像上应用相同的滤波器，并在在水平和垂直像素方向上移动。

因为在图像中使用相同的滤波器，所以我们只需要定义一组小的参数。权衡是滤波器只能在一个小像素邻域内吸收信息。换言之，卷积神经网络识别局部信息而不是全局信息。

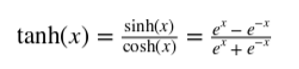
## 整流线性单元（Relu）变换

神经网络的输出通常通过另一个非线性变换，也称为激活函数。常见的选择是 tanh 函数（在 -1 和 1 之间有界的光滑非线性函数），sigmoid 函数（0 到 1 之间的平滑非线性函数），或者称为整流线性单元的函数（Relu）。Relu 是一个线性函数的简单变化，其中负部分被归零。换言之，它修剪了负值，但留下无穷的正边界。Relu 的范围从 0 到无穷大。

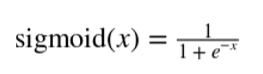
整流线性单元（Relu）是线性函数，负部分归零。

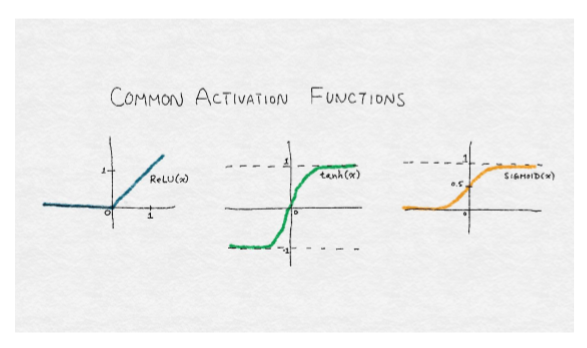
Macintosh HD:Users:apple:Downloads:1.png

tanh 函数，在 -1 和 1 之间有界的光滑非线性函数。



sigmoid 函数，0 到 1 之间的平滑非线性函数。





Relu 变换对原始图像或高斯滤波器等非负函数没有影响。然而，经过训练的神经网络，无论是完全连接的还是卷积的，都有可能输出负值。AlxNet 使用 Relu 代替其他变换，在训练过程中可以更快的收敛，它适用于每一个卷积和全连接层。

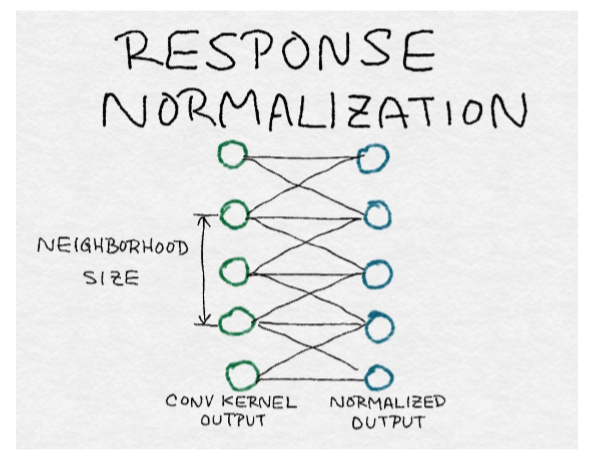
## 响应归一化层

在第 4 章和本章之前的讨论之后，归一化对大家来说应该是一个熟悉的概念。归一化将个体输出通过集体总响应的函数来划分。因此，理解归一化的另一种方式是，它在邻居之间产生竞争，因为现在每个输出的强度都相对于其邻居进行测量。AlexNet 在不同内核的每个位置上归一化输出。

局部响应归一化引起相邻核之间的竞争



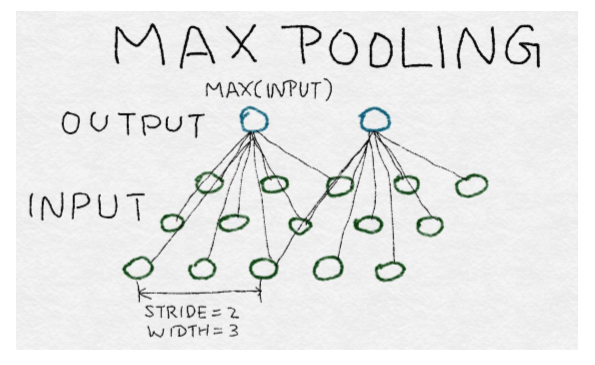
其中，Xk是第k个核的输出，yk是相对于邻域中的其他核的归一化响应。对每个输出位置分别执行归一化。换言之，对于每个输出位置(i, j)，在附近的卷积核输出上进行归一化。注意，这与在图像邻域或输出位置上的归一化不相同。内核邻域的尺寸c、α和β的大小超参数都是通过图像验证集调整的。



## 池化层

池化层将多个输入组合成单个输出。当卷积滤波器在图像上移动时，它为其尺寸下的每个邻域生成输出。池化层迫使局部图像邻域产生一个值而不是许多值。这减少了在深度学习网络的中间层中的输出数量，这有效地减少了过拟合训练数据的概率。

有多种方法汇集输入：平均，求和（或计算一个广义范数），或取最大值。池化层通过图像或中间输出层移动。Alxnet 使用最大池化层，以 2 像素（或输出）的步幅移动图像，并在 3 个邻居之间汇集。

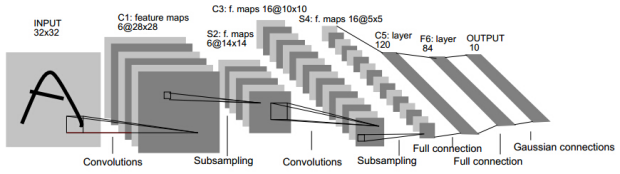


相应的我们可以综合采用的特征提取参考模型有：

1）Lenet（LeNet-5, 5大层）

文章：Gradient-Based Learning Applied to Document Recognition（1989）

下图是广为流传LeNet的网络结构，麻雀虽小，但五脏俱全，卷积层、pooling层、全连接层，这些都是现代CNN网络的基本组件。



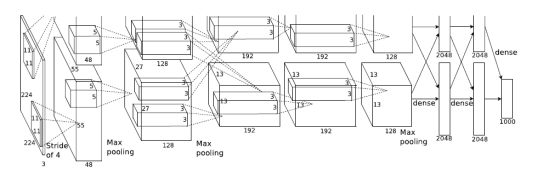
1）输入尺寸：32\*32；2）卷积层：3个；3）降采样层：2个；4）全连接层：1个；5）输出：10个类别（数字0-9的概率）

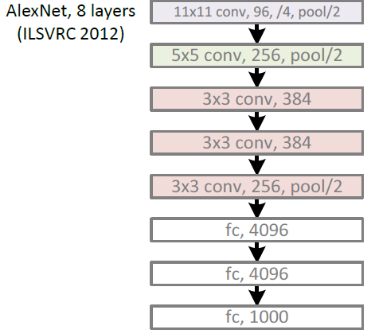
（2）Alexnet（7层）

文章：ImageNet Classification with Deep Convolutional Neural Network（2012）http://blog.csdn.net/whiteinblue/article/details/43202399（非常好）

见下图，别看只有寥寥八层（不算input层，mine：这里是把卷积层到pooling层放一起，认为是一层，认为是这层卷积层到下一层卷积层统一是一层，对于层的概念好像有点模糊，各处可能会不一样吧），但是它有60M（60 million）以上的参数总量，事实上在参数量上比后面的网络都大。

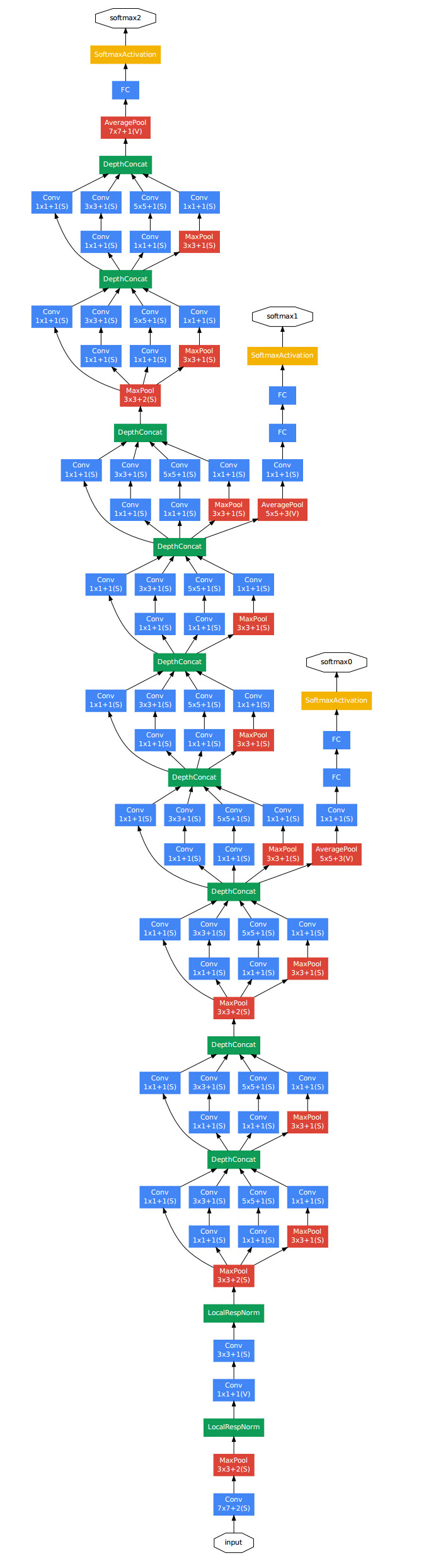
。





（3）GoogleNet(22层)

文章：Going deeper with convolutions（2014）

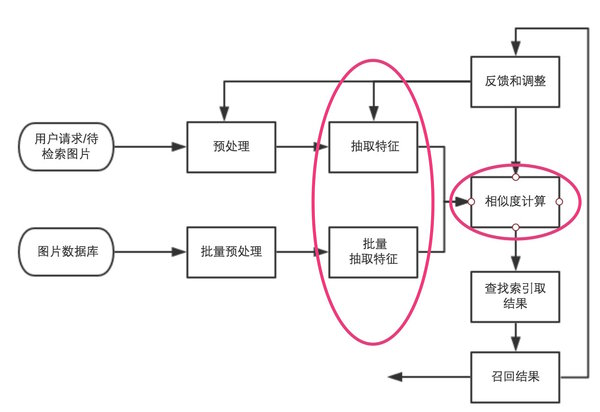


除此之外，还有VGG的多个版本的模型可以用于图像的特征提取

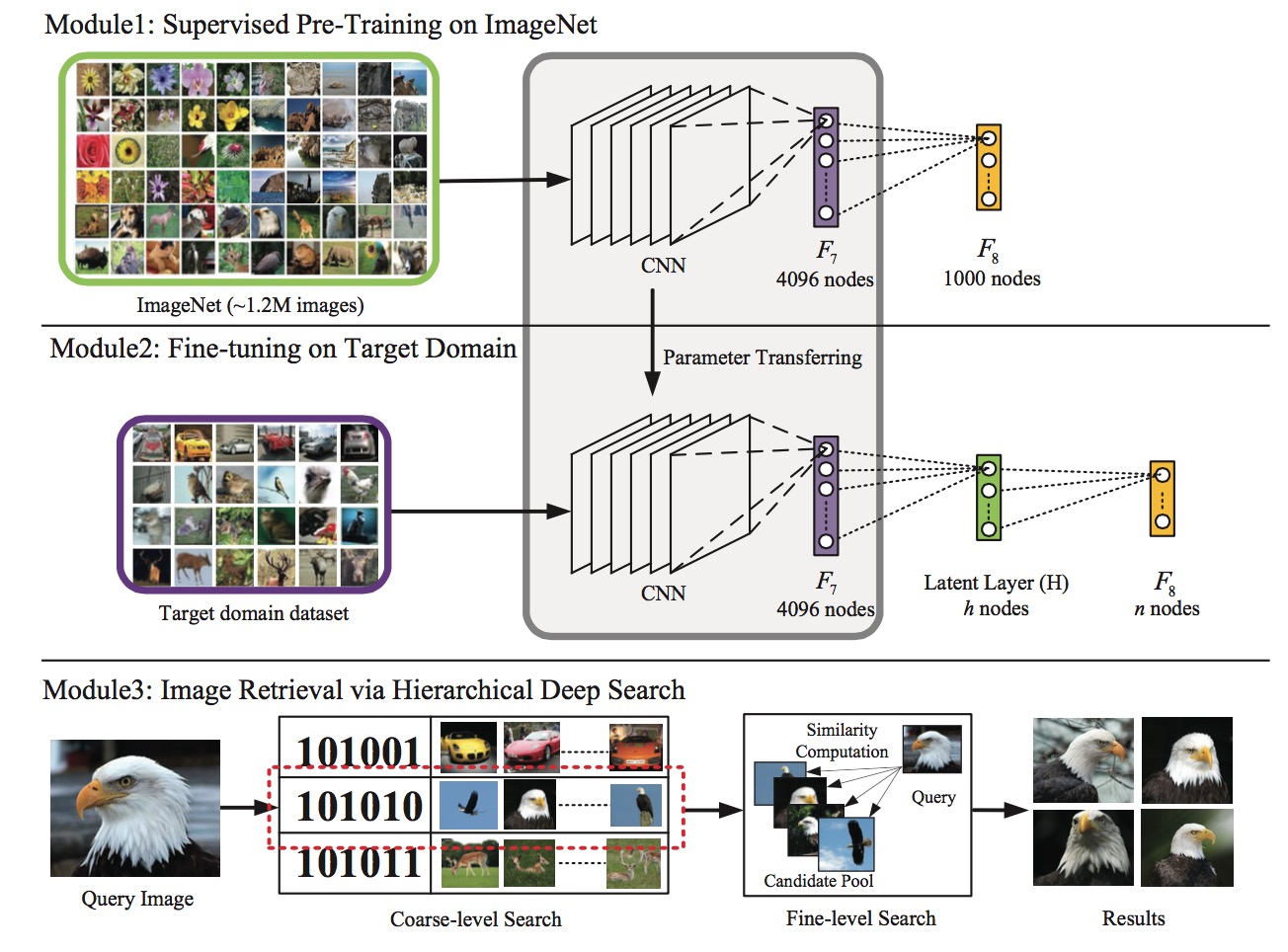
1.2 以图搜图架构设计

基于深度学习的以图搜图架构主要包括图像的特征提取和标签化两个组成部分。视频检索库构建采用深度学习方法将教育相关的图片进行预处理，提取其中的目标类型和深度特征;标签化模块通过提取待检目标的特征。最终与检索库中的特征标签和特征向量进行相似度比对，按相似度高低展现检索结果。

基于内容的图像检索系统最大的难点在于图片特征的表达能力及近似最近邻的查找，其一为大部分神经网络产出的中间层特征维度非常高，比如Krizhevsky等的在2012的ImageNet比赛中用到的AlexNet神经网，第7层的输出包含丰富的图像信息，但是维度高达4096维。4096维的浮点数向量与4096维的浮点数向量之间求相似度，运算量较大，因此Babenko等人在论文Neural codes for image retrieval中提出用PCA对4096维的特征进行PCA降维压缩，然后用于基于内容的图像检索，此场景中效果优于大部分传统图像特征。同时因为高维度的特征之间相似度运算会消耗一定的时间，因此线性地逐个比对数据库中特征向量是显然不可取的。大部分的ANN技术都是将高维特征向量压缩到低维度空间，并且以01二值的方式表达，因为在低维空间中计算两个二值向量的汉明距离速度非常快，因此可以在一定程度上缓解时效问题。ANN的这部分hash映射是在拿到特征之外做的，我们可以试图让卷积神经网在训练过程中学习出对应的『二值检索向量』，或者我们可以理解成对全部图先做了一个分桶操作，每次检索的时候只取本桶和临近桶的图片作比对，而不是在全域做比对，以提高检索速度。



我们实现『二值检索向量』借鉴如下：在Krizhevsky等2012年用于ImageNet中的卷积神经网络结构基础上，在第7层(4096个神经元)和output层之间多加了一个隐层(全连接层)。隐层的神经元激励函数，可以选用sigmoid，这样输出值在0-1之间值，可以设定阈值(比如说0.5)之后，将这一层输出变换为01二值向量作为『二值检索向量』，这样在使用卷积神经网做图像分类训练的过程中，会『学到』和结果类别最接近的01二值串，也可以理解成，我们把第7层4096维的输出特征向量，通过神经元关联压缩成一个低维度的01向量，但不同于其他的降维和二值操作，这是在一个神经网络里完成的，每对图片做一次完整的前向运算拿到类别，就产出了表征图像丰富信息的第7层output(4096维)和代表图片分桶的第8层output(神经元个数自己指定，一般都不会很多，因此维度不会很高)。引用论文中的图例解释就是如下结构：



上方图为ImageNet比赛中使用的卷积神经网络；中间图为调整后，在第7层和output层之间添加隐层(假设为128个神经元)后的卷积神经网络，我们将复用ImageNet中得到最终模型的前7层权重做fine-tuning，得到第7层、8层和output层之间的权重。下方图为实际检索过程，对于所有的图片做卷积神经网络前向运算得到第7层4096维特征向量和第8层128维输出(设定阈值0.5之后可以转成01二值检索向量)，对于待检索的图片，同样得到4096维特征向量和128维01二值检索向量，在数据库中查找二值检索向量对应『桶』内图片，比对4096维特征向量之间距离，做重拍即得到最终结果。图上的检索例子比较直观，对于待检索的"鹰"图像，算得二值检索向量为101010，取出桶内图片(可以看到基本也都为鹰)，比对4096维特征向量之间距离，重新排序拿得到最后的检索结果。

理论部分吹完了，后面就是具体应用了