**CNN发展史**

# 一、简述

最近一直在看深度学习相关的东西，而CNN是深度学习中的核心算法之一，也是2012年以来将人工智能推向风口浪尖的推手。今天我来与大家一起分享一下CNN的发展历史，看看各路大牛是如何克服难题，并造就了现在CNN网络结构百花齐放的景象。

1、卷积神经网络

首先来简单说一下，什么是CNN，**卷积神经网络**，它是一种**人工神经网络**的结构，是从猫的视觉神经结构中得到了灵感，进而模拟其结构设计出来的一种人工神经网络结构。至于什么是人工神经网络，我默认大家都是知道的，在这里就不解释了。（要了解CNN，需要先了解什么是人工神经网络算法，可以参考我的另一篇文章：[《【神经网络算法入门】详细推导全连接神经网络算法及反向传播算法+Python实现代码》](https://blog.csdn.net/cuiyuan605/article/details/84307323)）

这种结构通过卷积核来获取“感受野”范围内数据之间的关系特征。一张图片里，相邻的像素显然是有更强的相关性，相比于全连接神经网络，CNN突出了这种相邻的关系特征，因而更加准确的获取了图片内的有用信息。

具体卷积核是怎么工作的我找了几张图，大家感受一下：

[CNN1](CNN1.jpg) 卷积核的工作原理

[CNN2](CNN2.gif) 卷积核工作原理动图

[CNN3](CNN3.gif) 多通道多卷积核的工作原理

[CNN4](CNN4.gif) 多通道卷积的具体计算方法

所以说，**CNN的原理就是通过卷积核获取图像的空间关系特征**，并使用BP（反向传播）算法调整卷积核的参数，最终得到一个可以有效获取图片内有用信息的模型。

下面我们就从CNN的发展历史角度，来了解CNN是如何一步一步发展，在这过程中解决了什么问题。

2、比赛

讲历史就必须有一个主线，而我要说的这个历史主线就是**ILSVRC**，一个想了解CNN必须要了解的计算机视觉领域最权威的学术竞赛。

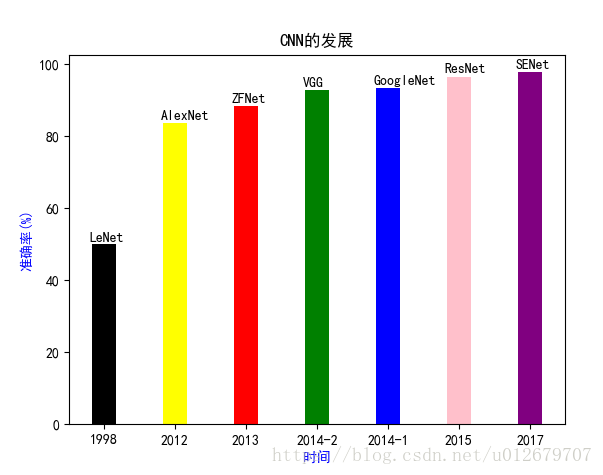
ILSVRC（ImageNet Large Scale Visual Recognition Challenge）竞赛(2010-2017)，这个竞赛所使用的数据集，是由斯坦福大学李飞飞教授主导，包含了超过1400万张全尺寸的有标记图片。ILSVRC比赛会每年从ImageNet数据集中抽出部分样本用于比赛，以2012年为例，比赛的训练集包含1281167张图片，验证集包含50000张图片，测试集为100000张图片。竞赛的项目主要包括图像的分类和定位（CLS-LOC）、目标检测（DET）、视频目标检测（VID）和场景分类（Scene）。

我下面讲的，主要是在最基本的分类问题上，历届表现优异的模型。不过ILSVRC从2017之后就停办了，另外一个计算机视觉领域的赛事，就成为了该领域权威的竞赛，它就是从2014年开始由微软举办的MS COCO竞赛。

MS COCO（Microsoft COCO: Common Objects in Context）竞赛（2014-） 物体检测、全景分割、人体关键点、姿态估计（DensePose）；2018年开始又新增了，街景检测、街景全景分割。在这里就不详细介绍了，感兴趣的可以自己研究研究，如果能在这个比赛中获得不错的名次，百万年薪不是梦！ [<传送门>](http://cocodataset.org/)

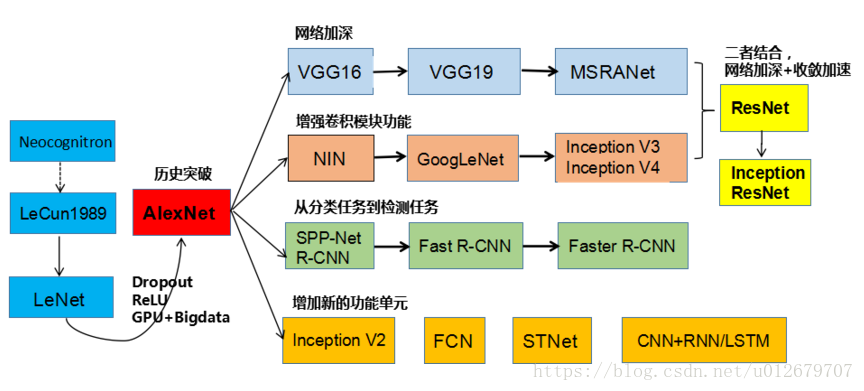
# 二、CNN发展史

先偷一张图给大家看看，就是一系列里程碑式的模型，分别在ImageNet数据集图像分类的准确率表现：



这里面除了LeNet以外，其他都是在ILSVRC竞赛里表现优异的模型，为什么把LeNet放在这里面呢，我想来作者的想法一是做个对比，另外一个就是对LeNet表示一下尊重，因为LeNet是其他所有CNN网络模型的鼻祖，它的作者就是CNN之父、2018年图灵奖得主、深度学习三主神之一的Yann Lecun扬·莱坎。

下面我们来捋一捋各个模型之间的关系，于是我又偷了一张图：



2012年AlexNet的发布使人工智能算法迎来了历史性的突破，才造就了之后的舆论关注和CNN的研究爆发式的增长。之后一系列著名的CNN模型都是在AlexNet的基础上演变出来的，它门主要分为两个流派，一个是以加深网络为主的VGG流派，另一个是以增强卷积模块功能为主的NIN流派。之后两个流派相结合产生了Inception ResNet。

如果说AlexNet是之后所有CNN模型的爸爸，那么LeNet就是所有这些模型的爷爷。下面我们就从LeNet开始一一介绍各个经典模型都做了哪些工作，解决了什么问题。

下面的介绍中会提到一些经典的论文，这些论文可以从这里获取：

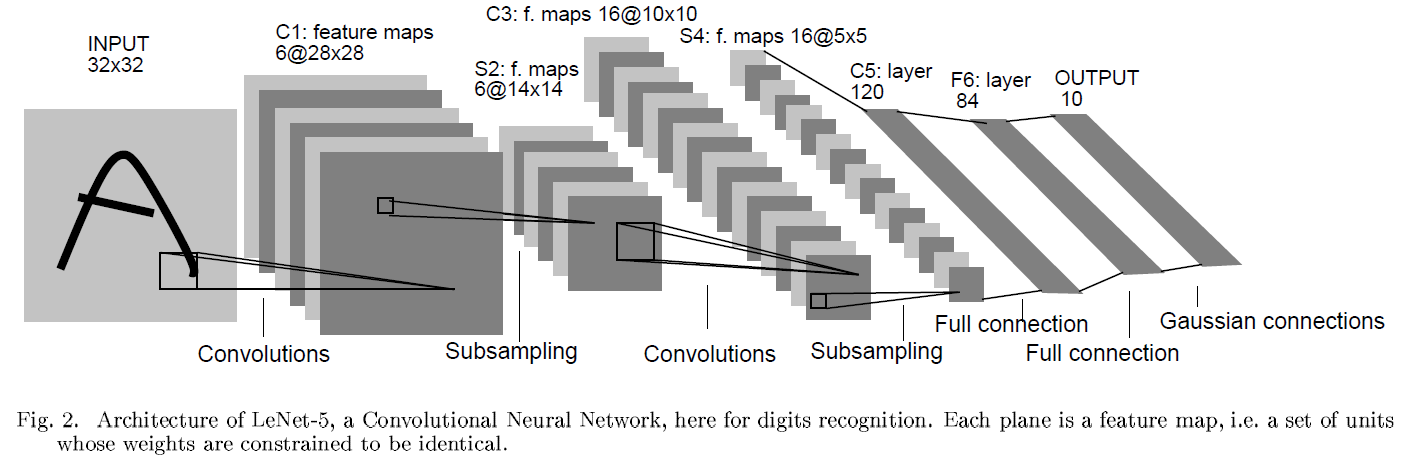
git@github.com:cuiyuan605/deep\_learning.git

供大家参考，另外，这里还包含了以下所述网络模型的部分实现代码。

## 1、LeNet（1998）--开山鼻祖

首先是LetNet，它是由CNN之父Lecun在1998年提出，用于解决手写数字识别任务的，著名的Tensorflow入门学习数据集MNIST，就是那时候做出来的。LeNet对应的论文是《Gradient-Based Learning Applied to Document Recognition》。

LeNet又叫LeNet-5，这个5是指它的网络结构中有5个表示层，具体结构如下图所示：



LeNet5包含Input、卷积层1、池化层1、卷积层2、池化层2、全连接层、输出层。

1998年的LeNet5模型的发布标注着CNN的真正面世，但是这个模型在后来的一段时间并未能火起来，主要原因是当时的计算力跟不上，而且其他的算法（比如SVM）也能达到类似的效果甚至超过。不过LeNet最大的贡献是：**定义了CNN卷积层、池化层、全连接层的基本结构，是CNN的鼻祖**。

**LeNet5当时的特性有如下几点：**

**（1）每个卷积层包含三个部分：卷积、池化和非线性激活函数**

**（2）使用卷积提取空间特征（起初被称为感受野，未提“卷积”二字）**

**（3）降采样（Subsample）的平均池化层（Average Pooling）**

**（4）双曲正切（Tanh）的激活函数**

**（5）MLP作为最后的分类器**

**（6）使用卷积令层与层之间稀疏连接，减少了计算复杂性**

不过，一般认为CNN的开山之作是Lecun 1989发表的另一篇论文《Backpropagation Applied to Handwritten Zip Code Recognition》。这篇论文写的比较短，通篇没有找到CNN模型相关的原理解释说明。当时也没有提到“卷积”二字，而是被称为“感受野”。

但是从它的参考文献中，我们可以找到一篇重要的文章《1980-Fukushima-Neocognitron A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position》。这是日本学者福岛邦彦，通过研究猫的视觉神经得到的启发，阐述了卷积和池化两个思想（当时还不叫卷积和池化）。

福岛邦彦的团队当时也开发了仿真模型Neocognitron（神经认知机），不过Neocognitron为什么没有发扬光大呢，感觉和CNN相比主要还是差在BP算法上，直到1998年Lecun提出了LeNet-5，把CNN推上了一个小高潮，之后Neocognitron也基本上放弃治疗了。

1998年到2012年世界在不断发生着巨大的变化，CNN也在默默的发展，其中有一篇论文要提一下，就是MIT脑科学研究中心的论文《Notes on Convolutional Neural Networks》，里面给了详细的CNN权值更新的公式。

## 2、AlexNet（2012）--王者归来

AlexNet在2012年ImageNet竞赛中以超过第二名10.9个百分点的绝对优势一举夺冠，这一突破是具有重大历史意义，因为它使计算机视觉模型跨过了从学术demo到商业化产品的门槛，从而将深度学习和CNN的名声突破学术界，在产业界一鸣惊人。AlexNet的出现可谓是卷积神经网络的王者归来。

在这里我们先八卦一下人工智能大神之间的关系。2018年图灵奖获得者是号称深度学习三巨头的Yann LeCun（扬·莱坎），Geoffrey Hinton（杰弗里·欣顿），和Yoshua Bengio（约舒亚·本希奥），我把他们称作深度学习三主神，就是下面这三位，我们先来一起瞻仰一下：



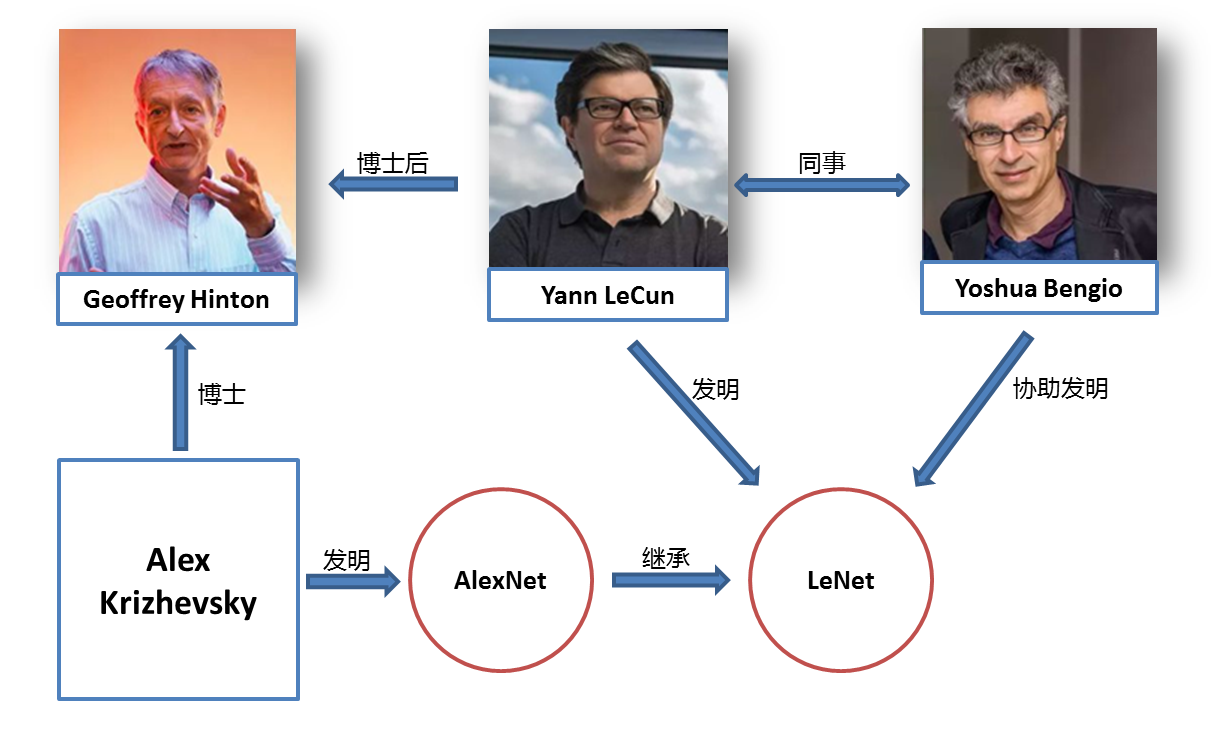
由于他们早年都在加拿大的大学工作过，是非常亲密的朋友和合作者，所以AI领域的人戏称他们“深度学习加拿大黑手党”。他们是上世纪七八十年代，深度学习的“寒冬”里，为数不多坚持研究神经网络的“顽固派”，是神经网络的坚定信仰者，也是他们一手把深度学习带出了“寒冬”。

其中Hinton年纪较大，1986年完成著名论文《Experiments on Learning by Back Propagation》第一个证明了广义反向传播算法可用于训练多层神经网络。

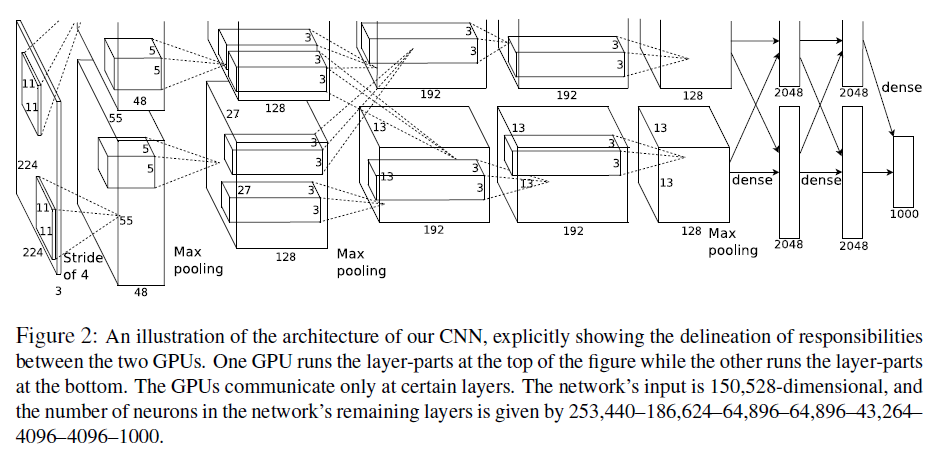
Lecun曾在多伦多大学跟随Hinton做博士后研究，期间完成了LeNet的雏形思想。

Bengio和LeCun年纪相仿，曾一起在AT&T贝尔实验室的共事，与Lucun一起完成了LeNet。

而AlexNet的作者Alex Krizhevsky，是Hinton的博士生，所以说AlexNet和LeNet可谓是一脉相承的。关系如下所示：



我们言归正传，AlexNet为8层深度的CNN网络，其中包括5个卷积层和3个全连接层（不包括LRN层和池化层），如下图所示：



**AlexNet的特点和贡献：**

**（1）使用ReLU作为激活函数**，由于ReLU是非饱和函数，也就是说它的导数在大于0时，一直是1，因此解决了Sigmoid激活函数在网络比较深时的梯度消失问题，提高SGD（随机梯度下降）的收敛速度。

**（2）使用Dropout方法避免模型过拟合**，该方法通过让全连接层的神经元（该模型在前两个全连接层引入Dropout）以一定的概率失去活性（比如0.5），失活的神经元不再参与前向和反向传播，相当于约有一半的神经元不再起作用。在预测的时候，让所有神经元的输出乘Dropout值（比如0.5）。这一机制有效缓解了模型的过拟合。

**（3）重叠的最大池化**，之前的CNN中普遍使用平均池化，而AlexNet全部使用最大池化，避免平均池化的模糊化效果。并且，池化的步长小于核尺寸，这样使得池化层的输出之间会有重叠和覆盖，提升了特征的丰富性。

**（4）提出LRN（局部响应归一化）层**，对局部神经元的活动创建竞争机制，使得响应较大的值变得相对更大，并抑制其他反馈较小的神经元，增强了模型的泛化能力。

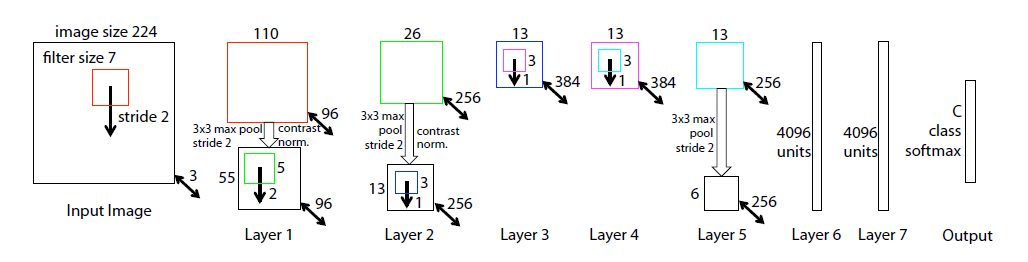
**（5）使用GPU加速**，加快了模型的训练速度，同时也意味着可以训练更大规模的神经网络模型。

**（6）数据增强**，随机从256\*256的原始图像中截取224\*224大小的区域（以及水平翻转的镜像），相当于增强了（256-224）\*（256-224）\*2=2048倍的数据量。使用了数据增强后，减轻过拟合，提升泛化能力。避免因为原始数据量的大小使得参数众多的CNN陷入过拟合中。

## 3、ZFNet（2013）--稳步前行

ZFNet是2013ImageNet分类任务的冠军，其网络结构没什么改进，只是调了调参，性能较Alex提升了不少。ZF-Net只是将AlexNet第一层卷积核由11变成7，步长由4变为2，第3，4，5卷积层转变为384，384，256。这一年的ImageNet还是比较平静的一届，其冠军ZF-Net的名堂也没其他届的经典网络架构响亮。

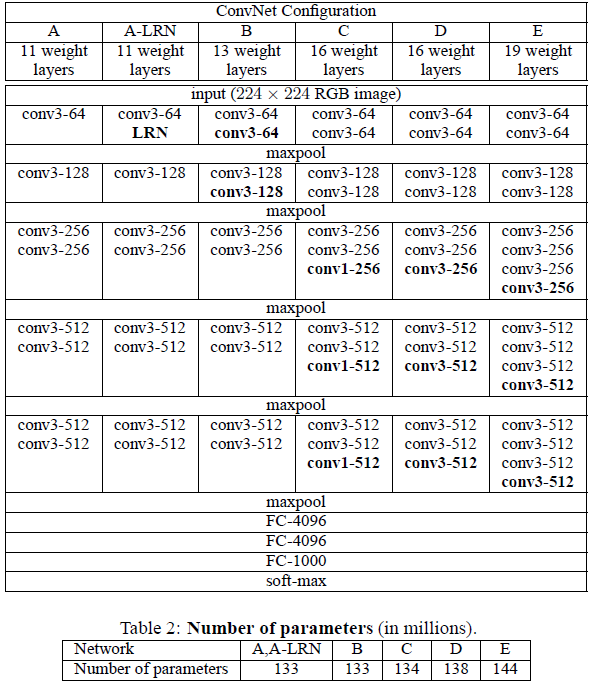
其论文是《Visualizing and Understanding Convolutional Networks》，网络结构如下：



## 4、VGG（2014 2rd）--越走越深

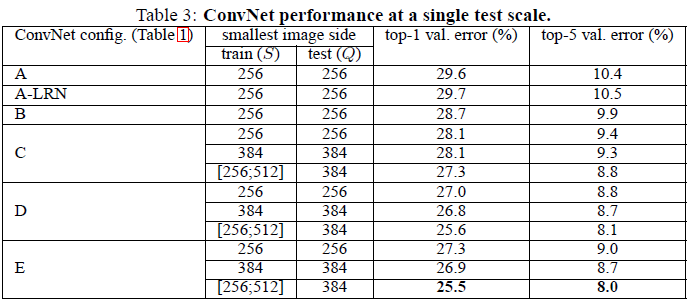
VGG-Nets是由牛津大学VGG（Visual Geometry Group）提出，是2014年ImageNet竞赛定位任务的第一名和分类任务的第二名的中的基础网络。VGG可以看成是加深版本的AlexNet. 都是卷积层+全连接层，在当时看来这是一个非常深的网络了，因为层数高达十多层，我们从其论文名字就知道了（《Very Deep Convolutional Networks for Large-Scale Visual Recognition》），当然以现在的目光看来VGG真的称不上是一个very deep的网络。

VGGNet探索了CNN的深度及其性能之间的关系，通过反复堆叠3\*3的小型卷积核和2\*2的最大池化层，VGGNet成功的构筑了16-19层深的CNN。VGGNet有A-E六种结构，从A-E网络逐步变深，但是参数量并没有增长很多，如下图所示，原因为：参数量主要消耗在最后3个全连接层，而前面的卷积层虽然层数多，但消耗的参数量不大。不过，卷积层的训练比较耗时，因为其计算量大。



其中，D和E是常说的VGGNet-16和VGGNet-19。C很有意思，相比于B多了几个1\*1的卷积层，1\*1卷积的意义在于增加非线性变换，而输入的通道数和输出的通道数不变，没有发生降维。

VGG的性能：



**VGG的特点：**

**（1）VGGNet拥有5段卷积，每段卷积内有2-3个卷积层，同时每段尾部都会连接一个最大池化层（用来缩小图片）。**

**（2）每段内的卷积核数量一样，越后边的段内卷积核数量越多**，依次为：64-128-256-512-512。

**（3）越深的网络效果越好。**

**（4）LRN层作用不大（作者结论）。**

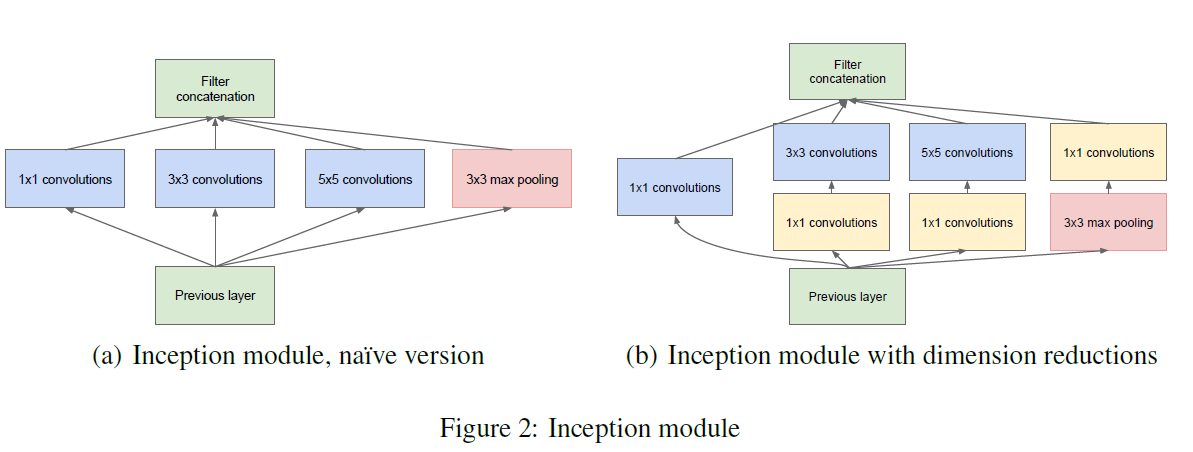
**（5）1x1的卷积也是很有效的，但是没有3\*3的卷积好**，大一些的卷积核可以学习更大的空间特征。不过1x1的卷积核可以用于增加模型的非线性变化，并可用于升维和降维。

**VGG的突出贡献在于，证明了使用小的卷积核，增加网络深度可以有效的提高模型效果**，而且VGGNet对其他数据集具有很好的泛化能力。到目前为止，VGGNet依然经常被用来提取图像特征。

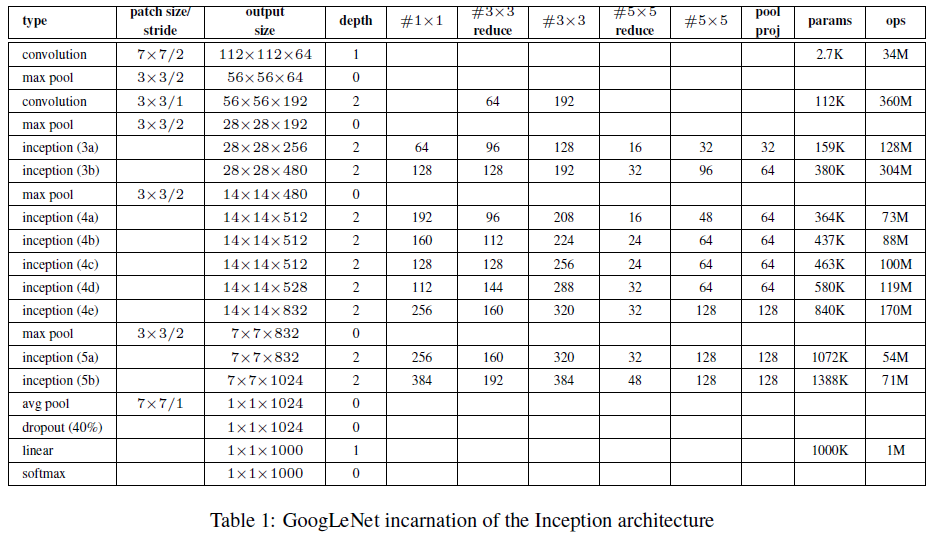
为什么说小的卷积核堆积能够提高模型效果呢？比如3x3的卷积核堆叠两层，则“感受野”就会变为5x5，堆叠三层，“感受野”就会变成7x7，而多层小的卷积核参数更少（3\*3\*3=27<1\*7\*7=49），且增加了更多的非线性变化（三层卷积有三次ReLU），这样增加了模型的表达能力。

## 5、GoogLeNet（2014） Inception V1 V2 V3 V4，Inception-ResNet（2016）--大浪推手

Google Inception Net首次出现在ILSVRC2014的比赛中(和VGGNet同年)，以较大的优势获得冠军。那一届的GoogLeNet通常被称为Inception V1，Inception V1的特点是控制了计算量的参数量的同时，获得了非常好的性能-top5错误率6.67%, 这主要归功于GoogLeNet中引入一个新的网络结构Inception模块（该设计灵感来源于《Network in Network》，也就是NIN），所以GoogLeNet又被称为Inception V1(后面还有改进版V2、V3、V4)架构中有22层深，V1比VGGNet和AlexNet都深，但是它只有500万的参数量，计算量也只有15亿次浮点运算，在参数量和计算量下降的同时保证了准确率，可以说是非常优秀并且实用的模型。Inception模块如下所示：



具体Inception V1的结构如下：



这里只放了一张表，因为模型示意图太长放不下。想要看模型图的可以去看下论文。GoogLeNet的论文为《Going deeper with convolutions》，在这篇论文里，我看到了几个华人的名字（其实NIN也是新加坡国立大学的三个华人发表的），说明大约从这个时候开始，华人科学家开始走到CNN研究的前列。其中有一个名字有点儿熟悉，就是前段时间上了新闻的贾扬清（2019年3月份加入阿里巴巴硅谷研究院），他开发了著名的机器学习框架Caffe，并在Google参与设计和开发了Tensorflow。

**GoogLeNet特点：**

**（1）引入Inception结构**，这是一种网中网（Network In Network）的结构，即原来的结点也是一个网络。

**（2）中间层的辅助LOSS单元**，GoogLeNet网络结构中有3个LOSS单元，这样的网络设计是为了帮助网络的收敛。在中间层加入辅助计算的LOSS单元，目的是计算损失时让低层的特征也有很好的区分能力，从而让网络更好地被训练。在论文中，这两个辅助LOSS单元的计算被乘以0.3，然后和最后的LOSS相加作为最终的损失函数来训练网络。

**（3）后面的全连接层全部替换为简单的全局平均池化层**，在最后参数会变的更少。而在AlexNet中最后3层的全连接层参数差不多占总参数的90%，使用大网络在宽度和深度上允许GoogleNet移除全连接层，但并不会影响到结果的精度，在ImageNet中实现93.3%的精度，而且要比VGG还要快。

之后Inception还出了V2 V3 V4，它们分别对应的论文是：

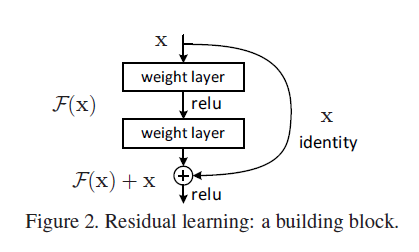
2015年，Inception V2，《Batch Normalization Accelerating Deep Network Training by Reducing Internal Covariate Shift》

2015年，Inception V3，《Rethinking the Inception Architecture for Computer Vision》

2016年，Inception V4，《Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning》

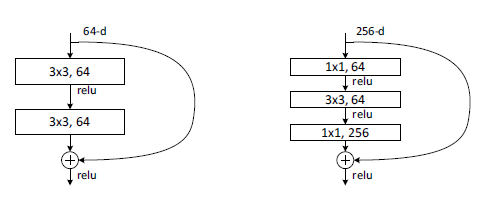
## 6、ResNet（2015）--里程碑式创新

2015年何恺明推出的ResNet在ISLVRC和COCO上横扫所有选手，获得冠军。ResNet在网络结构上做了大创新，而不再是简单的堆积层数，ResNet在卷积神经网络的新思路，绝对是深度学习发展历程上里程碑式的事件。它在VGGNet和MSRANet基础上进一步加深网络，并通过引入残差单元来解决网络过深引起的退化问题。而ResNet的主要创新就是**残差模块**，如下图所示：



残差模块借鉴了Highway Network思想的网络相当于旁边专门开个通道使得输入可以直达输出，而优化的目标由原来的拟合输出F(x)变成输出和输入的差F(x)-x，其中F(x)是某一层原始的期望映射输出，x是输入。

此外，如下图所示：

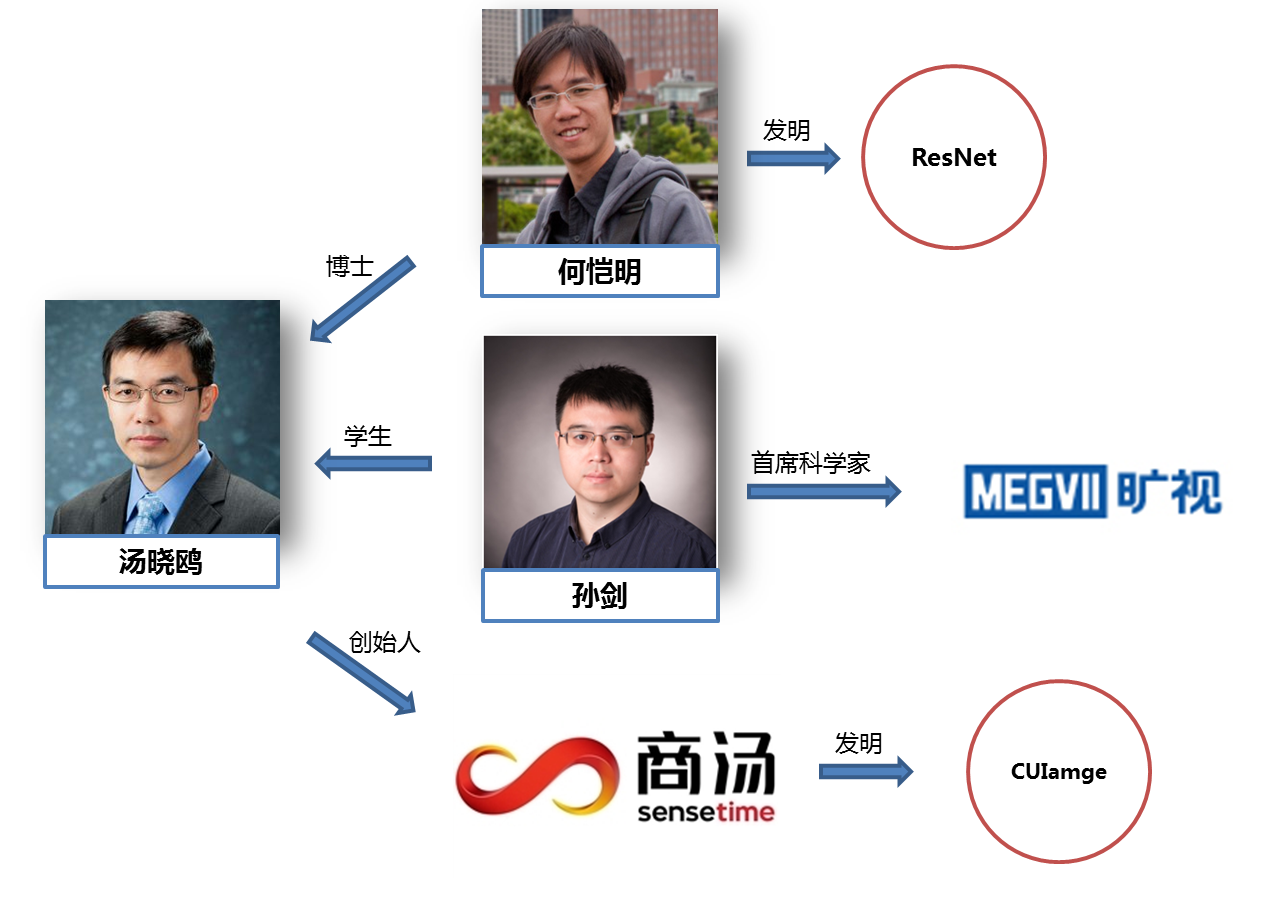


左图是常规残差模块，有两个3×3卷积核组成，但是随着网络进一步加深，这种残差结构在实践中并不是十分有效。针对这问题，右图的“瓶颈残差模块”（bottleneck residual block）可以有更好的效果，它依次由1×1、3×3、1×1这三个卷积层堆积而成，这里的1×1的卷积能够起降维或升维的作用，从而令3×3的卷积可以在相对较低维度的输入上进行，以达到提高计算效率的目的。

## 7、Trimps-Soushen（公安三所），CUImage（商汤和港中文）（2016）--百花齐放

2016年，深度学习的研究进入一个百花齐放的时期，涌现了各种各样的网络模型，中国的公司和研究机构，也开始刷榜机器学习相关竞赛。这一年ISLVRC获胜的队伍是公安三所的Trimps-Soushen（搜神），以及商汤和香港中文大学的CUImage，不过他们模型的相关资料不是很多，在这里就不详细说明了。

不过，在这里我们来八卦一下国内AI领域的大神，汤晓鸥，作为大神肯定有很多的头衔和身份，我在这里只介绍与本文主题最相关的三重身份，分别是：香港中文大学信息工程系主任，微软亚洲研究院视觉计算组主任（2005-2007），商汤科技创始人、董事长。



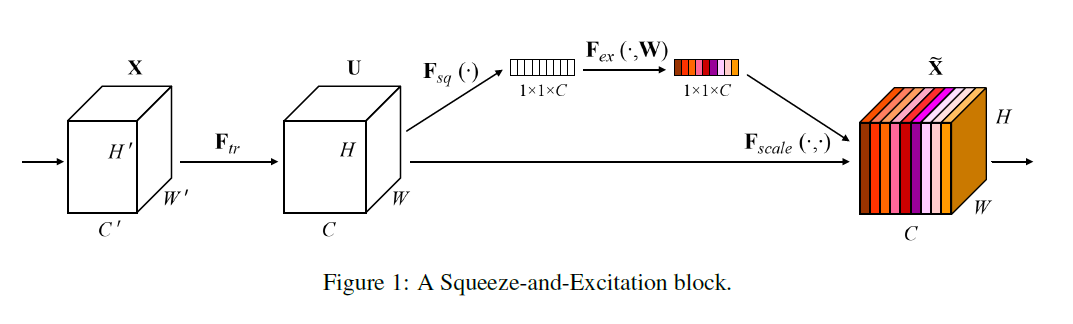
ResNet的作者何恺明，就是汤晓鸥在香港中文大学的博士生。此外，国内计算机视觉方面的另一家巨头独角兽，旷世科技的首席科学家孙剑，也是汤晓鸥在微软亚洲研究院的学生。

此外，国内很多家人工智能企业的骨干都和汤晓鸥有这千丝万缕的联系，可见其在国内AI领域的影响力。

## 8、SENet、DenseNet（2017）--另寻蹊径

**SENet**是2017年最后一届ILSVRC竞赛的图像分类和定位冠军。在2017年之前已经有很多工作在空间维度上来提升网络的性能。那么很自然想到，网络是否可以从其他层面来考虑去提升性能，比如考虑特征通道之间的关系？基于这一点并提出了 **Squeeze-and-Excitation Networks（简称 SENet）**。在SENet结构中，Squeeze 和 Excitation 是两个非常关键的操作，所以以此来命名。动机是希望显式地建模特征通道之间的相互依赖关系。另外，作者并不打算引入一个新的空间维度来进行特征通道间的融合，而是采用了一种全新的「特征重标定」策略。具体来说，就是通过学习的方式来自动获取到每个特征通道的重要程度，然后依照这个重要程度去提升有用的特征并抑制对当前任务用处不大的特征。

下图是提出的 SE 模块的示意图：



给定一个输入x，其特征通道数为c\_1，通过一系列卷积等一般变换后得到一个特征通道数为c\_2 的特征。与传统的 CNN 不一样的是，接下来我们通过三个操作来重标定前面得到的特征。

（1）首先是 Squeeze 操作，我们顺着空间维度来进行特征压缩，将每个二维的特征通道变成一个实数，这个实数某种程度上具有全局的感受野，并且输出的维度和输入的特征通道数相匹配。它表征着在特征通道上响应的全局分布，而且使得靠近输入的层也可以获得全局的感受野，这一点在很多任务中都是非常有用的。

（2）其次是 Excitation 操作，它是一个类似于循环神经网络中门的机制。通过参数 w 来为每个特征通道生成权重，其中参数 w 被学习用来显式地建模特征通道间的相关性。

（3）最后是一个 Reweight 的操作，我们将 Excitation 的输出的权重看作是经过特征选择后的每个特征通道的重要性，然后通过乘法逐通道加权到先前的特征上，完成在通道维度上的对原始特征的重标定。

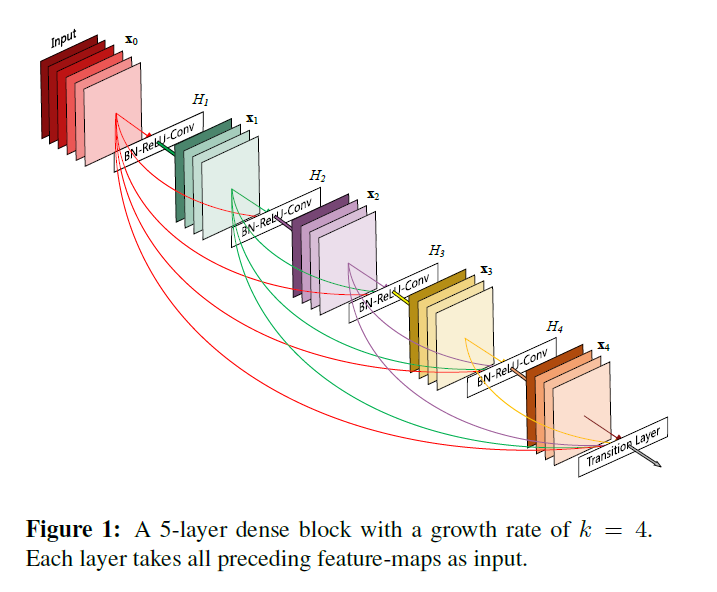
SENet作者的详细说明：《[专栏 | Momenta详解ImageNet 2017夺冠架构SENet](http://www.sohu.com/a/161633191_465975)》

实现代码：<https://github.com/hujie-frank/SENet>

**DenseNet**虽然没有参加ILSVRC竞赛，但它获得了计算机视觉顶级会议CVPR 2017最佳论文。自Resnet提出以后，ResNet的变种网络层出不穷，都各有其特点，网络性能也有一定的提升。DenseNet提出的网络模型（Dense Convolutional Network），主要也是和ResNet及Inception网络做对比，思想上有借鉴，但却是全新的结构，网络结构并不复杂，却非常有效，在CIFAR指标上全面超越ResNet。可以说DenseNet吸收了ResNet最精华的部分，并在此上做了更加创新的工作，使得网络性能进一步提升。

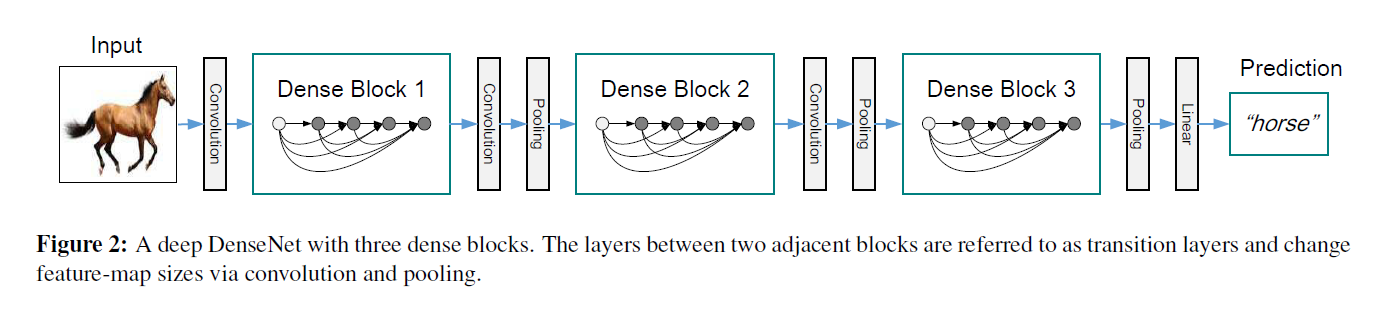
DenseNet的特点是：**密集连接，来缓解梯度消失问题，加强特征传播，鼓励特征复用，极大的减少了参数量**。DenseNet 是一种具有密集连接的卷积神经网络。在该网络中，任何两层之间都有直接的连接，也就是说，网络每一层的输入都是前面所有层输出的并集，而该层所学习的特征图也会被直接传给其后面所有层作为输入。

下图是 DenseNet 的一个dense block示意图：



一个block里面的结构如下，与ResNet中的BottleNeck基本一致： BN-ReLU-Conv(1×1)- BN-ReLU-Conv(3×3) ，而一个DenseNet则由多个这种block组成。每个DenseBlock的之间层称为Transition Layers，由BN−>Conv(1×1)−>AveragePooling(2×2)组成。

需要明确一点，dense connectivity 仅仅是在一个dense block里的，不同dense block 之间是没有dense connectivity的，比如下图所示：



那么问题来了，密集连接不会带来冗余吗？不会！密集连接这个词给人的第一感觉就是极大的增加了网络的参数量和计算量。但实际上 DenseNet 比其他网络效率更高，其关键就在于网络每层计算量的减少以及特征的重复利用。DenseNet则是让i层的输入直接影响到之后的所有层，它的输出为：xi=Hi([x0,x1,…,xi−1])，其中[x0,x1,...,xi−1]就是将之前的feature map以通道的维度进行合并。并且由于每一层都包含之前所有层的输出信息，因此其只需要很少的特征图就够了，这也是为什么DenseNet的参数量较其他模型大大减少的原因。这种dense connection相当于每一层都直接连接input和loss，因此就可以减轻梯度消失现象，这样更深网络不是问题。

天底下没有免费的午餐，网络自然也不例外。在同层深度下获得更好的收敛率，自然是有额外代价的。其代价之一，就是其恐怖如斯的内存占用。

# 三、参考资料

1、【卷积神经网络发展历程】从LeNet、AlexNet到ResNet、SENet： https://blog.csdn.net/u012679707/article/details/80870625

2、CNN网络架构演进：从LeNet到DenseNet： https://www.cnblogs.com/skyfsm/p/8451834.html

3、ILSVRC竞赛详细介绍（ImageNet Large Scale Visual Recognition Challenge）： https://blog.csdn.net/sophia\_11/article/details/84570177

4、CNN（卷积神经网络）最早是哪一年提出，是如何发展的？： https://www.zhihu.com/question/47705441?sort=created

5、图灵奖颁给熬过寒冬的人： http://tech.hexun.com/2019-03-28/196640717.html

6、刚刚！AI界的“加拿大黑手党”荣获 2018 年度图灵奖：https://baijiahao.baidu.com/s?id=1629174737290254826&wfr=spider&for=pc